Publish or Perish: Simulating the Impact of Publication Policies on Science

Marina Mancoridis[1] **(mm96@princeton.edu)**
Theodore R. Sumers[3] **(ted@anthropic.com)**
Thomas L. Griffiths[1,2] **(tomg@princeton.edu)**
[1]Department of Computer Science, Princeton University
[2]Department of Psychology, Princeton University
[3]Anthropic

Abstract

Science can be viewed as a collective, epistemic endeavor. However, a variety of factors — such as the publish-or-perish culture, institutional incentives, and publishers who favor novel and positive findings — may challenge the ability of science to accurately aggregate information about the world. Evidence of the shortcomings in the current structure of science can be seen in the replication crisis that faces psychology and other disciplines. We analyze scientific publishing through the lens of cultural evolution, framing the scientific process as a multi-generational interplay between scientists and publishers in a multi-armed bandit setting. We examine the dynamics of this model through simulations, exploring the effect that different publication policies have on the accuracy of the published scientific record. Our findings highlight the need for replications and caution against behaviors that prioritize factors uncorrelated with result accuracy.

Keywords: replication crisis; metascience; cognitive modeling; Bayesian analysis; multi-armed bandit problem

Introduction

Bad science is, in part, the product of incentive misalignment (Smaldino & McElreath, 2016). Scientists, especially young ones, face a growing expectation to publish frequently to advance their careers (Niles, Schimanski, McKiernan, & Alperin, 2020). In fact, the number of publications authored by newly-hired biologists has nearly *doubled* in the past decade alone (Brischoux & Angelier, 2015). The use of publication frequency as a measure of productivity can encourage shallow research (Grimes, Bauch, & Ioannidis, 2018). Publication policies that reward novel and surprising findings exacerbate the problem. Scientists are incentivized to publish and publishers are incentivized to break ground; no one is directly incentivized to be correct. This incentive misalignment challenges the validity of a system whose goal it is to collectively gather accurate knowledge of the world.

We see evidence of science's shortcomings in the replication crisis that faces psychology and other disciplines (Wiggins & Christopherson, 2019). Experts have estimated that up to half of scientific literature could be incorrect (Simmons, Nelson, & Simonsohn, 2011; Ioannidis, 2005; Pashler & Harris, 2012; McElreath & Smaldino, 2015). The Reproducibility Project, in which 270 psychologists analyzed 100 influential papers in the field, finds that only one third of the studies could be reproduced with statistically significant results (Aarts et al., 2015). Even among those that could be replicated, most show effects less than half the size indicated by the original reports. Some researchers argue that the magnitude of the replication crisis is exaggerated by limited, single-replication studies or that it can be explained by the field's high rate of false hypotheses (Maxwell, Lau, & Howard, 2015; Shrout & Rodgers, 2018; Bird, 2021). We consider another explanation: could the crisis be a natural consequence of the incentives and information aggregation processes involved in science?

It is difficult to offer a universal explanation for the emergence of false publications because of the heterogeneity of scientific research. Replication efforts can reveal the reason for a false publication about a *single* study, but lack transferability to other studies and are often limited by scope (Stevens, 2017). For an explanation that generalizes across experiments, one must consider the science of science itself, or metascience. Metascience researchers have developed innovations in scientific methodology and publishing policies that they argue have the potential to resolve the replication crisis (Schooler, 2014; Peterson & Panofsky, 2023). In this paper, we introduce a *meta*-metascience approach that explores the efficacy of different interventions of this type.

Our approach models science as the multi-generational interplay between scientists and publishers in a multi-armed bandit setting. We distill one generation of science into three stages: researchers gather evidence about the payoffs of different arms, they aggregate and report their findings, and journals choose reports to publish. Between generations, the public scientific record is updated. In our simulations, we search across publication policies to determine which improve the accuracy of published findings. In addition, we explore how hypothesis success rates affect the development of scientific knowledge by systematically varying the underlying arm probabilities in our bandit process under a fixed publication policy.

There are two primary branches of insights that emerge from this project. The first is our development of a novel, theoretical model of science. We build on the broader cultural evolution literature (Birch & Heyes, 2021) by offering a computational model of science as a collective, information-gathering process. Our mathematical model captures various elements of human behavior —such as decision-making over uncertain actions and reliance on previously aggregated knowledge —to simulate how science advances over time. The second branch of insights emerges from our findings themselves. We show that more important than prioritizing

In L. K. Samuelson, S. L. Frank, M. Toneva, A. Mackey, & E. Hazeltine (Eds.), *Proceedings of the 46th Annual Conference of the Cognitive Science Society.*

results by the amount of data that supports them is to *avoid* prioritizing results by factors that are uncorrelated with accuracy. Specifically, we show that favoring surprising or positive results leads to less efficient scientific progress. Moreover, we show that the data gathering and publishing structures in science can lead to inaccurate publications without sufficient multi-generational replications.

