eLIFE elifesciences.org

FEATURE ARTICLE

RESEARCH

Publication bias and the canonization of false facts

Abstract Science is facing a "replication crisis" in which many experimental findings cannot be replicated and are likely to be false. Does this imply that many scientific facts are false as well? To find out, we explore the process by which a claim becomes fact. We model the community's confidence in a claim as a Markov process with successive published results shifting the degree of belief. Publication bias in favor of positive findings influences the distribution of published results. We find that unless a sufficient fraction of negative results are published, false claims frequently can become canonized as fact. Data-dredging, p-hacking, and similar behaviors exacerbate the problem. Should negative results become easier to publish as a claim approaches acceptance as a fact, however, true and false claims would be more readily distinguished. To the degree that the model reflects the real world, there may be serious concerns about the validity of purported facts in some disciplines.
DOI: 10.7554/eLife.21451.001

SILAS BOYE NISSEN, TALI MAGIDSON, KEVIN GROSS*† AND CARL T BERGSTROM*†

Introduction

Science is a process of collective knowledge creation in which researchers use experimental, theoretical and observational approaches to develop a naturalistic understanding of the world. In the development of a scientific field, certain claims stand out as both significant and stable in the face of further experimentation (***Ravetz, 1971***). Once a claim reaches this stage of widespread acceptance as true, it has transitioned from claim to *fact*. This transition, which we call *canonization*, is often indicated by some or all of the following: a canonized fact can be taken for granted rather than treated as an open hypothesis in the subsequent primary literature; tests that do no more than to confirm previously canonized facts are seldom considered publication-worthy; and canonized facts begin to appear in review papers and textbooks without the company of alternative hypotheses. Of course the veracity of so-called facts may be called back into question (***Arbesman, 2012***; ***Latour, 1987***), but for time being the issue is considered to be settled. Note that we consider facts to be epistemological rather than ontological: a claim is a fact because it is accepted by the relevant community, not because it accurately reflects or represents underlying physical reality (***Ravetz, 1971***; ***Latour, 1987***).

But what is the status of these facts in light of the widely reported replication crisis in science? Large scale analyses have revealed that many published papers in fields ranging from cancer biology to psychology to economics cannot be replicated in subsequent experiments (***Begley and Ellis, 2012***; ***Open Science Collaboration, 2015***; ***Errington et al., 2014***; ***Ebrahim et al., 2014***; ***Chang and Li, 2015***; ***Camerer et al., 2016***; ***Baker, 2016***). One possible explanation is that many published experiments are not replicable because many of their conclusions are ontologically false (***Ioannidis, 2005***; ***Higginson and Munafò, 2016***).

If many experimental findings are ontologically false, does it follow that many scientific facts are ontologically untrue? Not necessarily. Claims of the sort that become facts are rarely if ever tested directly in their entirety. Instead, such claims typically comprise multiple subsidiary hypotheses which must be individually verified. Thus multiple experiments are usually required to establish a claim. Some of these may include direct replications, but more typically an ensemble of distinct experiments will produce

***For correspondence:** krgross@ncsu.edu (KG); cbergst@u.washington.edu (CTB)

†These authors contributed equally to this work

Reviewing editor: Peter Rodgers, eLife, United Kingdom

multiple lines of evidence before a claim is accepted by the community.

For example, as molecular biologists worked to unravel the details of the eukaryotic RNA interference (RNAi) pathway in the early 2000s, they wanted to understand how the RNAi pathway was initiated. Based on work with *Drosophila* cell lines and embryo extracts, one group of researchers made the claim that the RNAi pathway is initiated by the Dicer enzyme which slices double-stranded RNA into short fragments of 20–22 amino acids in length (**Bernstein et al., 2001**). Like many scientific facts, this claim was too broad to be validated directly in a single experiment. Rather, it comprised a number of subsidiary assertions: an enzyme called Dicer exists in eukaryotic cells; it is essential to initiate the RNAi pathway; it binds dsRNA and slices it into pieces; it is distinct from the enzyme or enzyme complex that destroys targeted messenger RNA; it is ubiquitous across eukaryotes that exhibit RNAi pathway. Researchers from numerous labs tested these subsidiary hypotheses or aspects thereof to derive numerous lines of convergent evidence in support of the original claim. While the initial breakthrough came from work in *Drosophila melanogaster* cell lines ((**Bernstein et al., (2001**), subsequent research involved in establishing this fact drew upon in vitro and in vivo studies, genomic analyses, and even mathematical modeling efforts, and spanned species including the fission yeast *Schizosaccharomyces pombe*, the protozoan *Giardia intestinalis*, the nemotode *Caenorhabditis elegans*, the flowering plant *Arabidopsis thaliana*, mice, and humans (**Jaskiewicz and Filipowicz, 2008**). Ultimately, sufficient supporting evidence accumulated to establish as fact the original claim about Dicer's function.

Requiring multiple studies to establish a fact is no panacea, however. The same processes that allow publication of a single incorrect result can also lead to the accumulation of sufficiently many incorrect findings to establish a false claim as fact (**McElreath and Smaldino, 2015**).

This risk is exacerbated by *publication bias* (**Sterling, 1959**; **Rosenthal, 1979**; **Newcombe, 1987**; **Begg and Berlin, 1988**; **Dickersin, 1990**; **Easterbrook et al., 1991**; **Song et al., 2000**; **Olson et al., 2002**; **Chan and Altman, 2005**; **Franco et al., 2014**). Publication bias arises when the probability that a scientific study is published is not independent of its results (**Sterling, 1959**). As a consequence, the findings from published tests of a claim will differ in a systematic way from the findings of all tests of the same claim (**Song et al., 2000**; **Turner et al., 2008**).

Publication bias is pervasive. Authors have systematic biases regarding which results they consider worth writing up; this is known as the "file drawer problem" or "outcome reporting bias" (**Rosenthal, 1979**; **Chan and Altman, 2005**). Journals similarly have biases about which results are worth publishing. These two sources of publication bias act equivalently in the model developed here, and thus we will not attempt to separate them. Nor would separating them be simple; even if authors' behavior is the larger contributor to publication bias (**Olson et al., 2002**; **Franco et al., 2014**), they may simply be responding appropriately to incentives imposed by editorial preferences for positive results.

What kinds of results are most valued? Findings of statistically significant differences between groups or treatments tend to be viewed as more worthy of submission and publication than those of non-significant differences. Correlations between variables are often considered more interesting than the absence of correlations. Tests that reject null hypotheses are commonly seen as more noteworthy than tests that fail to do so. Results that are interesting in any of these ways can be described as "positive".

A substantial majority of the scientific results published appear to be positive ones (**Csada et al., 1996**). It is relatively straightforward to measure the fraction of published results that are negative. One extensive study found that in 2007, more than 80% of papers reported positive findings, and this number exceeded 90% in disciplines such as psychology and ecology (**Fanelli, 2012**). Moreover, the fraction of publications reporting positive results has increased over the past few decades. While this high prevalence of positive results could in principle result in part from experimental designs with increasing statistical power and a growing preference for testing claims that are believed likely to be true, publication bias doubtless contributes as well (**Fanelli, 2012**).

How sizable is this publication bias? To answer that, we need to estimate the fraction of negative results that are published, and doing so can be difficult because we rarely have access to the set of findings that go unpublished. The best available evidence of this sort comes from registered clinical trials. For example, a 2008 meta-analysis examined 74 FDA-registered studies of antidepressants (**Turner et al., 2008**). In

that analysis, 37 of 38 positive studies were published as positive results, but only 3 of 24 negative studies were published as negative results. An additional 5 negative studies were re-framed as positive for the purposes of publication. Thus, negative studies were published at scarcely more than 10% the rate of positive studies.

We would like to understand how the possibility of misleading experimental results and the prevalence of publication bias shape the creation of scientific facts. Mathematical models of the scientific process can help us understand the dynamics by which scientific knowledge is produced and, consequently, the likelihood that elements of this knowledge are actually correct. In this paper, we look at the way in which repeated efforts to test a scientific claim establish this claim as fact or cause it to be rejected as false.

We develop a mathematical model in which successive publications influence the community's perceptions around the likelihood of a given scientific claim. Positive results impel the claim toward fact, whereas negative results lead in the opposite direction. Describing this process, ***Latour, (1987)*** compared the fate of a scientific claim to that of a rugby ball, pushed alternatively toward fact or falsehood by the efforts of competing teams, its fate determined by the balance of their collective actions. Put in these terms, our aim in the present paper is to develop a formal model of how the ball is driven up and down the epistemological pitch until one of the goal lines is reached. In the subsequent sections, we outline the model, explain how it can be analyzed, present the results that we obtain, and consider its implications for the functioning of scientific activity.

Model

In this section, we will develop a simplified model of scientific activity, designed to capture the important qualitative features of fact-creation as a dynamic process.

Model description

We explore a simple model in which researchers sequentially test a single claim until the scientific community becomes sufficiently certain of its truth or falsehood that no further experimentation is needed. Our model is conceptually related to those developed in refs. (***Rzhetsky et al., 2006***; ***McElreath and Smaldino, 2015***), though it is considerably simpler than either since we only consider a single claim at a time.

Figure 1 provides a schematic illustration of the experimentation and publication process. We begin with a claim which is ontologically either true or false. Researchers sequentially conduct experiments to test the claim; these experiments are typically not direct replications of one another, but rather distinct approaches that lend broader support to the claim. Each experiment returns either a positive outcome supporting the claim, or a negative outcome contravening it. For mathematical simplicity, we assume all tests to have the same error rates, in the sense that if the claim under scrutiny is false, then investigators obtain false positives with probability α. Conversely, when the claim is true, investigators obtain false negatives with probability β. We take these error rates to be the ones that are conventionally associated with statistical hypothesis testing, so that α is equivalent to the significance level (technically, the *size*) of a statistical test and $1-\beta$ is the test's power. We assume that, as in any reasonable test, a true claim is more likely to generate a positive result than a negative one: $1-\beta>\alpha$. A broader interpretation of α and β beyond statistical error does not change the interpretation of our results.

After completing a study, the investigators may attempt to publish their experimental results. However, publication bias occurs in that the result of the experiment influences the chance that a study is written up as a paper and accepted for publication. Positive results are eventually published somewhere with probability ρ_1 whereas negative results are eventually published somewhere with probability ρ_0. Given the reluctance of authors to submit negative results and of journals to publish them, we expect that in general $\rho_1>\rho_0$.

