nature human behaviour

ARTICLES

https://doi.org/10.1038/s41562-020-01040-1

Check for updates

Competition for priority harms the reliability of science, but reforms can help

Leonid Tiokhin[1]✉, Minhua Yan[2,3] and Thomas J. H. Morgan[2,3]

Incentives for priority of discovery are hypothesized to harm scientific reliability. Here, we evaluate this hypothesis by developing an evolutionary agent-based model of a competitive scientific process. We find that rewarding priority of discovery causes populations to culturally evolve towards conducting research with smaller samples. This reduces research reliability and the information value of the average study. Increased start-up costs for setting up single studies and increased payoffs for secondary results (also known as scoop protection) attenuate the negative effects of competition. Furthermore, large rewards for negative results promote the evolution of smaller sample sizes. Our results confirm the logical coherence of scoop protection reforms at several journals. Our results also imply that reforms to increase scientific efficiency, such as rapid journal turnaround times, may produce collateral damage by incentivizing lower-quality research; in contrast, reforms that increase start-up costs, such as pre-registration and registered reports, may generate incentives for higher-quality research.

Academic science is a culturally evolved social institution with formal rules, norms and conventions. However, in recent years, scientists have begun to examine the utility of even longstanding characteristics of this institution[1–3]. For example, it is now widely recognized that preferentially valuing positive over negative results can generate publication bias, which distorts the published literature[4,5]; evaluating scientists based on their number of publications can cause a myopic focus on productivity at the expense of rigour[6]; and rewarding scientists based on the prestige of the journal in which they publish may incentivize scientists to present their work in an overly positive light, submit low-quality papers to high-impact journals and engage in other questionable research practices[7–12].

The priority rule is a particularly longstanding scientific norm, in which individuals who are first to make discoveries receive disproportionate credit relative to all other individuals who provide solutions to the same problem[13,14]. Famously, Charles Darwin was motivated to publish his writings on evolution by means of natural selection in part because of a concern that he would lose priority to Alfred Russel Wallace, who had developed a similar idea. In his famous letter to Charles Lyell, Darwin proclaimed "I rather hate the idea of writing for priority, yet I certainly should be vexed if any one were to publish my doctrines before me"[15]. Rewards for priority take on various forms, including eponymy (that is, naming a scientific discovery after the scientist who discovered it), financial prizes (for example, the Nobel prize), an increased probability of publishing in high-impact journals, and better professional positions and speaking engagements[3,13,16,17]. Little research explicitly documents the career repercussions of losing a priority race (that is, getting scooped). However, one survey of physical and biological scientists found that over 60% of scientists reported being scooped at some point in their careers[18], and a recent study among structural biologists found that scooped papers received 28% fewer citations and were 18% less likely to appear in a top-ten journal[19]. This suggests that scientists have significant incentives to compete over priority of discovery.

Given its role as a major incentive, how does rewarding priority of discovery affect scientific inquiry? Rewards for priority can certainly be useful. For example, they may incentivize scientists to quickly solve problems, share findings with the scientific community and efficiently distribute themselves among multiple research problems[13,14]. The prospect of losing out in a competitive system may also increase individual effort, task performance and innovation relative to a system in which individuals are rewarded for each unit of output regardless of order (refs. [20–22], but see ref. [23]). However, rewarding priority has potential repercussions. For example, it may disincentivize replication if scientists obtain higher expected payoffs by moving on to new research problems after being scooped on existing ones. One particularly longstanding concern is that rewards for priority may cause scientists to rush their work in an attempt to avoid being scooped[13,24–26]. Such rushed research could harm the research process by increasing the probability of mistakes or by reducing the information value (for example, sample size) of the final research product.