Background

Selecting Papers for Publication

The landscape of academic publishing is shaped by reviewer biases. When selecting papers for publication, reviewers consider factors like experiment design, scientific significance, and writing clarity (NeurIPS, 2023; Fiske, 2004). However, they have also been shown to exhibit selection biases towards positive and surprising results (Coursol & Wagner, 1986). The study of selection bias in scientific journal publication has broadened to the term *publication bias*, which considers both the direction and statistical significance of the results (Rothstein, Sutton, & Borenstein, 2005).

Academic publishing is also shaped by the incentive structures facing scientists. Colloquially, phrases like '*publish or perish*' and '*funding or famine*' capture the pressure to produce that faces individual scientists (Van Dalen, 2021). This pressure has grown in recent years, creating a culture that increasingly prioritizes output over quality. Scientists have expressed concern about the growing use of quantitative metrics when evaluating productivity. Scientific behaviors that have been shown to emerge as a result include p-hacking, salami slicing, and selective reporting (Head, Holman, Lanfear, Kahn, & Jennions, 2015; Aschwanden & King, 2015; Grimes et al., 2018; Beaufils & Karlsson, 2013).

The Cultural Evolution of Scientific Knowledge

The cultural evolution of knowledge is the process by which collective beliefs are accumulated by individuals across generations (Birch & Heyes, 2021). Relying on findings made by independent agents across generations to further collective knowledge is in no way foreign to the human experience. Social interaction and knowledge transmission are key features of human intelligence, as any single individual is limited by time, cognitive resources, and effort (Hardy, Krafft, Thompson, & Griffiths, 2022).

Researchers have used mathematical and computational models to explore how the structure of science influences its outcomes. Basic questions about how scientists choose the problems they study can be framed from the perspective of decision theory and game theory (Kitcher, 1990). This approach makes it possible to explore how modifying the incentive structure of science, such as how credit is awarded for discoveries, should change the behavior of scientists (Strevens, 2003, 2017, 2020).

This kind of approach can be extended to model science as a process of cultural evolution. One example of such a model, developed by Smaldino and McElreath (2016), explores the ability of replications and incentives to curtail poor science. Their model adds and removes labs over time according to the accumulated payoffs of each lab's publications. In a separate body of work, they develop a model in which researchers update the tally of positive and negative findings for different hypotheses over generations of scientists. They vary hypothesis base rates to highlight the importance of quality theorizing (McElreath & Smaldino, 2015). A final example of model that frames science through the lens of cultural evolution was developed by Grimes et al. (2018). Their model compares cohorts of scientists with different levels of diligence in a mathematical model that varies funding resources, publication metrics, and more. These projects and others define computational models of science. However, none have investigated the impact of different publication policies on science.

Simulating Scientific Publishing

Our goal is to explore how scientific publication policies affect the accumulation of knowledge. To pursue this goal, we need a way of formalizing the scientific process that allows us to search over this policy space. We do this by framing science as a *distributed, multi-armed bandit problem*. The multi-armed bandit problem originates from the analogy of a gambler who pulls different arms on a slot machine (Robbins, 1952). The goal is to maximize reward, which requires discovering the payoff distribution of a set of actions (the different arms) through experimentation. In a *Bernoulli* multi-armed bandit problem, an agent who selects arm a receives a payoff of 1 with a probability θ_a and a payoff of 0 with a probability of $1 - \theta_a$. A distributed multi-armed bandit problem involves multiple agents who take independent actions to collectively learn the payoff distribution (Zhu & Liu, 2023).

How can the multi-armed bandit problem be viewed as a simplistic characterization of the problem that science solves? Multi-armed bandits and the scientific process both involve adaptive information search. In science, independent researchers who are guided by previous knowledge work simultaneously to uncover information about the world. In our multi-armed bandit framework, we represent the unknown world state by the unknown payoff distribution of arms. We represent findings by information about this payoff distribution. Our *distributed* framework captures the idea that scientists can work in parallel to collectively discover information. Finally, we capture the idea that scientists are guided by existing knowledge by giving agents a prior probability distribution over the success probabilities of the arms.