Finally, readers attempt to judge whether a claim is true by consulting the published literature only. For modeling purposes, we will consider a best-case scenario, in which the false positive and false negative rates α and β are established by disciplinary custom or accepted benchmarks, and readers perform Bayesian updating of their beliefs based upon these known values. In practice, these values may not be as well standardized or widely reported as would be desirable. Moreover, readers are unlikely to be this sophisticated in drawing their inferences. Instead readers are likely to form subjective beliefs in an informal fashion based on a general assessment of the accumulated positive and negative results and the strength of each. But the Bayesian updating case provides a

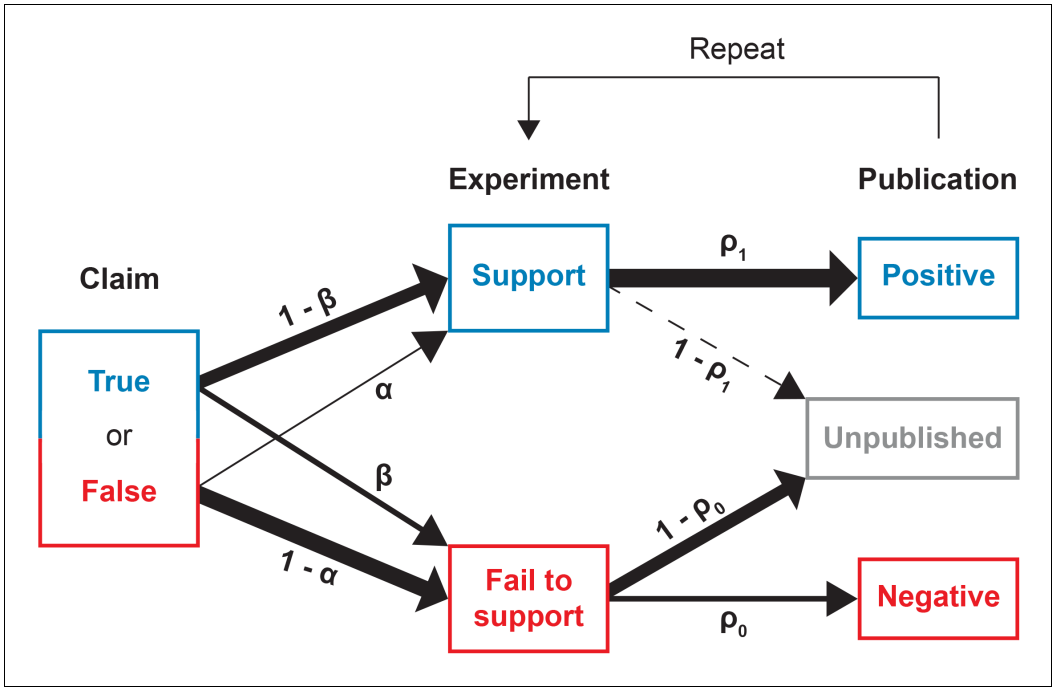

Figure 1. Conducting and reporting the test of a claim. In our model, a scientific claim is either true or false. Researchers conduct an experiment which either supports or fails to the support the claim. True claims are correctly supported with probability $1-\beta$ while false claims are incorrectly supported with probability α. Next, the researchers may attempt to publish their results. Positive results that support the claim are published with probability ρ_1 whereas negative results that fail to support the claim are published with probability ρ_0. This process then repeats, with additional experiments conducted until the claim is canonized as fact or rejected as false.
DOI: 10.7554/eLife.21451.002

well-defined model under which to explore the distortion of belief by publication bias.

The problem is that the results described in the published literature are now biased by the selection of which articles are drafted and accepted for publication. We assume that readers are unaware of the degree of this bias, and that they fail to correct for publication bias in drawing inferences from the published data. It may seem pessimistic that researchers would fail to make this correction, but much of the current concern over the replication crisis in science is predicated on exactly this. Moreover, it is usually impossible for a researcher to accurately estimate the degree of publication bias in a given domain.

Model dynamics

Consider a claim that the community initially considers to have probability q_0 of being true. Researchers iteratively test hypotheses that bear upon the claim until it accumulates either sufficient support to be canonized as fact, or sufficient counter-evidence to be discarded as false. If the claim is true, the probability that a single test leads to a positive publication is $(1-\beta)\rho_1$, and the corresponding probability of a negative publication is $\beta\rho_0$. The remaining probability corresponds to results of the test not being published. If the claim is false, these probabilities are $\alpha\rho_1$ and $(1-\alpha)\rho_0$ for positive and negative published outcomes, respectively. Given that a claim is true, the probability that a published test of that claim reports a positive outcome is therefore

$$\omega_T = \frac{(1-\beta)\rho_1}{(1-\beta)\rho_1 + \beta\rho_0}. \quad (1)$$

For a false claim, the probability that a published test is positive is

$$\omega_F = \frac{\alpha\rho_1}{\alpha\rho_1 + (1-\alpha)\rho_0}. \quad (2)$$

Because only the ratio of ρ_1 to ρ_0 matters for the purposes of our model, we set ρ_1 to 1 for the remainder of the paper. We initially assume that ρ_0 is constant, but will relax this assumption later.

To formalize ideas, consider a sequence of published outcomes X, and let Y_k be the number of positive published outcomes in the first k terms of X. When the probability of publishing a negative result ρ_0 is constant, the outcomes of published experiments are exchangeable random variables. Thus after k published tests, the distribution of Y_k for a true claim is the binomial distribution $\mathrm{Bin}(k,\omega_T)$ and for a false claim is $\mathrm{Bin}(k,\omega_F)$. Moreover, the sequence $\{Y_k\}_{k=1}^{\infty}$ is a Markov chain. When the extent of publication bias is known, we can compute the conditional probability that a claim is true, given $Y_k = y$, as

$$\frac{\omega_T^y(1-\omega_T)^{k-y}q_0}{\omega_T^y(1-\omega_T)^{k-y}q_0+\omega_F^y(1-\omega_F)^{k-y}(1-q_0)}. \quad (3)$$

We now consider the consequences of drawing inferences based on the published data alone, without correcting for publication bias. For model readers who do not condition on publication bias, let $q_k(y)$ be the perceived, conditional probability that a claim is true given $Y_k = y$. We say "perceived" because these readers use Bayes' Law to update q_k, but do so under the incorrect assumption that there is no publication bias, i.e., that $\rho_0 = \rho_1 = 1$. To ease the narrative, we refer to the perceived conditional probability that a claim is true as the "belief" that the claim is true. Expressing this formally,

$$q_k(y) = \frac{(1-\beta)^y\beta^{k-y}q_0}{(1-\beta)^y\beta^{k-y}q_0+\alpha^y(1-\alpha)^{k-y}(1-q_0)}. \quad (4)$$

Note that without publication bias, we have $\omega_T = (1-\beta)$ and $\omega_F = \alpha$, and thus ***Equation 3*** coincides with ***Equation 4***.

From the perspective of an observer who is unaware of any publication bias, the pair (Y_k,k) is a sufficient statistic for the random variable $A \in \{\mathrm{True},\mathrm{False}\}$ representing the truth value of the claim in question. This follows from the definition of statistical sufficiency and the fact that

$$\begin{aligned}\mathrm{prob}\ &\left[A=\mathrm{True}|Y_k,k,\{Y_i\}_{i=1}^{k}\right]=q_k(y)\\ &=\mathrm{prob}[A=\mathrm{True}|Y_k,k].\end{aligned}$$

By analogous logic, the pair (Y_k,k) is also a sufficient statistic for an observer aware of the degree of publication bias provided that the publication probabilities ρ_0 and ρ_1 are constant.

We envision science as proceeding iteratively until the belief that a claim is true is sufficiently close to 1 that the claim is canonized as fact, or until belief is sufficiently close to 0 that the claim is discarded as false. We let τ_0 and τ_1 be the belief thresholds at which a claim is rejected or canonized as fact, respectively ($0<\tau_0<\tau_1<1$), and refer to these as evidentiary standards. In our analysis, we make the simplifying assumption that the evidentiary standards are symmetric, i.e., $\tau_0 = 1-\tau_1$. We describe the consequences of relaxing this assumption in the Discussion.

Thus, mathematically, we model belief in the truth of a claim as a discrete-time Markov chain $\{q_k\}_{k=0}^{\infty}$ with absorbing boundaries at the evidentiary standards for canonization or rejection (***Figure 2A***). When the Markov chain represents belief, its possible values lie in the interval from 0 to 1. For mathematical convenience, however, it is often helpful to convert belief to the log odds scale, that is, $\ln(q_k/(1-q_k))$. Some algebra shows that the log odds of belief $q_k(y)$ can be written as

$$\begin{aligned}\ln\left(\frac{q_k(y)}{1-q_k(y)}\right) &= y\ln\left(\tfrac{1-\beta}{\alpha}\right)+(k-y)\ln\left(\tfrac{\beta}{1-\alpha}\right)\\ &\quad+\ln\left(\tfrac{q_0}{1-q_0}\right).\end{aligned} \quad (5)$$

The log odds scale is convenient because, as ***Equation. 5*** shows, each published positive outcome increases the log odds of belief by a constant increment

$$d_1 = \ln\left(\frac{1-\beta}{\alpha}\right)>0$$

(***Figure 2B***). Each published negative outcome decreases the log odds of belief by

$$d_0 = \ln\left(\frac{\beta}{1-\alpha}\right)<0.$$

Below, we will see that much of the behavior of our model can be understood in terms of the expected change in the log odds of belief for each published outcome. For a true claim, the expected change in the log odds of belief is

$$d_1\omega_T + d_0(1-\omega_T) \quad (6)$$

whereas for a false claim, the expected change in the log odds of belief is

$$d_1\omega_F + d_0(1-\omega_F). \quad (7)$$

Computing canonization and rejection probabilities

In general, we cannot obtain a closed-form expression for the probability that a claim is canonized as fact or for the probability that it is rejected as false. We can, however, derive recursive expressions for the probabilities that after k published experiments a claim has been