Several lines of evidence suggest that rewarding priority may cause scientists to sacrifice the quality of their research. In qualitative interviews, scientists admit to cutting corners in order to outcompete rivals[27]. In laboratory experiments using simple information-sampling paradigms, rewarding priority causes individuals to spend less time on exploration before making decisions between uncertain options[23,28]. More broadly, optimization models of scientists' behaviour suggest that, when novelty is disproportionately valued, scientists optimize their expected payoffs by conducting studies with low statistical power[6,29]. Concerns about rewarding priority in particular are so substantial that the academic journals *eLife* and *PLoS Biology* began to offer scoop protection (that is, allowing researchers to publish findings identical to those already published) in attempts to reduce the disproportionate payoffs to scientists who publish first[30,31], a policy that has recently been adopted by all *PLoS* journals[32].

Although such research is suggestive, it has several limitations. By relying on general information-sampling paradigms in dyadic settings (for example, picking balls from urns[28] or revealing tiles to guess the majority colour on a grid[23]), past experiments have missed critical features of scientific priority races, including the fact

[1]Human-Technology Interaction Group, Eindhoven University of Technology, Eindhoven, the Netherlands. [2]School of Human Evolution and Social Change, Arizona State University, Tempe, AZ, USA. [3]Institute of Human Origins, Arizona State University, Tempe, AZ, USA. ✉e-mail: leotiokhin@gmail.com

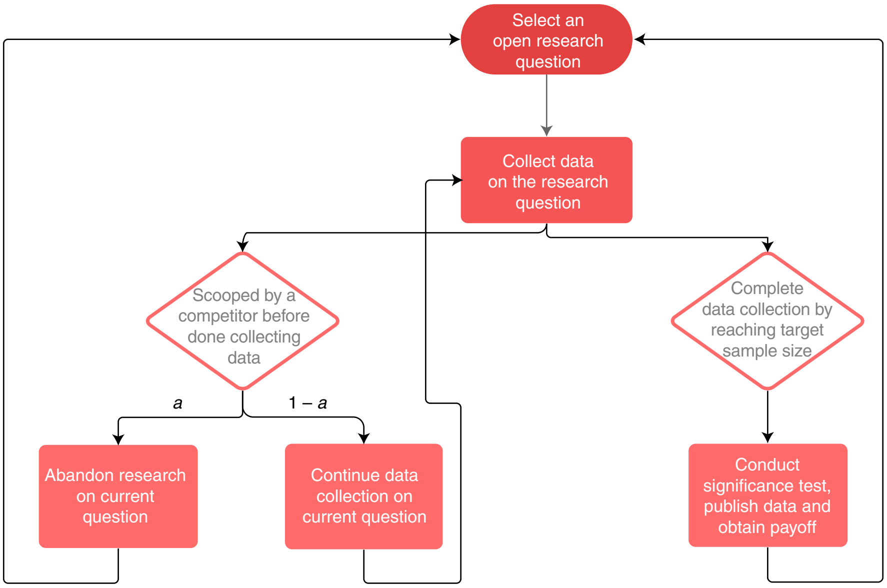
Select an open research question
Collect data on the research question
Scooped by a competitor before done collecting data
Complete data collection by reaching target sample size
a
1 – a
Abandon research on current question
Continue data collection on current question
Conduct significance test, publish data and obtain payoff

Fig. 1 | Scientists' flow of behaviour within a single generation. Scientists begin their career by being assigned to an open research question. They then collect data on that question until reaching their pre-specified sample size, at which point they conduct a significance test, publish their data and obtain a payoff as a function of the type of result (positive or negative) and the number of previously published results on that research question (novelty of the result). After publication, scientists move to a new open research question. Scientists who are in the process of data collection when another scientist publishes on that question (scooped scientists) probabilistically abandon that question every time they are scooped, where *a* determines the probability of abandonment. Scientists who abandon then move to a new research question that is not currently being studied by the scientist who scooped them. Scientists who do not abandon continue data collection on the same question until either reaching their pre-specified sample size or abandoning the question after being scooped by a competitor. This process continues until scientists reach the end of their careers for that generation, at which point all scientists retire.

that scientists can face multiple competitors, can abandon research problems upon being scooped, face time costs to start-up new studies and may receive larger payoffs for certain findings (for example, statistically significant results). Models of priority races have been subject to similar limitations[33]. Other models exploring the relationship between rewarding novelty and research quality have assumed that scientists face optimization problems[6,29,34,35]. This assumption precludes the possibility that scientists' payoffs depend on the strategies of other scientists studying the same questions, which is an essential component of priority races. Moreover, no work has evaluated the logic of whether policy changes offering scoop protection can improve the quality of scientific research. Thus, there is a surprising disconnect between claims about the repercussions of rewarding priority in science and the strength of the evidence that underlies these claims.