Just as simulations have shed light on the dynamics of cultural evolution (Lewis & Laland, 2012), we provide a simulation that models the process of collecting, reporting, publishing, and updating scientific data.

We model data gathering as taking *samples* from arms. In our Bernoulli setting, the result of a sample is either a success or a failure. We encode scientists as virtual agents who independently take a limited number of samples from arms

of their choice. After sampling, each scientist chooses *one* arm and publishes a single report containing the data sampled from that arm.

After data gathering and reporting, the publishing policy chooses a subset of the submitted reports to publish. To each report, the policy ascribes a utility value that is a function of three features: (1) the amount of data associated with the report, (2) the surprisingness of the report given the published scientific record, and (3) a publication bias that favors positive findings (probability of success $\geq \frac{1}{2}$). The publishing policy employs a softmax function to determine which reports to publish, prioritizing those with higher utility values. The scientific record consists of the number of successes and failures that have been published for each arm. It is available to all agents and is updated after each generation.

Our first analysis varies the utility function that publishers ascribe to reports. To explore the effects of different publication policies, we search across weights of the three features used in the utility function. This allows us to identify the publishing behaviors that lead to accurate scientific records. Our second analysis explores the impact of arm success probabilities on our simulation dynamics, asking the question of how variations in the base rate of hypotheses affect the ability of science to converge on accurate publications.

Methods

Figure 1 shows a schematic representation of our simulation framework.

Selecting Experiments

Each scientist has access to the scientific record, which consists of the number of successes and failures *published* for each arm. The record represents the scientist's prior knowledge over the success rate of different hypotheses. In the sampling stage, scientists take a limited number of samples from a set of arms. Each scientist takes the same number of samples over the same set of arms in each generation. The scientific record is initialized with one success and one failure published for each arm.

Our simulation models scientists as rational Bayesian actors who sample from the arm that maximizes their expected value of information (EVI). We define EVI as follows. Let p_a be the prior probability that arm a returns a success; this is calculated directly from the published scientific record. Let $u_s(x)$ be the utility that the scientist receives from drawing a value of $x \in \{0,1\}$ from arm a. Under this regime, $EVI = p_a \cdot u_s(1) + (1-p_a) \cdot u_s(0)$. Scientists draw from the arm with the highest EVI value, breaking ties randomly.

The following question remains: how is $u_s(x)$ defined? Prior work suggests that both scientists and publishers ascribe more utility to surprising findings (Vinkers, Tijdink, & Otte, 2015; Grimes et al., 2018). To capture this notion, we define $u_s(x)$ as the KL divergence between the published scientific record and the record assuming they publish x and their observed findings so far (Kullback & Leibler, 1951). If q_a is the posterior mean of θ_a after observing x and p_a is the prior mean (the predicted probability the arm delivers a payoff), we define the KL divergence as follows:

$$KL_a = p_a \log \frac{p_a}{q_a} + (1-p_a) \log \frac{1-p_a}{1-q_a} \quad (1)$$

In this formulation, the prior distribution is the arm success probabilities from the existing record. The posterior distribution is the arm success probabilities assuming that the scientist publishes the existing record *and* their observed findings. Note the direction of the KL divergence equation. We sample from the posterior distribution and code it with the codebook used from the prior distribution because we are interested in quantifying the cost (surprise) incurred by using the wrong codebook (the existing scientific record).

Reporting Findings

Scientists must publish the number of successes and failures sampled from a single arm of their choice. They choose the arm that maximizes the KL divergence between the existing scientific record (ie. p_a) and the record assuming their new findings are published (ie. the new q_a). Again, this makes the assumption that scientists will submit their most interesting and novel findings for publication. This assumption finds support within the current scientific culture, which prioritizes the frequency of publication and the novelty of findings over methodology (Anderson, Ronning, De Vries, & Martinson, 2007; Vinkers et al., 2015; Grimes et al., 2018). We assume that scientists report all successes and failures observed in their selected arm and do not fabricate data.

Choosing Reports for Publication

The publishing layer is constructed to mimic the real process that journals use to publish papers. We introduce the idea of a *pragmatic publisher* that acts as a rational (Bayesian) actor. As such, a publisher simulates taking an action, evaluates its utility, and chooses actions based on their utilities. Publisher actions are restricted to binary decisions of whether or not to publish a submitted report. After each generation, they publish one fifth of the submissions (Hargens, 1988).