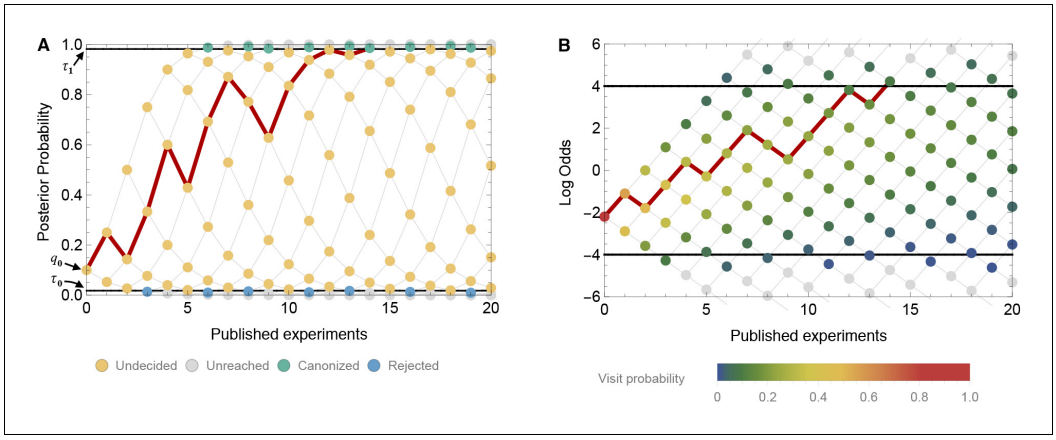

Figure 2. A time-directed graph represents the evolution of belief over time. In panel **A**, the horizontal axis indicates the number of experiments published and the vertical axis reflects the observer's belief, quantified as the probability that the claim is true. The process begins at the single point at far left with an initial belief q_0. Each subsequent experiment either supports the claim, moving to the next node up and right, or contradicts the claim, moving to the next node down and right. At yellow nodes, the status of the claim is as yet undecided. At green nodes, it is canonized as fact, and at blue nodes, it is rejected as false. The black horizontal lines show the evidentiary standards (τ_0 and τ_1). The red path shows one possible trajectory, in which a positive experiment is followed by a negative, then two positives, then a negative, etc., ultimately becoming canonized as fact when it reaches the upper boundary. Panel **B** shows the same network, but with the vertical axis representing log odds and using color to indicate the probability that the process visits each node. In log-odds space, each published positive result shifts belief by the constant distance $d_1>0$ and each negative result by a different distance $d_0<0$. Shown here (in both panel **A** and **B**) is a false claim with false positive rate $\alpha = 0.2$, false negative rate $\beta = 0.4$, publication probabilities $p_0 = 0.1$ and $p_1 = 1$, and initial belief $q_0 = 0.1$. In this case, the claim is likely to be canonized as fact, despite being false.

DOI: 10.7554/eLife.21451.003

canonized as fact, has been discarded as false, or remains undecided. From these, it is straightforward to compute the canonization and rejection probabilities numerically to any desired level of precision.

For each number of published experiments k, the state space for Y_k is simply $Y_k \in \{0, 1, \ldots, k\}$. Partition this state space as follows:

$$\mathcal{C}_k = \{y : q_k(y) > \tau_1\}$$
$$\mathcal{I}_k = \{y : q_k(y) \in [\tau_0, \tau_1]\}$$
$$\mathcal{R}_k = \{y : q_k(y) < \tau_0\}.$$

That is, $\mathcal{C}_k$ is the set of outcomes corresponding to a belief greater than the evidentiary standard for canonization, $\mathcal{R}_k$ is the set of outcomes corresponding to a belief less than the evidentiary standard for rejection, and $\mathcal{I}_k$ is the set of outcomes corresponding to belief in between these two standards (the "interior"). Let T be the number of publications until a claim is either canonized or rejected. Formally,

$$T = \min\{k : Y_k \in C_k \cup R_k\}.$$

For a true claim, we calculate the probability of canonization as follows. (A parallel set of equations gives the probability of canonization for a false claim.) For each $y \in \mathcal{I}_k$, define $p_k(y) = \text{Prob}\{Y_k = y, T\}$. That is, $p_k(y)$ is the probability that there are exactly y positive outcomes in the first k publications, and the claim has yet to be canonized or rejected by publication k. Suppose these probabilities are known for each $y \in \mathcal{I}_k$. Then for each $y \in \mathcal{I}_{k+1}$, these probabilities can be found recursively by

$$p_{k+1}(y) = \omega_T p_k(y-1) + (1-\omega_T) p_k(y).$$

For computational purposes, in the recursion above we define $p_k(y) = 0$ whenever $y \notin \mathcal{I}_k$. The probability that the claim has yet to be canonized or rejected by publication k is simply

$$\text{Prob}\{T > k\} = \sum_{y \in \mathcal{I}_k} p_k(y).$$

Let ϕ_k be the probability that a claim is first canonized at publication k. Formally,

$$\phi_k = \sum_{y : y-1 \in \mathcal{I}_{k-1} \text{ and } y \in \mathcal{C}_k} \omega_T p_{k-1}(y-1).$$

Let $k^\star$ be the smallest value of k for which

$\mathrm{Prob}\{T\} \leq \epsilon$. To calculate the probability of canonization, we calculate $p_k(y)$ for all $k = 1, \ldots, k^\star$. The probability of canonization is then $\sum_{k=1}^{k^\star} \phi_k$. For the analyses in this paper, we have set $\epsilon = 10^{-4}$.

Results

We focus throughout the paper on the dynamic processes by which false claims are canonized as facts, and explore how the probability of this happening depends on properties of the system such as the publication rate of negative results, the initial beliefs of researchers, the rates of experimental error, and the degree of evidence required to canonize a claim. In principle, the converse could be a problem as well: true claims could be discarded as false. However, this is rare in our model. Publication bias favors the publication of positive results and therefore will not tend to cause true claims to be discarded as false, irrespective of other parameters. We first establish this, and then proceed to a detailed examination of how scientific experimentation and publication influences the rate at which false claims are canonized as fact.

True claims tend to be canonized as facts

In our model, true claims are almost always canonized as facts. ***Figure 3*** illustrates this result in the form of a receiver operating characteristic (ROC) curve. Holding the other parameters constant, the curve varies the negative publication rate ρ_0, and uses the vertical and horizontal axes to indicate the probabilities that true and false claims respectively are canonized as fact.

One might fear that as the probability ρ_0 of publishing negative results climbs toward unity, the risk of rejecting a true claim would increase dramatically as well. This is not the case. Even as the probability of publishing negative results approaches 1, the risk of rejecting a true claim is low when evidentiary standards are lax (***Figure 3A***), and negligible when evidentiary standards are strict (***Figure 3B***).

This result turns out to be general across a broad range of parameters. Assuming the mild requirements that (i) tests of a true claim are more likely to result in positive publications than negative publications (i.e., $\omega_T > 1/2$, or equivalently $(1-\beta) > \beta\rho_0$), and (ii) positive published outcomes increase belief that the claim is true ($d_1 > 0$, or equivalently $(1-\beta) > \alpha$), true claims are highly likely to be canonized as facts. The exceptions occur only when minimal evidence is

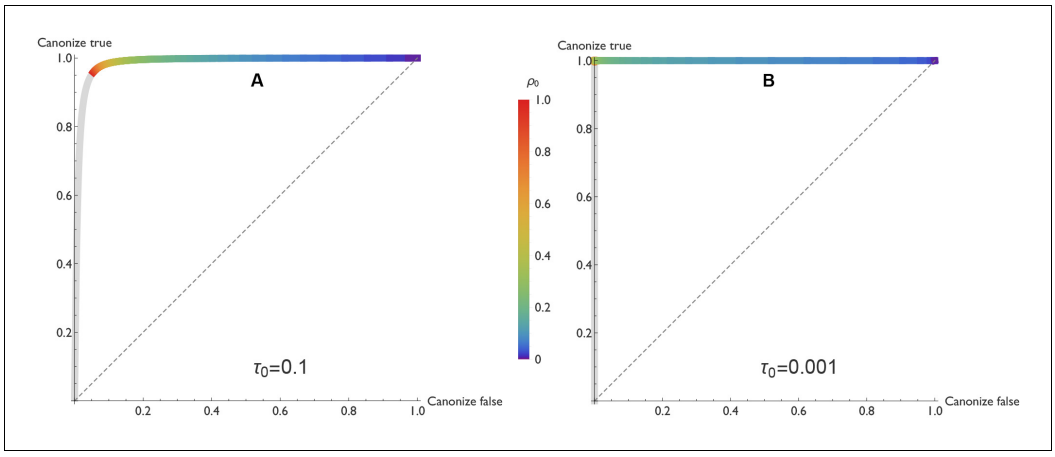

Figure 3. ROC curves reveal that true claims are almost always canonized as fact. In the receiver operating characteristic (ROC) curves shown here, the vertical axis represents the probability that a true claim is correctly canonized as fact, and the horizontal axis represents the probability that a false one is incorrectly canonized as fact. Panel **A**: lax evidentiary standards $\tau_0 = 0.1$ and $\tau_1 = 0.9$. Panel **B**: strict evidentiary standards $\tau_0 = 0.001$ and $\tau_1 = 0.999$. Error rates and initial belief are $\alpha = 0.05$, $\beta = 0.2$, and $q_0 = 0.5$. Each point along the ROC curve corresponds to a different value of the negative publication rate, ρ_0, as indicated by color. Grey regions of the curve correspond to the unlikely situations in which $\rho_0 > \rho_1 = 1$, i.e., negative results are more likely to be published than positive ones. The figures reveal two important points. First, when negative results are published at any rate $\rho_0 \leq 1$, the vast majority of true claims are canonized as fact. Second, when negative results are published at a low rate (ρ_0 less than 0.3 or 0.2 depending on evidentiary standards), many false claims will also be canonized as true.
DOI: 10.7554/eLife.21451.004

needed to discard a claim, i.e., when initial belief is small ($q_0 \approx \tau_0$). In such cases a bit of bad luck—the first one or two published experiments report false negatives, for example—can cause a true claim to be rejected. But otherwise, truth is sufficient for canonization.

Unfortunately, truth is not required for canonization. The risk of canonizing a false claim—shown on the horizontal axis in ***Figure 3***—is highly sensitive to the rate at which negative results are published. When negative results are published with high probability, false claims are seldom canonized. But when negative results are published with lower probability, many false claims are canonized.