Here, we address these issues by developing an evolutionary agent-based model to test the effect of rewarding priority of discovery on the scientific research process. Our model advances the existing literature in several ways. We incorporate critical aspects of real-world priority races, including the possibility of multiple competitors, the ability for scientists to abandon research problems upon being scooped, the fact that new problems have start-up costs and the possibility for differential payoffs for positive and negative results. As it turns out, several of these factors have significant effects on how competition for priority affects scientific reliability. Our model also evaluates the logical coherence of scoop protection reforms and identifies the conditions under which scoop protection increases scientific reliability. We find that, although scoop protection generally increases reliability, this effect is negligible when there are low start-up costs to single studies or when negative results are highly valuable. Finally, our model identifies start-up costs as a heretofore overlooked mechanism to reduce the negative effects of competition. This mechanism has direct implications for the unintended consequences of emerging reforms, such as pre-registration and registered reports.

Results

See the Methods for full model details and https://osf.io/cbftz/ for a code-review report. Consider a population of $n = 120$ scientists. Each scientist is characterized by two parameters representing their characteristic methods: the sample size of their conducted research studies, s, and their probability of abandoning a research question when another scientist publishes a result on that question, a. Scientists transmit their methods to trainees, so the distributions of these parameters can evolve across generations. On any given question, a scientist's statistical power, pwr, is a function of three parameters: sample size, s, the false positive rate, α, and the size of the effect being studied, e. There are an infinite number of research questions, each of which is characterized by an effect size (rounded to one decimal place) drawn from an exponential distribution with a rate parameter, λ, of 5. A maximum of m scientists can work on any given question.

A scientist begins their career on the smallest-numbered open research question. Once their career has started, a scientist collects data until they reach their desired sample size (dictated by their s value). Once a scientist has completed a study, they perform a significance test and obtain a positive result with probability pwr or α for questions with a true effect or no true effect, respectively. The results of all completed studies are published, but there may be bias against negative results (see below). Once a scientist publishes a result, the scientist's payoff is determined by v (the novelty of the result) and whether the result is positive (that is, significant) or negative (that is, non-significant). v is a function of the number of previously published results on a research question and d (the severity of the cost of being scooped). Supplementary Fig. 1 illustrates the function that determines the payoff for a published result.

Table 1 | Parameter definitions and values

Parameter	Definition	Value [range]
n	Population size	120
s	Scientist's target sample size	Uniform [2–1,000]
a	Scientist's probability of abandoning a research question when scooped	Uniform [0–1]
α	False positive rate	0.05
e	Effect size	Exponential (λ)
λ	Rate parameter characterizing distribution of effect sizes	5
t	Scientists' career length	15,000 if $c>10$; 5,000 if $c=10$
c_s	Sample cost (number of time steps to acquire one data point)	1
c	Start-up cost (number of time steps to set up a study)	10, 100, 200 or 400
m	Maximum number of scientists per research question	1, 2, 4 or 8
d	Decay parameter determining the penalty for being scooped	0, 0.15, 0.4, 1 or 10
b_n	Payoff from publishing negative results, relative to positive results	0, 0.25, 0.50, 0.75 or 1.00

s and *a* are unique to each scientist, whereas all of the other parameter values are true for all scientists. Where parameters could take on multiple values, we explored all possible combinations, with 50 repeat simulations for each combination.

Figure 1 provides a visualization of scientists' behaviour within our model. Upon retiring, each scientist's fitness is calculated as proportional to their accumulated payoffs. A new (non-overlapping) generation of scientists is then created, with their s and a values sampled from members of the previous generation, weighted by fitness. This evolutionary component corresponds to the assumption that successful scientists are more likely to pass on their research strategies to subsequent generations. The evolutionary process proceeds for 500 generations. Table 1 summarizes all of the model parameter values. The following section describes the model results. The qualitative patterns presented below hold across all parameter values explored in our model, unless noted otherwise.