We capture variations in publisher choices through the parameters in the utility function that publishers apply to each report. We define utility as a function of three features: [1] the amount of data associated with the report, [2] the surprisingness of the report given the published scientific record, and [3] a publication bias that favors positive findings (Van Aert, Wicherts, & Van Assen, 2019).

We represent the amount of data associated with a report r by $d(r)$, the percentage of samples used for the selected arm. We represent the surprisingness of the report by $s(r)$, the KL divergence between the published record and what the record would be if it were to include the findings from r. Finally, we represent the publication bias by $b(r)$, the percentage difference between the actual and expected number of successes given the number of draws in the selected arm.

Let r be a report and $u_p(r)$ be the utility that the publishing layer ascribes to r. We define $u_p(r)$ as the scaled sum of our

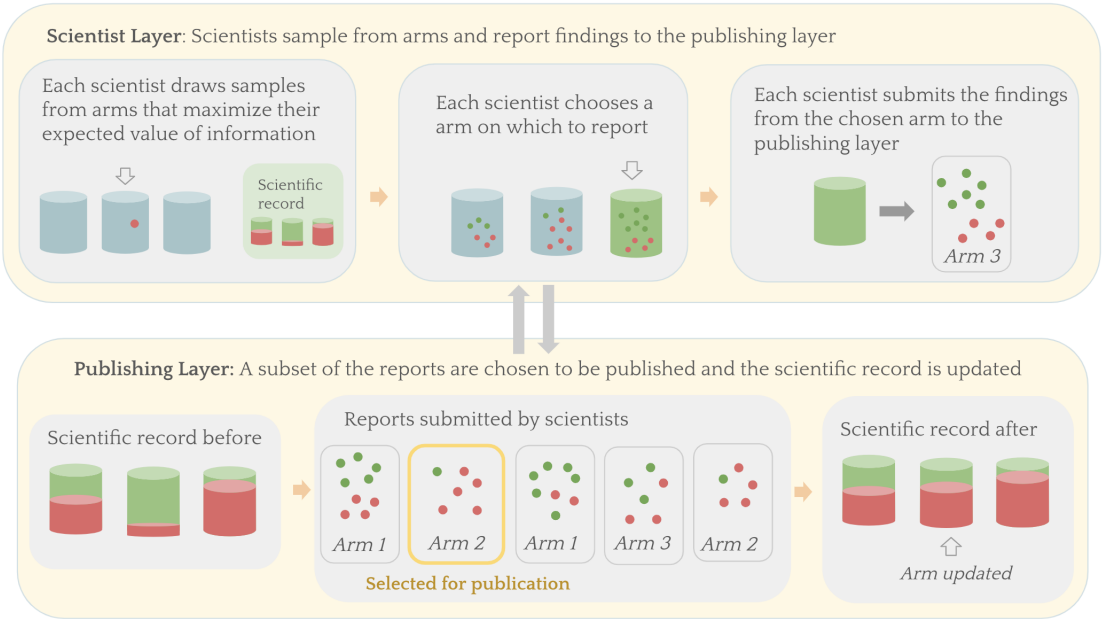

Figure 1: This figure provides a schematic representation of our experimental framework.

three input features. The weighting parameters α, β, and γ are used as dependent variables in our simulations.

$$u_p(r) = \alpha \cdot d(r) + \beta \cdot s(r) + \gamma \cdot b(r)$$

Next, the publishing layer must select a subset of one fifth of the submitted reports to be published. At this stage, each report r has a utility value of $u_p(r)$. We use the softmax function to convert the real number utility values to a probability distribution over publishing actions. Let $\mathbf{u}$ be the vector containing the utility values associated with each report. The formula that we use for our softmax selection is as follows:

$$\sigma(\mathbf{u}) = \frac{\exp(u_i)}{\sum_{j=1}^{N} \exp(u_j)}$$

Reports are selected for publication using the softmax function until one fifth of the submitted reports are published, forbidding duplicate publications.

Evaluating Performance

Our first analysis varies the utility function that publishers ascribe to reports. This objective of the analysis is to identify the publishing behaviors that lead to accurate scientific records. In our simulations, ten scientists each take ten draws from thirty arms. There are fifteen generations of scientists. All arm probabilities are drawn from a uniform distribution. To explore the effects of different publication policies, we search across the relative weights of the three parameters used in the utility function: α, β, and γ. We search over twenty values evenly distributed from 0 to 10 for each parameter.