Thus we see that the predominant risk associated with publication bias is the canonization of false claims. In the remainder of this analysis, we focus on this risk of incorrectly establishing a false claim as a fact.

Publication of negative results is essential

As we discussed in the introduction, authors and journals alike tend to be reluctant to publish negative results, and as we found in the previous subsection, when most negative results go unpublished, science performs poorly. Here, we explore this relationship in further detail.

Figure 4 shows how the probability of erroneously canonizing a false claim as fact depends on the probability ρ_0 that a negative result is published. False claims are likely to be canonized below a threshold rate of negative publication, and unlikely to be canonized above this threshold. For example, when the false positive rate α is 0.05, the false negative rate β is 0.4,

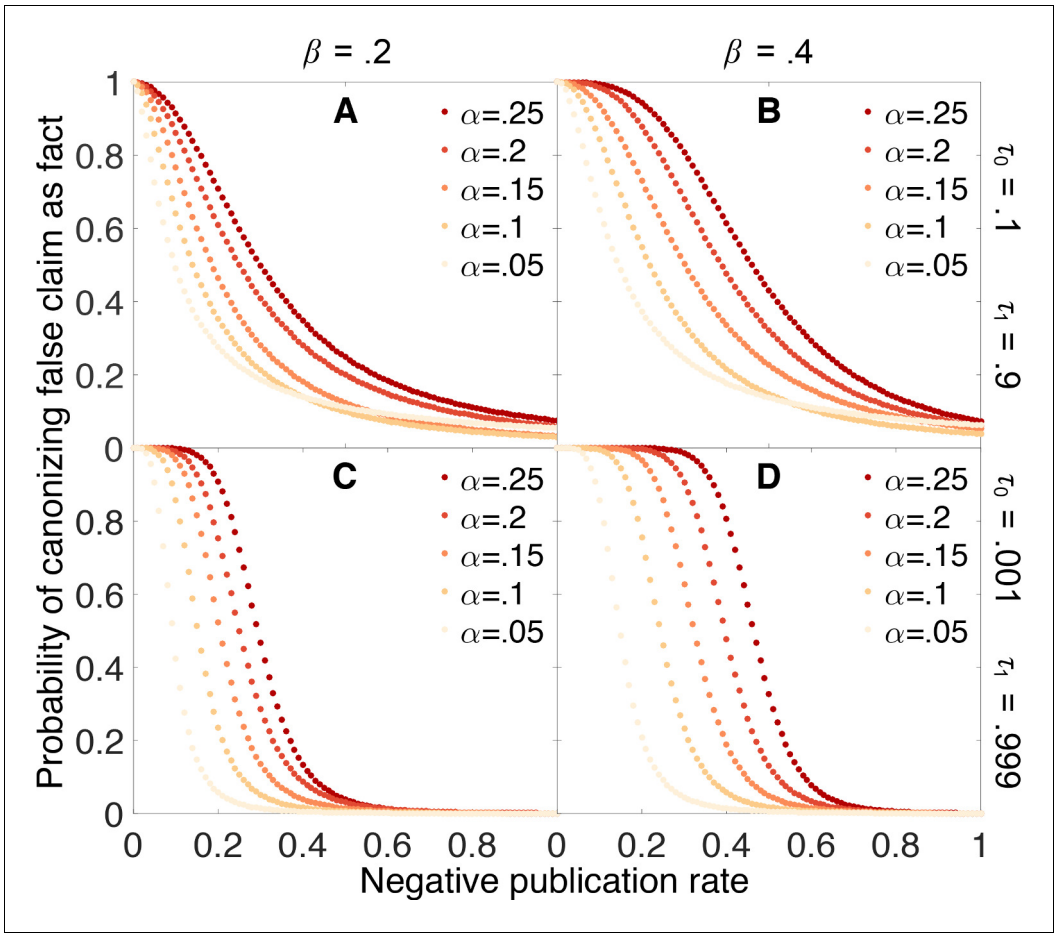

Figure 4. Publishing negative outcomes is essential for rejecting false claims. Probability that a false claim is incorrectly canonized, as a function of the negative publication rate. Throughout, initial belief is $q_0 = 0.5$, and individual data series show false positive rates $\alpha = 0.05$ (yellow), $0.10, \ldots, 0.25$ (red). Top row: weak evidentiary standards $\tau_0 = 0.1$ and $\tau_1 = 0.9$. Panel **A**: false negative rate $\beta = 0.2$. Panel **B**: $\beta = 0.4$. Panels **C**–**D**: similar to panels **A**–**B**, with more demanding evidentiary standards $\tau_0 = 0.001$ and $\tau_1 = 0.999$.
DOI: 10.7554/eLife.21451.005

and the evidentiary requirements are strong (yellow points in Panel 4D), a false claim is likely to be canonized as fact unless negative results are at least 20% as likely as positive results to be published.

Figure 4 also reveals that the probability of canonizing false claims as facts depends strongly on both the false positive rate and the false negative rate of the experimental tests. As these error rates increase, an increasingly large fraction of negative results must be published to preserve the ability to discriminate between true and false claims.

Initial beliefs usually do not matter much

If the scientific process is working properly, it should not automatically confirm what we already believe, but rather it should lead us to change our beliefs based on evidence. Our model indicates that in general, this is the case.

Figure 5 shows how the probability that a false claim is canonized as true depends on the initial belief q_0 that the claim is true. Under most circumstances, the probability of canonization is relatively insensitive to initial belief. False canonization rates depend strongly on initial belief only when evidentiary standards are weak and experiments are highly prone to error (*Figure 5B*). In this case, belief is a random walk without a systematic tendency to increase or decrease with each published outcome, and thus the odds of canonization or rejection depend most strongly on the initial belief.

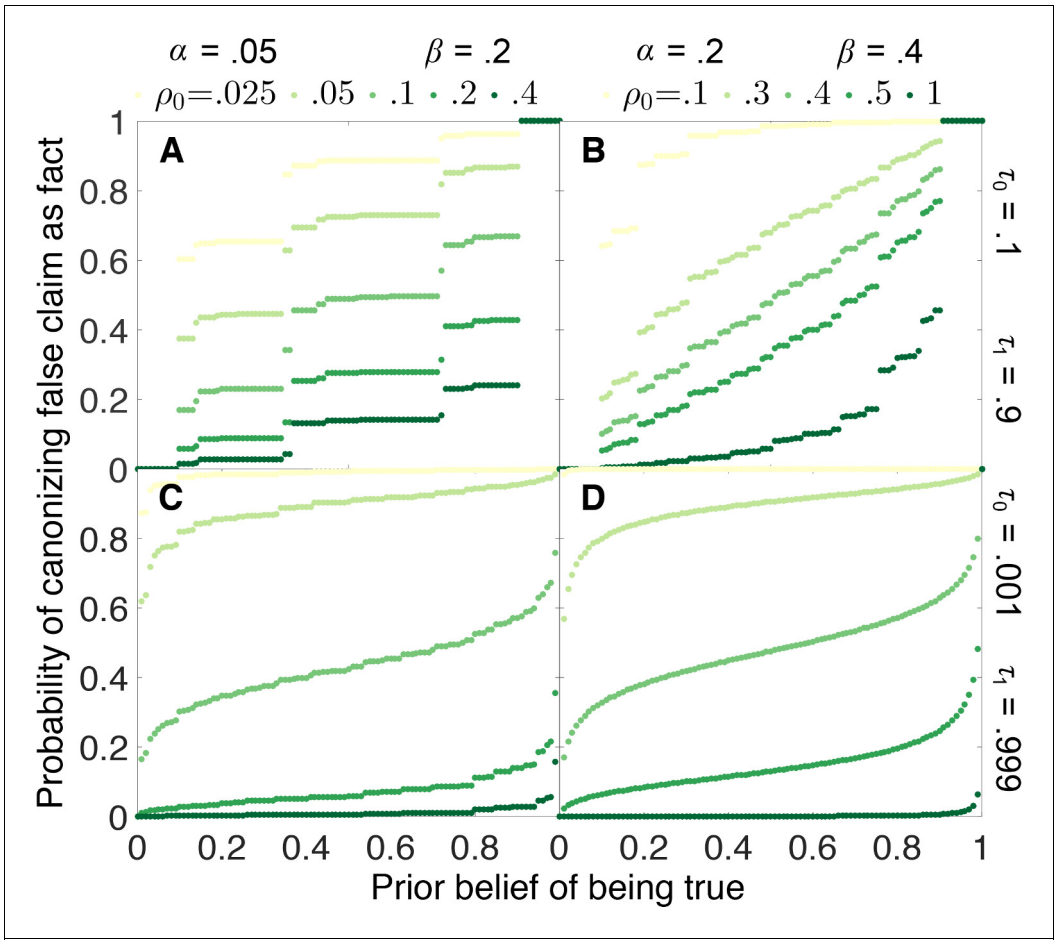

Figure 5. False canonization rates are relatively insensitive to initial belief, unless experimental tests are inaccurate and evidentiary standards are weak. Probability that a false claim is mistakenly canonized as a true fact vs. prior belief for various negative publication rates. Top row: weak evidentiary standards $\tau_0 = 0.1$ and $\tau_1 = 0.9$. Panel **A**: false positive rate $\alpha = 0.05$, false negative rate $\beta = 0.2$, and publication rate of negative results $\rho_0 = 0.025$ (light green), $0.05, 0.1, 0.2, 0.4$ (dark green). Panel **B**: $\alpha = 0.2$, $\beta = 0.4$, and $\rho_0 = 0.1$ (light green), $0.3, 0.4, 0.5, 1$ (dark green). Panels **C**–**D**: similar to panels **A**–**B**, with more demanding evidentiary standards $\tau_0 = 0.001$ and $\tau_1 = 0.999$.
DOI: 10.7554/eLife.21451.006

The step-function-like appearance of some of the results in ***Figure 5***, particularly ***Figure 5A***, is a real property of the curves in question and not a numerical artifact. The "steps" arise because, when evidentiary standards are weak, canonization or rejection often happens after a small number of experiments. Because the number of experiments must be integral, probabilities of false canonization can change abruptly when a small change in initial belief increases or decreases the number of experiments in the most likely path to canonization or rejection.