More competitors promote the cultural evolution of smaller sample sizes. Figure 2 plots equilibrium sample size as a function of the maximum number of competitors for each research question (m), the relative benefit of negative results (b_n) and the cost of being scooped (d). For illustrative purposes, Fig. 2 depicts a scenario in which the start-up cost, $c=400$. Similar qualitative results occur for all start-up costs (see Fig. 3 and Supplementary Section 1).

The more competitors, the smaller the equilibrium sample size. More competitors increase the probability that any given scientist will be scooped, which favours scientists who conduct research with smaller sample sizes. To illustrate, consider a case where the effect size = 0.2, start-up cost = 0, benefit to negative results = 0 and decay = 10. That is, only positive first publications generate a tangible benefit. Imagine two competitors with sample sizes 50 and 200, respectively. The scientists have statistical power of 0.17 and 0.51, respectively, to detect $e=0.2$. The $s=50$ scientist can conduct four studies (at time periods 50, 100, 150 and 200), while the $s=200$ scientist can only conduct one during the same time (at time period 200). The $s=50$ scientist's probability of detecting at least one statistically significant result before the $s=200$ scientist finishes sampling is $1-0.83^3=0.43$ (the complement of obtaining three non-significant results). In this case, the $s=200$ scientist has a 43% probability of being scooped before completing their study. Now consider a case where the $s=200$ scientist faces seven other competitors, all of whom have $s=50$. In this case, the probability that at least one competitor obtains at least one statistically significant result before the $s=200$ scientist finishes sampling is $1-0.57^7=0.98$ (the complement of all seven competitors obtaining only non-significant results).

Scoop protection promotes larger sample sizes. As scooped results become more beneficial (smaller values of d), populations of scientists evolve towards larger equilibrium sample sizes. In other words, scoop protection favours larger studies. Larger benefits to publishing scooped results allow scientists who are most likely to get scooped (that is, those with larger sample sizes) to receive larger payoffs. This reduces the relative payoff difference between scooped scientists and those who are fastest to finish sampling (that is, those with smaller sample sizes).

Rewarding negative results promotes smaller sample sizes. As negative results become more beneficial (larger values of b_n), populations of scientists evolve towards smaller equilibrium sample sizes (Fig. 2). When positive and negative results are equally valuable ($b_n=1$), the effect of the other parameters is minimal: populations rarely evolve to sample sizes larger than 10. This occurs because scientists have little incentive to conduct large studies—conducting a small, underpowered study usually produces a negative result, but this result is worth just as much a result from a larger, well-powered study. However, conducting many small studies produces results at a higher rate than conducting fewer large studies. This favours scientists who conduct studies with smaller sample sizes.

Larger start-up costs promote larger sample sizes. Figure 3 plots equilibrium sample size as a function of the number of competitors, the relative benefit of negative results and the start-up cost to single studies. For illustrative purposes, we depict only two values for the cost of being scooped (see Supplementary Section 1.1).

When start-up costs are small, populations of scientists evolve towards very small sample sizes. Larger start-up costs increase equilibrium sample sizes. The reason for this effect is as follows. Scientists who conduct studies with small sample sizes have low statistical power—their expected probability of obtaining a statistically significant result in a single study is low. Instead, their success depends on performing many studies as quickly as possible. This is most profitable when start-up costs are low because scientists can perform multiple successive studies quickly. When the goal is to obtain at least one statistically significant finding, scientists use a simple statistical test to compare the means of two groups (for example, a t-test), and effect sizes are small to medium, running many underpowered studies is a more efficient strategy than running a single well-powered study[36]. Large start-up costs disincentivize scientists from pursuing such a quantity strategy because they impose a time cost on the scientist every time they start (or restart) a study. Such a time cost disproportionately affects scientists who conduct more, smaller-sample-size studies.