We calculate the KL divergence between the true (ie. θ_a) and published (ie. q_a) arm probabilities for each combination of values. Higher values correspond to a scientific record that is farther from reality and thus less accurate. We record the average value across ten experiments in each setting. In total, this results in $10 \cdot 15 \cdot 20^3 \cdot 10 = 12$ million experiments.

Our second analysis explores whether variations in arm success rates affect the ability of science to converge on accurate publications. Again, we simulate ten scientists who each take ten samples across thirty arms. In this experiment, there are *fifty* generations of scientists. We hold the value $\alpha = 1$ and $\beta = 1$ constant to simulate a publication policy that equally weighs the amount of data supporting a finding and its surprisingness. We consider ten success probabilities spaced evenly between 0 and 1 and assert that all arms have the same success probability in each experiment. For each probability, we record the KL divergence between the true (ie. θ_a) and published arm probabilities (ie. q_a) in *each generation*. This analysis allows us to compare how arm success probabilities affect the dynamics of publication accuracy over time.

Results

Exploration of Publication Policies

Figure 2 shows the results of our search across the space of publication policies. Recall that higher KL divergence values correspond to a scientific record that is farther from reality and thus less accurate. Sample trajectories of mean KL divergence over generations of scientists under representative publication policies are shown in Figure 2A. The remaining panels show the results of manipulating the weight assigned to each feature of the publication policy.

First, we observe that there is no strong relationship between the amount that supporting data is prioritized in the publishing utility function and the KL divergence between the actual and published arm probability distributions (see Figure

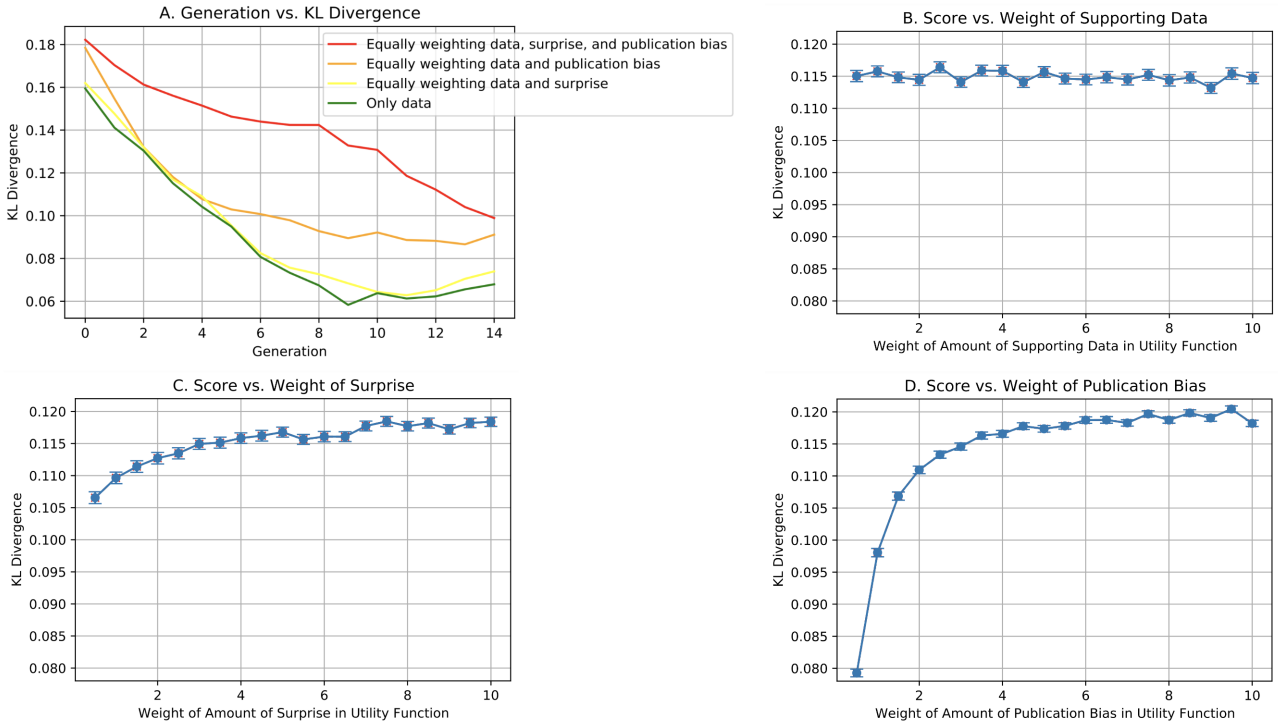

Figure 2: Factors affecting the accuracy of the scientific record. (A) Evolution of the accuracy of the published scientific record, measured by KL divergence between the published and actual arm probability distributions, over time. (B)-(D) Consequences of manipulating the weight assigned to each feature of the publication policy (supporting data, surprise, positive publication bias), averaged over all simulations (error bars show one standard error). Recall that lower KL divergence is better: it corresponds to a scientific record that is closer to reality and thus more accurate.