Stronger evidentiary standards do not reduce the need to publish negative outcomes

We have seen in the previous sections that the scientific process struggles to distinguish true from false claims when the rate of publishing negative results is low. We might hope that we could remedy this problem simply by demanding more evidence before accepting a claim as fact. Unfortunately, this is not only expensive in terms of time and effort—sometimes it will not even help.

Figure 6 illustrates the problem. In this figure, we see the probability of canonizing a false claim as a function of negative publication rate for three different evidentiary standards: $\tau_0 = 0.1$, $\tau_0 = 0.01$, and $\tau_0 = 0.001$. When the false positive rate α is relatively low (***Figure 6A***), increasing the evidentiary requirements reduces the chance of canonizing a false claim for negative publication rates above 0.1 or so, but below this threshold there is no advantage to requiring stronger evidence. When the false positive rate is higher (***Figure 6B***), the situation is even worse: for negative publication rates below 0.3 or so, increasing evidentiary requirements actually increases the chance of canonizing a false fact.

The limited benefits of strengthening evidentiary standards can be understood through the mathematical theory of random walks (***Norris, 1998***). In short, the thresholds of belief for canonizing or rejecting a claim are *absorbing boundaries* such that once belief attains either boundary, the walk terminates and beliefs will not change further. Increasing the evidentiary standards for canonization or rejection is tantamount to increasing the distance between these boundaries and the initial belief state. Basic results from the theory of random walks suggest that, as the distance between the initial state and the boundaries increases, the probability of encountering one boundary before the other depends increasingly strongly on the average change in the log odds of belief at each step (experiment), and less on the random

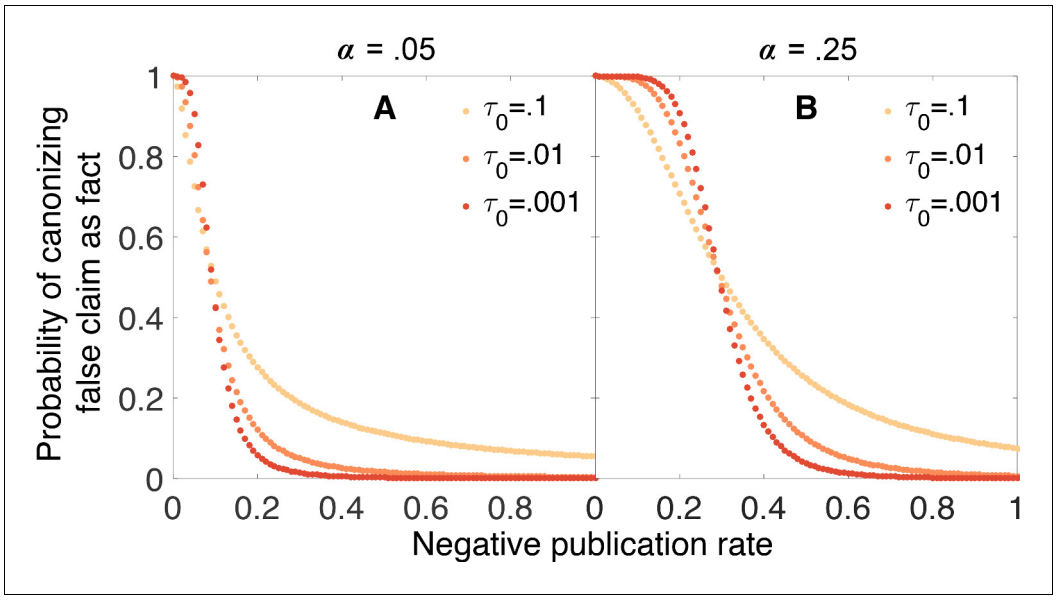

Figure 6. Strengthening evidentiary requirements does not necessarily decrease canonization of false facts. In panel **A**, the false positive rate is $\alpha = 0.05$, the false negative rate is $\beta = 0.2$, the original belief in the claim is $q_0 = 0.5$, and the evidentiary standards are symmetric $\tau_1 = 1 - \tau_0$. In panel **B**, the false positive rate is increased to $\alpha = 0.25$ while the other parameters remain unchanged. Particularly in this latter case, increasing evidentiary standards does not necessarily decrease the rate at which false claims are canonized as facts.
DOI: 10.7554/eLife.21451.007

fluctuations in belief that arise from the stochasticity of the walk. Thus, for exacting evidentiary standards, the probability of eventual canonization or rejection depends critically on the average change in the log odds of belief for each experiment. These are given by ***Equation 6*** for a true claim and ***Equation 7*** for a false one.

Figure 7 shows how the expected change in log odds of belief varies in response to changes in the publication rate of negative outcomes, for both false and true claims. Critically, for false claims, if too few negative outcomes are published, then on average each new publication will *increase* the belief that the claim is true, because there is a high probability this publication will report a positive result. Thus, paradoxically, scientific activity does not help sort true claims from false ones in this case, but instead promotes the erroneous canonization of false claims. The only remedy for this state of affairs is to publish enough negative outcomes that, on average, each published result moves belief in the "correct" direction, that is, towards canonization of true claims (a positive average change per experiment in log odds of belief) and rejection of false ones (a negative average change per experiment in log odds of belief).

Two additional points are in order here. First, for true claims, under most circumstances the expected change in the log odds of belief is positive (***Figure 7B***). That is, on average, scientific activity properly increases belief in true claims, and thus the risk of incorrectly rejecting a true claim is small (under reasonable evidentiary standards). Second, the observation that more exacting evidentiary standards can occasionally increase the chance of incorrectly canonizing a false claim is not much of an argument in favor of weaker evidentiary standards. In short, weaker standards cause canonization or rejection to depend more strongly on the happenstance of the first several published experiments. When scientific activity tends to increase belief in a false claim, weaker evidentiary standards appear beneficial because they increase the chance that a few initial published negatives will lead to rejection and bring a halt to further investigation. While this is a logical result of the model, it is somewhat tantamount to stating that, if scientific activity tends to increase belief in false claims, then the best option is to weaken the dependence on scientific evidence. More robust practices for rejecting false claims seem desirable.

P-hacking dramatically increases the probability of canonizing false claims

Our model has been based on the optimistic premise that the significance levels reported in each study accurately reflect the actual false

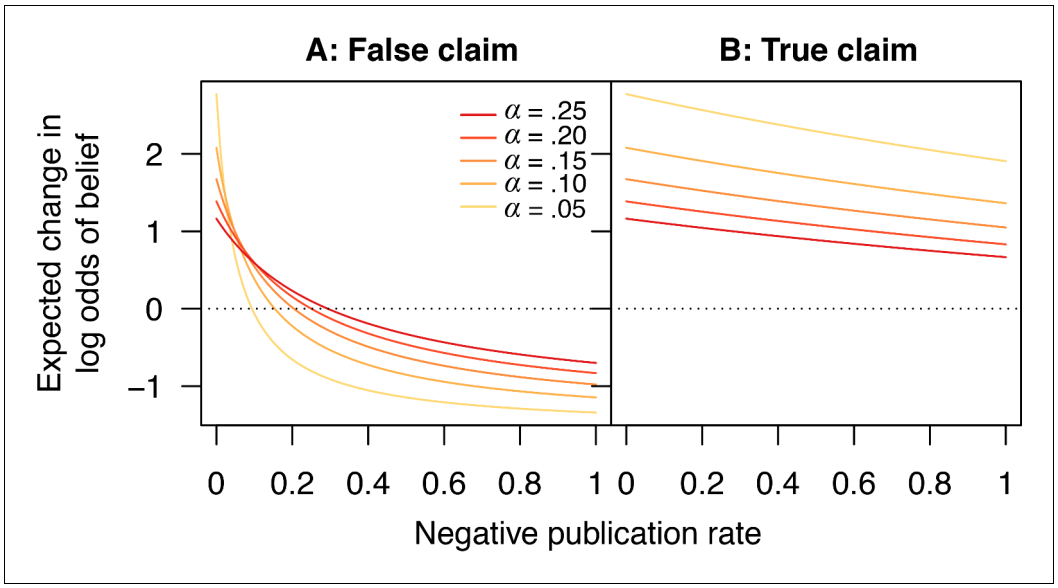

Figure 7. Scientific activity will tend to increase belief in false claims if too few negative outcomes are published. Expected change in log odds of belief vs. negative publication rate for (**A**) false and (**B**) true claims. Lines show false positive rates $\alpha = 0.05$ (yellow), $0.10, \ldots, 0.25$ (red). Other parameter values are false negative rate $\beta = 0.2$ and positive publication rate $\rho_1 = 1$.

DOI: 10.7554/eLife.21451.008

positive rates. This means that there is only a 5% chance that a false claim will yield a positive result at the $\alpha = 0.05$ level.

In practice, reported significance levels can be misleading. Questionable research practices of all sorts can result in higher-than-reported false positive rates; these include p-hacking (**Head et al., 2015**), outcome switching (**Le Noury et al., 2015**), unreported multiple comparisons (**Tannock, 1996**), data dredging (**Smith and Ebrahim, 2002**), HARKing—hypothesizing after the results are known (**Kerr, 1998**), data-dependent analysis (**Gelman and Loken, 2014**), and opportunistic stopping or continuation (**Pocock, 1987**). Insufficient validation of new technologies, or even software problems can also drive realized error rates far above what is expected given stated levels of statistical confidence (see e.g. ref. **Eklund et al., 2016**). Research groups may be positively disposed toward their prior hypotheses or reluctant to contradict the work of closely allied labs. Finally, industry-sponsored clinical trials often allow the sponsors some degree of control over whether results are published (**Kasenda et al., 2016**), resulting in an additional source of publication bias separate from the journal acceptance process.

To understand the consequences of these problems and practices, we can extend our model to distinguish the actual false positive rate α_{act} from the nominal false positive rate α_{nom} which is reported in the paper and used by readers to draw their inferences. We assume the actual false positive rate is always at least as large as the nominal rate, that is, $\alpha_{act} \geq \alpha_{nom}$. In this scenario, the probability that a false claim leads to a positive published outcome depends on the actual false positive rate, i.e.,

$$\omega_F = \frac{\alpha_{act}\rho_1}{\alpha_{act}\rho_1 + (1-\alpha_{act})\rho_0}.$$

However, the change in belief following a positive or negative published outcome respectively depends on the nominal false positive rate:

$$d_1 = \ln\left(\frac{1-\beta}{\alpha_{nom}}\right)$$
$$d_0 = \ln\left(\frac{\beta}{1-\alpha_{nom}}\right)$$

An inflated false positive rate makes it much more likely that false claims will be canonized as true facts (**Figure 8**). For example, suppose the false negative rate is $\beta = 0.2$ and the nominal false positive rate is $\alpha_{nom} = 0.05$, but the actual false positive rate is $\alpha_{act} = 0.25$. Even eliminating publication bias against negative outcomes (i.e., $\rho_0 = 1$) and using strong evidentiary standards does not eliminate the possibility that false claims will be canonized as facts under these circumstances (**Figure 8**). Less dramatic inflation of the false positive rate leaves open the possibility that true vs. false claims can be distinguished, but only if a higher percentage of negative outcomes is published.

Increasing negative publication rates as a claim approaches canonization greatly increases accuracy

Thus far we have told a troubling story. Without high probabilities of publication for negative results, the scientific process may perform poorly at distinguishing true claims from false ones. And there are plenty of reasons to suspect that negative results are not always be likely to be published.

However, authors, reviewers, and editors are all drawn to unexpected results that challenge or modify prevalent views—and for a claim widely believed to be true, a negative result

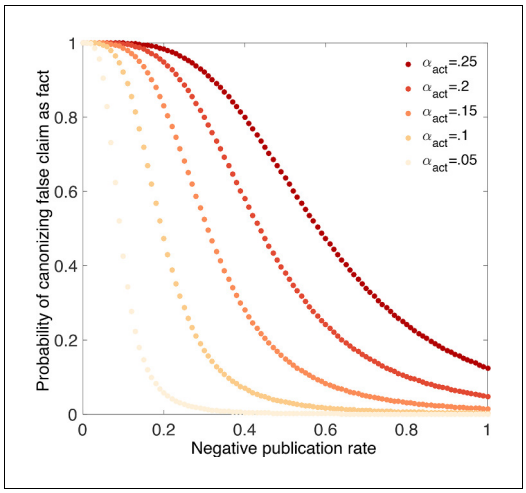

Figure 8. *p*-hacking dramatically increases the chances of canonizing false claims. Probability that a false claim is canonized as fact vs. fraction of negative outcomes. Throughout, all positive outcomes are published ($p_1 = 1$), and the nominal false positive rate is $\alpha_{nom} = 0.05$, the false negative rate is $\beta = 0.2$, and evidentiary standards are strong ($\tau_0 = 0.001$ and $\tau_1 = 0.999$). Curves show actual false positive rates $\alpha_{act} = 0.05$ (yellow), $0.10, \ldots, 0.25$ (red). Compared with **Figure 4C**, in which the nominal rates are equal to the actual rates, the probability of canonizing a false claim as fact is substantially higher.

DOI: 10.7554/eLife.21451.009

from a well-designed study is surprising. As a consequence, the probability of publishing a negative result may be higher for claims that are already considered likely to be true (*Silvertown and McConway, 1997*; *Ioannidis, 2005*)

In a simulation of point estimation by successive experimentation, de Winter and Happee considered an even more extreme situation in which it is only possible to publish results that contradict the prevailing wisdom (*de Winter and Happee, 2013*). They argue that this has efficiency benefits, but their results have been challenged persuasively by van Assen and colleagues (*van Assen et al., 2014*). In any event, such a publication strategy would not work in the framework we consider here, because a claim could neither be canonized nor rejected if each new published result were required to contradict the current beliefs of the community.

Some meta-analyses have revealed patterns consistent with this model (*Poulin, 2000*). For example, when the fluctuating asymmetry hypothesis was proposed in evolutionary ecology, the initial publications exclusively reported strong associations between symmetry and attractiveness or mating success. As time passed, however, an increasing fraction of the papers on this hypothesis reported negative findings with no association between these variables (*Simmons et al., 1999*). A likely interpretation is that initially journals were reluctant to publish results inconsistent with the hypothesis, but as it became better established, negative results came to be viewed as interesting and worthy of publication (*Simmons et al., 1999*; *Palmer, 2000*; *Jennions and Møller, 2002*).

To explore the consequences of this effect, we consider a model in which the probability of publishing a negative outcome increases linearly from a baseline value ρ_b when belief in the claim is weak, to a maximum value of $\rho_0 = 1$ when belief in the claim is strong. We assume that the probability of publishing a negative outcome is $\rho_0 = \rho_b + q(1 - \rho_b)$, where ρ_b is the baseline probability for publishing negative outcomes, and q is the current belief. As before, our agents are unaware of any publication bias in updating their own beliefs.

Figure 9 indicates that dynamic publication rates can markedly reduce (though not eliminate) the false canonization rate under many scenarios. In particular, *Figure 9* suggests that even if it is difficult to publish negative outcomes for claims already suspected to be false, we can still accurately sort true claims from false ones provided that negative outcomes are more readily published for claims nearing canonization. In practice, this mechanism may play an important role in preventing false results from becoming canonized more frequently.

Discussion

In the model of scientific inquiry that we have developed here, publication bias creates serious problems. While true claims will seldom be rejected, publication bias has the potential to cause many false claims to be mistakenly canonized as facts. This can be avoided only if a substantial fraction of negative results are published. But at present, publication bias appears to be strong, given that only a small fraction of the published scientific literature presents negative results. Presumably many negative results are going unreported. While this problem has been noted before (*Knight, 2003*), we do not know of any previous formal analysis of its consequences regarding the establishment of scientific facts.

Should scientists publish all of their results?

There is an active debate over whether science functions most effectively when researchers publish all of their results, or when they publish only a select subset of their findings *Nelson et al,, 2012*; *de Winter and Happee, 2013*; *van Assen et al., 2014*; *McElreath and Smaldino, 2015*). In our model, we observe no advantage to selective publication; in all cases treated we find that false canonization decreases monotonically with increasing publication of negative results. This seems logical enough. Decision theory affirms that in the absence of information costs, withholding information cannot on average improve performance in a decision problem such as the classification task we treat here (*Savage, 1954*; *Good, 1967*; *Ramsey, 1990*). As *Good (1967)* notes, a decision-maker "should use all the evidence *already* available, provided that the cost of doing so is negligible."

Nonetheless, several research groups have argued that selective publishing can be more efficient than publishing the results of all studies. Clearly they must be implicitly or explicitly imposing costs of some sort on the acts of publishing papers or reading them, and it can be instructive to see where these costs lie. One source of such costs comes simply from the increased volume of scientific literature that

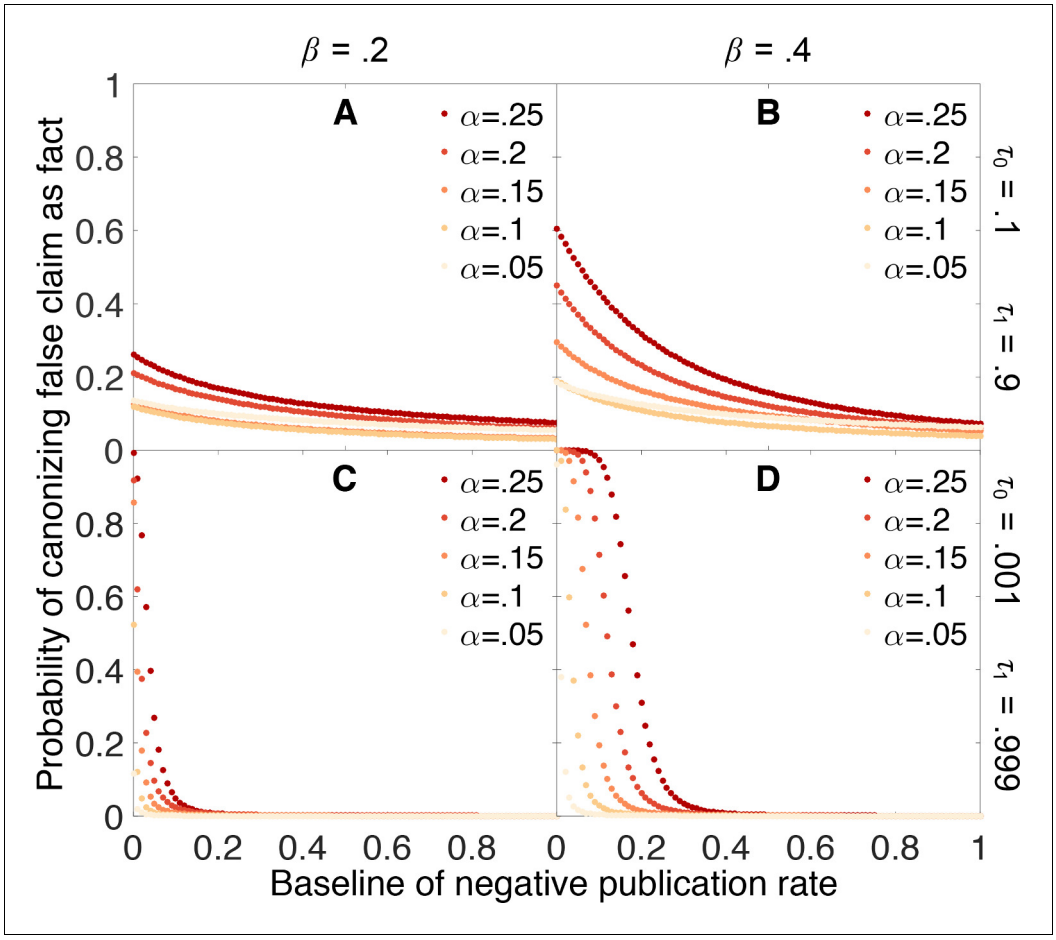

Figure 9. Publishing a larger fraction of negative outcomes as belief increases lessens the chances of canonizing false claims. Probability that a false claim is mistakenly canonized as a true fact vs. baseline probability of publishing a negative outcome. The baseline probability of publishing a negative outcome is the probability that prevails when belief in the claim is weak. The actual probability of publishing a negative outcome increases linearly from the baseline rate when belief is 0 to a value of 1 when belief is 1. All other parameters are the same as in ***Figure 4***.

DOI: 10.7554/eLife.21451.010

ensues when all results are published; this is sometimes known as the "cluttered office" effect (***Nelson et al., 2012***). If we envision that search costs increase with the volume of literature, for example, it may be beneficial not to publish everything.