2B). In other words, we find that prioritizing the quantity of supporting data does little to improve the correspondence between the scientific record and reality.

Second, we observe that surprise, unsurprisingly, has a different effect (see Figure 2C). The more the publisher emphasizes the surprisingess of a report, the farther the published record deviates from true values. The slight plateau observed in Figure 2C may indicate that there is a point after which increasing the priority of surprising results will not significantly worsen the accuracy of the published scientific record.

Finally, we observe that publication bias shows a similar trend to surprise (see Figure 2D). Our results indicate that the more the publisher emphasizes publication bias, the farther the published scientific record deviates from the truth.

Dynamics and Evolution of Scientific Accuracy

Figure 3 shows the evolution of scientific accuracy over time for different arm probabilities.

First, let us examine the case where arm probabilities are very low or very high ($\theta_a < 0.3$ and $\theta_a > 0.7$). We observe that science almost monotonically converges on accurate publications. The explanation for this phenomenon is simple. Let us consider the case where $\theta_a = 0.9$. It is unlikely that any scientist samples a failure from any arm, let alone multiple failures from the same arm. As a result, publications are unlikely to deviate from true probabilities and the weighting of surprise in the publication policy is unlikely to play a large role in the selection of reports for publication. True results are published by early generations and the scientific process only makes the results more accurate as time progresses.[1]

Second, let us examine the case where arm probabilities fall within a range of middle values ($0.3 \leq \theta_a \leq 0.7$). Here, a different phenomenon emerges. In early generations, the published scientific record becomes increasingly *inaccurate*. Only in *later* generations is it able to converge on accurate values. Why does this occur? In settings where sampling a success or a failure from an arm are equally likely or close to equally likely, there are enough scientists conducting experiments from enough arms that one can expect spurious results. As a result, early generations of publishers publish the spurious results into the broader scientific record. It takes time and replications (scientists sampling from well-supported arms) to recover from the incorrect initial publications.

We see that *across all arm probabilities*, our computational model suggests that science will eventually converge on the true state of the world. The variation in arm probabilities affects only the speed and dynamics of this process.

[1]Note that high KL divergence values in early generations are a direct product of the 1-1 prior on each arm in the scientific record.

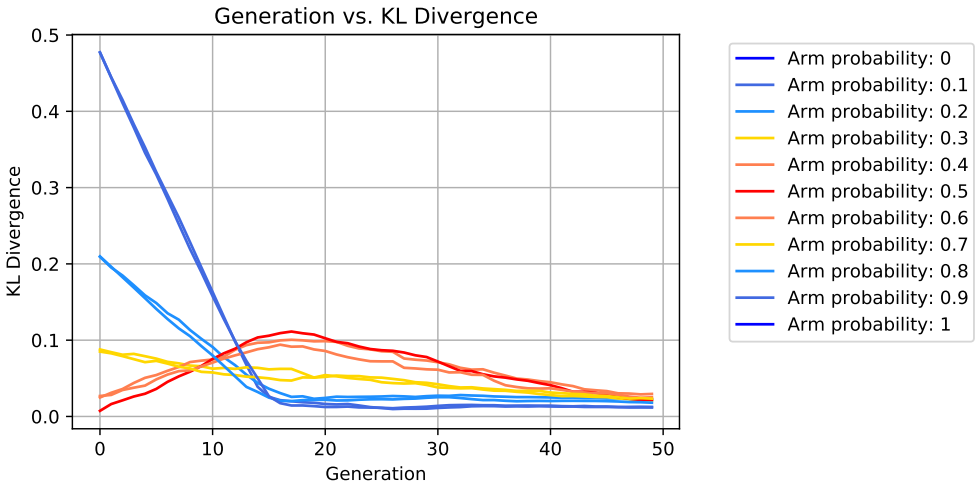

Figure 3: Impact of arm probability – the chance a study produces positive results – on the accuracy of the published scientific record, measured by KL divergence between the published and actual arm probability distributions, over time.