Another possible cost is that of actually writing a paper and going through the publication process. If preparing a paper for publication is costly relative to doing the experiments which would be reported, it may be advantageous to publish only a subset of all experimental results. This is the argument that de Winter and Happee make when, in a mathematical model of point estimation, they find that selective publication minimizes the variance *given the number of publications* (as opposed to the number of experiments conducted). Note, however, that they assume a model of science in which experiments are only published when they contradict the prevailing wisdom—and that their results have been roundly challenged in a followup analysis (***van Assen et al., 2014***).

McElreath and Smaldino (2015) analyzed a model that is more similar to ours in structure. As we do, they consider repeated tests of binary-valued hypotheses. But rather than focusing on a single claim at a time, they model the progress of a group of scientists testing a suite of hypotheses. Based on this model, McElreath and Smaldino conclude that there can be advantages to selective publication under certain conditions.

While selective publication certainly can ameliorate the cluttered office problem—observed in their model as the flocking of researchers to questions already shown likely to be false—we are skeptical about the other advantages to selective publication. McElreath and Smaldino's model and results appear to rely in part on their assumption that "the only information relevant for judging the truth of a hypothesis is its *tally*, the difference between the number of published positive and the number of published negative findings" (p. 3).

As a mathematical claim, this is incorrect. Presumably the claim is instead intended to be a tactical assumption that serves to simplify the analysis. But this assumption is severely limiting. The tally is often an inadequate summary of the evidence in favor of a hypothesis. One can see the problem with looking only at the tally by considering a simple example in which false positive rates are very low, false negative rates are high, and all studies are published. There is mild evidence that a hypothesis is false if no positive studies and one negative study have been published, but there is strong evidence that the hypothesis is true if three positive and four negative studies have been published. Yet both situations share the same tally: -1. The same problem arises when publication bias causes positive and negative findings to be published at different rates.

If one is forced to use only the tally to make decisions, an agent can sometimes make better inferences by throwing away some of the data (i.e., under selective publication). For example, when false negatives are common it may be beneficial to suppress some fraction of the negative results lest they swamp any signal from true positive findings. This is not the case when the agent has access to complete information about the number of positive and the number of negative results published. As a result, it is unclear whether most of McElreath and Smaldino's arguments in favor of selective publication are relevant to optimal scientific inference, or whether they are consequences of the assumption that agents draw their inferences based on the tally alone.

What do we do about the problem of publication bias?

Several studies have indicated that much of the publication bias observed in science can be attributed to authors not writing up null results, rather than journals rejecting null results (***Dickersin et al., 1992***; ***Olson et al., 2002***; ***Franco et al., 2014***). This does not necessarily exonerate the journals; authors may be responding to incentives that the journals have put in place (***Song et al., 2000***). Authors may be motivated by other reputational factors as well. It would be a very unusual job talk, promotion seminar, or grant application that was based primarily upon negative findings.

So what can we as a scientific community do? How can we avoid canonizing too many false claims, so that we can be confident in the veracity of most scientific facts? In this paper, we have shown that strengthening evidentiary standards does not necessarily help. In the presence of strong publication bias, false claims become canonized as fact not so much because of a few misleading chance results, but rather because on average, misleading results are more likely to be published than correct ones.

Fortunately, this problem may be ameliorated by several current aspects of the publication process. In this paper, we have modeled claims that have only one way of generating "positive" results. For many scientific claims, e.g. those like our Dicer example that propose particular mechanisms, this may be appropriate. In other cases, however, results may be continuous: not only do we care whether variables X and Y are correlated, but also we want to know about the strength of the correlation, for example. This does not make the problem go away, if stronger or highly significant correlations are seen as more worthy of publication than weaker or non-significant correlations. However, one advantage of framing experimental results as continuous-valued instead of binary is that there may be multiple opposing directions in which a result could be considered positive. For example, the expression of two genes could be correlated, uncorrelated, or anticorrelated. Both correlation and anticorrelation might be seen as positive results, whereas the null result of no correlation could be subject to publication bias. But suppose there is truly no effect: what does publication bias do in this case? We would expect to see false positives in both directions. Meta-analysis would readily pick up the lack of a consistent direction of the effect, and (if the authors avoid mistakenly inferring population heterogeneity) it is unlikely that correlations in either direction would be falsely canonized as fact.

Our model assumes that research continues until each claim is either rejected or canonized as fact. In practice, researchers can and do lose interest in certain claims. False claims might generate more conflicting results, or take longer to reach one of the evidentiary thresholds; either

mechanism could lead researchers to move on to other problems and leave the claim as unresolved. If this is the case, we might expect that instead of being rejected or canonized as fact, many false claims might simply be abandoned.

Another possible difference between the model and the real world is that we model the evidentiary standards as symmetric, but in practice it may require less certainty to discard a claim as false than it requires to accept the same claim as fact. In this case, the probability of rejecting false claims would be higher than predicted in our model—possibly with only a very small increase in the probability of rejecting true claims.

The scientific community could also actively respond to the problem of canonizing false claims. One of the most direct ways would be to invest more heavily in the publication of negative results. A number of new journals or collections within journals have been established to specialize in publishing negative results. These include Elsevier's *New Negatives in Plant Science*, *PLOS One's* Positively Negative collection, Biomed Central's *Journal of Negative Results in Biomedicine*, and many others (**Editors, 2016**). Alternatively, peer reviewed publication may be unnecessary; simply publishing negative results on preprint archives such as the arXiv, bioRxiv, and SocArXiv may make these results sufficiently visible. In either case, we face an incentive problem: if researchers accrue scant credit or reward for their negative findings, there is little reason for them to invest the substantial time needed in taking a negative result from a bench-top disappointment to a formal publication.

Another possibility—which may already be in play—involves shifting probabilities of publishing negative results. We have shown that if negative results become easier to publish as a claim becomes better established, this can greatly reduce the probability of canonizing false claims. One possibility is that negative results may become easier to publish as they become more surprising to the community, i.e., as researchers become increasingly convinced that a claim is true. Referees and journal editors could make an active effort to value papers of this sort. At present, however, our experience suggests that negative results or even corrections of blatant errors in previous publications rarely land in journals of equal prestige to those that published the original positive studies (**Matosin et al., 2014**).

A final saving grace is that even after false claims are established as facts, science can still self-correct. In this paper, we have assumed for simplicity that claims are independent propositions, but in practice claims are entangled in a web of logical interrelations. When a false claim is canonized as fact, inconsistencies between it and other facts soon begin to accumulate until the field is forced to reevaluate the conflicting facts. Results that resolve these conflicts by disproving accepted facts then take on a special significance and suffer little of the stigma placed upon negative results. Until the scientific community finds more ways to deal with publication bias, this may be an essential corrective to a process that sometimes loses its way.

We conclude with a note on what this work tells us about the value of science as a means of comprehending the natural world. Science denialists on both ends of the ideological spectrum might be tempted to invoke our findings as justification for their world-views. This would be a mistake. The facts that science denialists target are almost always very different from the types of facts we are modeling. We are modeling small-scale facts of modest import, the kind that would be established based on one or two dozen studies and then considered settled. The reality of anthropogenic climate change, the lack of connection between vaccination and autism, or the causative role of smoking in cancer are very different. Facts of this sort have enormous practical importance; they are supported by massive volumes of research; and they have been established despite well-funded groups with powerful incentives to expose any evidence that might give cause for skepticism. The process by which false claims can become canonized as fact in our model simply would not operate under these circumstances.

Of all the institutions and methods that humankind have developed to make sense of our universe, science has proven unparalleled in its power to generate useful models of physical phenomena. Nothing that we have written here changes this. The point of asking questions such as those in the present paper is not to de-legitimize science, but rather to improve the accuracy and efficiency of scientific inquiry.

Acknowledgements

The authors thank Jacob G Foster, Frazer Meacham and Kevin Zollman for helpful comments and suggestions on the manuscript. This work was supported by the Danish National Research Foundation and by a generous grant to the Metaknowledge Network from the John Templeton Foundation. KG thanks the University of

Washington Department of Biology for sabbatical support.

Silas Boye Nissen is in the Niels Bohr Institute, University of Copenhagen, Copenhagen, Denmark
Tali Magidson is in the Department of Computer Science, University of Washington, Seattle, United States
Kevin Gross is in the Department of Statistics, North Carolina State University, Raleigh, United States
Carl T Bergstrom is in the Department of Biology, University of Washington, Seattle, United States
http://orcid.org/0000-0002-2070-385X

Author contributions: SBN, TM, KG, CTB, Conception and design, Acquisition of data, Analysis and interpretation of data, Drafting or revising the article

Competing interests: CTB: Reviewing editor, *eLife*. The other authors declare that no competing interests exist.

Received 12 September 2016
Accepted 28 November 2016
Published 20 December 2016

Funding

Funder	Author
John Templeton Foundation	Carl T Bergstrom
Danish National Research Foundation	Silas Boye Nissen

The funders had no role in study design, data collection and interpretation, or the decision to submit the work for publication.

References

Arbesman S. 2012. *The Half-Life of Facts: Why Everything We Know Has an Expiration Date*. Penguin.