Discussion

This research provides a novel computational framework for exploring the dynamics of the scientific process. Our model frames the scientific process as an iterative interaction between scientists and publishers who collectively update a global record of published information. We design scientists and publishers as virtual agents in a distributed, multi-armed bandit simulation. This approach offers a view of science as the interaction between rational (Bayesian) entities. It also sheds light onto the shortcomings of modern day science and provides a formal account of the replication crisis.

First, we find that the way that publishers prioritize different reports significantly influences the accuracy of the scientific record. When publishers prioritize surprising results and employ a positive publication bias, they hinder the ability of the scientific process to collectively accumulate accurate knowledge. Although both are harmful, small levels of bias towards surprising findings have larger consequences on the development of an accurate scientific record than small levels of bias towards positive findings. Our results suggest that the best policies are those that promote accurate *early* publications that may have otherwise required generations of replication to overturn. It is not what we *don't* know that encourages the poor and inefficient advancement of science, it is what we *think* we know that is *wrong*. This work sheds light on the need for institutional changes at the publishing level to prioritize credible scientific practices.

Another finding is the role of arm success probabilities (or, analogously, hypothesis success rates) on the evolution of the scientific record. We find that many generations of replications may be required for our model to converge on accurate results. In particular, we find that without replications, the integration of spurious results in the scientific record by early generations of scientists can have a lasting effect. This is particularly true in settings where the probability of success lies in middle ranges (between 0.3 and 0.7). Our model offers an explanation for this phenomenon: a rational scientist with limited resources is not inclined to spend them hoping, with what they believe is a low probability, to overturn an established finding. This result supports calls for scientists to pursue and publishers to promote replication studies.

Limitations and Future Directions

Our study is not without its limitations. Our model assumes that all parties involved behave as rational Bayesian actors. In practice, the dynamics of the scientific process can be further distorted by post-hoc hypothesis formation, data manipulation, and blatant fraud (Kerr, 1998). Moreover, heterogeneity in methods used by different scientists to approach the distributed multi-armed bandit problem could be considered in future work. Our paradigm addresses this limitation by offering the ability to independently explore different models of each stage, such as swapping behavioral models of individual scientists. Finally, generational behavioral data to support the simulations could be explored in future research.

Conclusion

We provide a novel, theoretical model of science as a collective, information-gathering process and explore this model through simulations. Building off of the cultural evolution literature, our computational approach highlights behavioral practices at the scientist and publisher levels that either encourage or discourage good science. Our results underscore the value of replications and caution against behaviors that prioritize factors uncorrelated with result accuracy. In science and beyond, the pursuit of truth and dissemination of evidence is critical to the advancement of knowledge.

Acknowledgments. This research project and related results were made possible with the support of the NOMIS Foundation.

References

Aarts, A. A., Anderson, C. J., Anderson, J., van Assen, M. A., Attridge, P. R., Attwood, A. S., ... others (2015). Reproducibility project: psychology. *OSF*.

Anderson, M. S., Ronning, E. A., De Vries, R., & Martinson, B. C. (2007). The perverse effects of competition on scientists' work and relationships. *Science and Engineering Ethics*, *13*, 437–461.

Aschwanden, C., & King, R. (2015). *Science isn't broken.* Retrieved from `https://fivethirtyeight.com/features/science-isnt-broken/`

Beaufils, P., & Karlsson, J. (2013). Legitimate division of large datasets, salami slicing and dual publication. where does a fraud begin? *Orthopaedics & Traumatology: Surgery & Research*, *2*(99), 121–122.

Birch, J., & Heyes, C. (2021). The cultural evolution of cultural evolution. *Philosophical Transactions of the Royal Society B*, *376*(1828), 20200051.

Bird, A. (2021). Understanding the replication crisis as a base rate fallacy. *The British Journal for the Philosophy of Science*, *74*(4).

Brischoux, F., & Angelier, F. (2015). Academia's never-ending selection for productivity. *Scientometrics*, *103*, 333–336.

Coursol, A., & Wagner, E. E. (1986). Effect of positive findings on submission and acceptance rates: a note on meta-analysis bias. , *17*(2), 136–137.

Fiske, S. T. (2004). Mind the gap: In praise of informal sources of formal theory. *Personality and Social Psychology Review*, *8*(2), 132–137.

Grimes, D. R., Bauch, C. T., & Ioannidis, J. P. (2018). Modelling science trustworthiness under publish or perish pressure. *Royal Society Open Science*, *5*(1), 171511.

Hardy, M. D., Krafft, P. M., Thompson, B., & Griffiths, T. L. (2022). Overcoming individual limitations through distributed computation: Rational information accumulation in multigenerational populations. *Topics in Cognitive Science*, *14*(3), 550–573.

Hargens, L. L. (1988). Scholarly consensus and journal rejection rates. *American Sociological Review*, *53*, 139–151.

Head, M. L., Holman, L., Lanfear, R., Kahn, A. T., & Jennions, M. D. (2015). The extent and consequences of p-hacking in science. *PLoS biology*, *13*(3), e1002106.

Ioannidis, J. P. (2005). Why most published research findings are false. *PLoS Medicine*, *2*(8), e124.

Kerr, N. L. (1998). Harking: Hypothesizing after the results are known. *Personality and Social Psychology Review*, *2*(3), 196–217.

Kitcher, P. (1990). The division of cognitive labor. *The Journal of Philosophy*, *87*(1), 5–22.

Kullback, S., & Leibler, R. A. (1951). On information and sufficiency. *The Annals of Mathematical Statistics*, *22*(1), 79–86.

Lewis, H. M., & Laland, K. N. (2012). Transmission fidelity is the key to the build-up of cumulative culture. *Philosophical Transactions of the Royal Society B: Biological Sciences*, *367*(1599), 2171–2180.

Maxwell, S. E., Lau, M. Y., & Howard, G. S. (2015). Is psychology suffering from a replication crisis? what does "failure to replicate" really mean? *American Psychologist*, *70*(6), 487–498.

McElreath, R., & Smaldino, P. E. (2015). Replication, communication, and the population dynamics of scientific discovery. *PloS one*, *10*(8), e0136088.

NeurIPS. (2023). *2023 reviewer guidelines.* Retrieved from `https://neurips.cc/Conferences/2023/ReviewerGuidelines`

Niles, M. T., Schimanski, L. A., McKiernan, E. C., & Alperin, J. P. (2020). Why we publish where we do: Faculty publishing values and their relationship to review, promotion and tenure expectations. *PLoS One*, *15*(3), e0228914.

Pashler, H., & Harris, C. R. (2012). Is the replicability crisis overblown? three arguments examined. *Perspectives on Psychological Science*, *7*(6), 531–536.

Peterson, D., & Panofsky, A. (2023). Metascience as a scientific social movement. *Minerva*, *61*, 1–28.

Robbins, H. (1952). Some aspects of the sequential design of experiments.

Rothstein, H. R., Sutton, A. J., & Borenstein, M. (2005). Publication bias in meta-analysis: Prevention, assessment and adjustments. *Psychometrika*, 1–7.

Schooler, J. W. (2014). Metascience could rescue the 'replication crisis'. *Nature*, *515*(7525), 9.

Shrout, P. E., & Rodgers, J. L. (2018). Psychology, science, and knowledge construction: Broadening perspectives from the replication crisis. *Annual Review of Psychology*, *69*(1), 487-510.

Simmons, J. P., Nelson, L. D., & Simonsohn, U. (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science*, *22*(11), 1359–1366.

Smaldino, P. E., & McElreath, R. (2016). The natural selection of bad science. *Royal Society Open Science*, *3*(9), 160384.

Stevens, J. R. (2017). *Replicability and reproducibility in comparative psychology* (Vol. 8). Frontiers Media SA.

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

Strevens, M. (2017). Scientific sharing: Communism and the social contract. *Scientific collaboration and collective knowledge*, 3–33.

Strevens, M. (2020). *The knowledge machine: How irrationality created modern science*. Liveright Publishing.

Van Aert, R. C., Wicherts, J. M., & Van Assen, M. A. (2019). Publication bias examined in meta-analyses from psychology and medicine: A meta-meta-analysis. *PloS One*, *14*(4), e0215052.

Van Dalen, H. P. (2021). How the publish-or-perish principle divides a science: The case of economists. *Scientometrics*, *126*(2), 1675–1694.

Vinkers, C. H., Tijdink, J. K., & Otte, W. M. (2015). Use of positive and negative words in scientific pubmed abstracts between 1974 and 2014: retrospective analysis. *British Medical Journal*, *351*.

Wiggins, B. J., & Christopherson, C. D. (2019). The replication crisis in psychology: An overview for theoretical and philosophical psychology. *Journal of Theoretical and Philosophical Psychology*, *39*(4), 202–217.

Zhu, J., & Liu, J. (2023). Distributed multi-armed bandits. *IEEE Transactions on Automatic Control*, *68*, 3025-3040.