Baker M. 2016. 1,500 scientists lift the lid on reproducibility. *Nature* **533**:452–454. doi: 10.1038/533452a, PMID: 27225100

Begg CB, Berlin JA. 1988. Publication bias: a problem in interpreting medical data. *Journal of the Royal Statistical Society. Series A* **151**:419–463. doi: 10.2307/2982993

Begley CG, Ellis LM. 2012. Drug development: Raise standards for preclinical cancer research. *Nature* **483**:531–533. doi: 10.1038/483531a, PMID: 22460880

Bernstein E, Caudy AA, Hammond SM, Hannon GJ. 2001. Role for a bidentate ribonuclease in the initiation step of RNA interference. *Nature* **409**:363–366. doi: 10.1038/35053110, PMID: 11201747

Camerer CF, Dreber A, Forsell E, Ho TH, Huber J, Johannesson M, Kirchler M, Almenberg J, Altmejd A, Chan T, Heikensten E, Holzmeister F, Imai T, Isaksson S, Nave G, Pfeiffer T, Razen M, Wu H. 2016. Evaluating replicability of laboratory experiments in economics. *Science* **351**:1433–1436. doi: 10.1126/science.aaf0918, PMID: 26940865

Chan AW, Altman DG. 2005. Identifying outcome reporting bias in randomised trials on PubMed: review of publications and survey of authors. *BMJ* **330**:753. doi: 10.1136/bmj.38356.424606.8F, PMID: 15681569

Chang AC, Li P. 2015. Is economics research replicable? sixty published papers from thirteen journals say "usually Not. *Finance and Economics Discussion Series* **083**:1–26. doi: 10.17016/feds.2015.083

Csada RD, James PC, Espie RHM. 1996. The "file drawer problem" of non-significant results: does it apply to biological research? *Oikos* **76**:591–593. doi: 10.2307/3546355

de Winter J, Happee R. 2013. Why selective publication of statistically significant results can be effective. *PLoS One* **8**:e66463. doi: 10.1371/journal.pone.0066463

Dickersin K, Min YI, Meinert CL. 1992. Factors influencing publication of research results. follow-up of applications submitted to two institutional review boards. *JAMA* **267**:374–378. doi: 10.1001/jama.267.3.374, PMID: 1727960

Dickersin K. 1990. The existence of publication bias and risk factors for its occurrence. *JAMA* **263**:1385. doi: 10.1001/jama.1990.03440100097014, PMID: 2406472

Easterbrook PJ, Gopalan R, Berlin JA, Matthews DR. 1991. Publication bias in clinical research. *The Lancet* **337**:867–872. doi: 10.1016/0140-6736(91)90201-Y

Ebrahim S, Sohani ZN, Montoya L, Agarwal A, Thorlund K, Mills EJ, Ioannidis JP. 2014. Reanalyses of randomized clinical trial data. *JAMA* **312**:1024–1032. doi: 10.1001/jama.2014.9646, PMID: 25203082

Editors. 2016. Go forth and replicate!. *Nature* **536**:373. doi: 10.1038/536373a, PMID: 27558028

Eklund A, Nichols TE, Knutsson H. 2016. Cluster failure: Why fMRI inferences for spatial extent have inflated false-positive rates. *PNAS* **113**:7900–7905. doi: 10.1073/pnas.1602413113

Errington TM, Iorns E, Gunn W, Tan FE, Lomax J, Nosek BA. 2014. An open investigation of the reproducibility of cancer biology research. *eLife* **3**:e04333. doi: 10.7554/eLife.04333

Fanelli D. 2012. Negative results are disappearing from most disciplines and countries. *Scientometrics* **90**:891–904. doi: 10.1007/s11192-011-0494-7

Franco A, Malhotra N, Simonovits G. 2014. Social science. publication bias in the social sciences: unlocking the file drawer. *Science* **345**:1502–1505. doi: 10.1126/science.1255484, PMID: 25170047

Gelman A, Loken E. 2014. The statistical crisis in science. *American Scientist* **102**:460–465. doi: 10.1511/2014.111.460

Good IJ. 1967. On the principle of total evidence. *The British Journal for the Philosophy of Science* **17**:319–321. doi: 10.1093/bjps/17.4.319

Higginson AD, Munafò MR. 2016. Current incentives for scientists lead to underpowered studies with erroneous conclusions. *PLoS Biology* **14**:e2000995. doi: 10.1371/journal.pbio.2000995, PMID: 27832072

Ioannidis JP. 2005. Why most published research findings are false. *PLoS Medicine* **2**:e124. doi: 10.1371/journal.pmed.0020124, PMID: 16060722

Ioannidis JP, Trikalinos TA. 2005. Early extreme contradictory estimates may appear in published research: the Proteus phenomenon in molecular genetics research and randomized trials. *Journal of Clinical Epidemiology* **58**:543–549. doi: 10.1016/j.jclinepi.2004.10.019, PMID: 15878467

Jaskiewicz L, Filipowicz W. 2008. Role of dicer in posttranscriptional RNA silencing. *Current Topics in Microbiology and Immunology* **320**:77–97. doi: 10.1007/978-3-540-75157-1_4 , PMID: 18268840

Jennions MD, Moller AP. 2002. Relationships fade with time: a meta-analysis of temporal trends in publication in ecology and evolution. *Proceedings of the Royal Society B: Biological Sciences* **269**:43–48. doi: 10.1098/rspb.2001.1832

Kasenda B, von Elm E, You JJ, Blümle A, Tomonaga Y, Saccilotto R, Amstutz A, Bengough T, Meerpohl JJ, Stegert M, Olu KK, Tikkinen KA, Neumann I, Carrasco-Labra A, Faulhaber M, Mulla SM, Mertz D, Akl EA, Bassler D, Busse JW, et al. 2016. Agreements between industry and academia on publication rights: A retrospective study of protocols and publications of randomized clinical trials. *PLoS Medicine* **13**:e1002046. doi: 10.1371/journal.pmed.1002046, PMID: 27352244

Kerr NL. 1998. HARKing: hypothesizing after the results are known. *Personality and Social Psychology Review* **2**:196–217. doi: 10.1207/s15327957pspr0203_4, PMID: 15647155

Knight J. 2003. Negative results: Null and void. *Nature* **422**:554–555. doi: 10.1038/422554a, PMID: 12686968

Latour B. 1987. *Science in Action: How to Follow Scientists and Engineers Through Society*. Harvard University Press.

Matosin N, Frank E, Engel M, Lum JS, Newell KA. 2014. Negativity towards negative results: a discussion of the disconnect between scientific worth and scientific culture. *Disease Models & Mechanisms* **7**:171–173. doi: 10.1242/dmm.015123, PMID: 24713271

McElreath R, Smaldino PE. 2015. Replication, communication, and the population dynamics of scientific discovery. *PLoS One* **10**:e0136088. doi: 10.1371/journal.pone.0136088, PMID: 26308448

Head ML, Holman L, Lanfear R, Kahn AT, Jennions MD, Megan L. 2015. The extent and consequences of p-hacking in science. *PLoS Biology* **13**:e1002106. doi: 10.1371/journal.pbio.1002106, PMID: 25768323

Nelson LD, simmons JP, simonsohn U. 2012. let's publish fewer papers. *Psychological Inquiry* **233**:291–293 . doi: 10.1080/1047840x.2012.705245

Newcombe RG. 1987. Towards a reduction in publication bias. *BMJ* **295**:656–659. doi: 10.1136/bmj.295.6599.656

Norris JR. 1998. *Markov Chains*. Cambridge University Press.

Le Noury J, Nardo JM, Healy D, Jureidini J, Raven M, Tufanaru C, Abi-Jaoude E. 2015. Restoring study 329: efficacy and harms of paroxetine and imipramine in treatment of major depression in adolescence. *BMJ* **351**:h4320. doi: 10.1136/bmj.h4320, PMID: 26376805

Olson CM, Rennie D, Cook D, Dickersin K, Flanagin A, Hogan JW, Zhu Q, Reiling J, Pace B. 2002. Publication bias in editorial decision making. *JAMA* **287**:2825–2828. doi: 10.1001/jama.287.21.2825, PMID: 12038924

Open Science Collaboration. 2015. PSYCHOLOGY. Estimating the reproducibility of psychological science. *Science* **349**:aac4716. doi: 10.1126/science.aac4716, PMID: 26315443

Palmer AR. 2000. Quasi-replication and the contract of error: lessons from sex ratios, heritabilities and fluctuating asymmetry. *Annual Review of Ecology and Systematics* **31**:441–480. doi: 10.1146/annurev.ecolsys.31.1.441

Pocock SJ, Hughes MD, Lee RJ. 1987. Statistical problems in the reporting of clinical trials. *New England Journal of Medicine* **317**:426–432. doi: 10.1056/NEJM198708133170706

Poulin R. 2000. Manipulation of host behaviour by parasites: a weakening paradigm? *Proceedings of the Royal Society B: Biological Sciences* **267**:787–792. doi: 10.1098/rspb.2000.1072

Ramsey FP. 1990. Weight or the value of knowledge. *The British Journal for the Philosophy of Science* **41**:1–4. doi: 10.1093/bjps/41.1.1

Ravetz JR. 1971. Scientific knowledge and its social problems. *British Society for the Philosophy of Science* **7**:72. doi: 10.1017/s0007087400012875

Rosenthal R. 1979. The file drawer problem and tolerance for null results. *Psychological Bulletin* **86**:638–641. doi: 10.1037/0033-2909.86.3.638

Rzhetsky A, Iossifov I, Loh JM, White KP. 2006. Microparadigms: Chains of collective reasoning in publications about molecular interactions. *PNAS* **103**:4940–4945. doi: 10.1073/pnas.0600591103

Savage LJ. 1954. *The Foundations of Statistics*. John Wiley & Sons.

Silvertown J, McConway KJ. 1997. Does "publication bias" lead to biased science? *Oikos* **79**:167–168. doi: 10.2307/3546101

Simmons LW, Tomkins JL, Kotiaho JS, Hunt J. 1999. Fluctuating paradigm. *Proceedings of the Royal Society B: Biological Sciences* **266**:593–595. doi: 10.1098/rspb.1999.0677

Smith GD, Ebrahim S. 2002. Data dredging, bias, or confounding. *BMJ* **325**:1437–1438. doi: 10.1136/bmj.325.7378.1437, PMID: 12493654

Song F, Eastwood AJ, Gilbody S, Duley L, Sutton AJ. 2000. Publication and related biases. *Health Technology Assessment* **4**:1–115. PMID: 10932019

Sterling TD. 1959. Publication decisions and their possible effects on inferences drawn from tests of significance –Or vice versa. *Journal of the American Statistical Association* **54**:30–34. doi: 10.1080/01621459.1959.10501497

Tannock IF. 1996. False-positive results in clinical trials: Multiple significance tests and the problem of unreported comparisons. *Journal of the National Cancer Institute* **88**:206–207. doi: 10.1093/jnci/88.3-4.206

Turner EH, Matthews AM, Linardatos E, Tell RA, Rosenthal R. 2008. Selective publication of antidepressant trials and its influence on apparent efficacy. *New England Journal of Medicine* **358**:252–260. doi: 10.1056/NEJMsa065779, PMID: 18199864

van Assen MA, van Aert RC, Nuijten MB, Wicherts JM. 2014. Why publishing everything is more

effective than selective publishing of statistically significant results. *PLoS One* **9**:e84896. doi: 10.1371/journal.pone.0084896, PMID: 24465448