Thus far, we have focused on how different reward structures affect the optimum sample size for individual scientists. However, these individual strategies have consequences for the efficiency and reliability of science as a whole. We assess these consequences by computing several population-level outcomes: (1) positive predictive value (PPV); (2) the proportion of research questions with more true than false results; and (3) the average change in log-odds belief per study. We also explore: (4) the proportion of time spent productively; (5) the proportion of results that are true; (6) the proportion

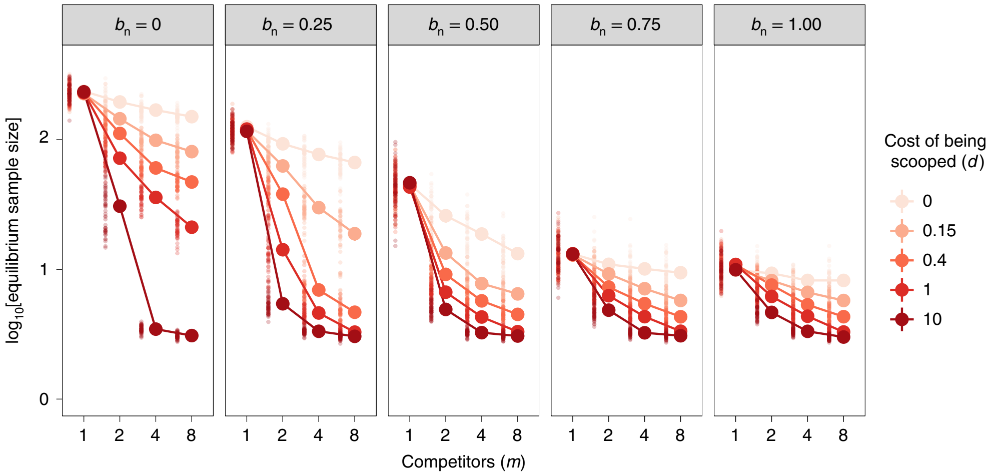

Fig. 2 | Equilibrium sample size for individual scientists. Equilibrium sample size for individual scientists (500 generations and 50 repeats) as a function of m, b_n and d, plotted for a start-up cost, c, of 400. Error bars represent two standard errors. When $b_n = 0$, negative results have no value. When $b_n = 1$, negative results are as valuable as positive results. Equilibrium sample size decreases when there are more competitors, when the cost of being scooped (that is, decay) is large and when negative results are more valuable.

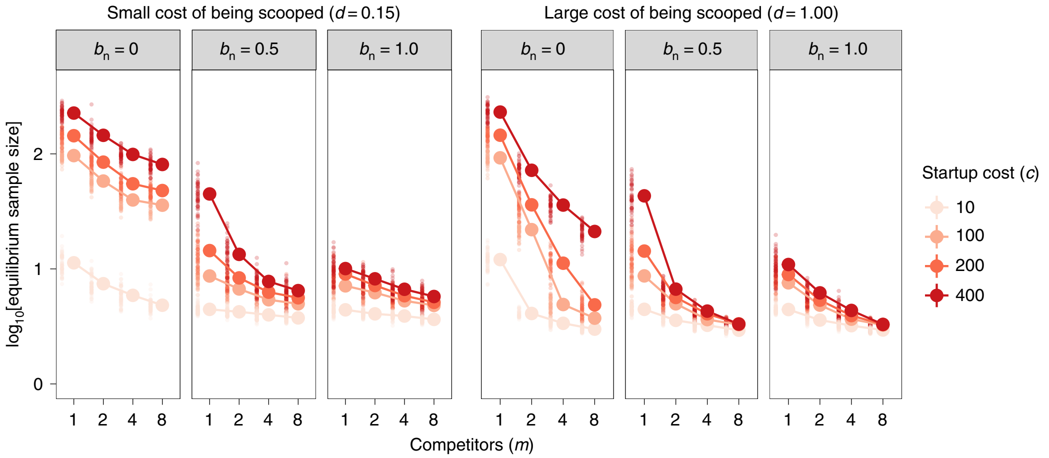

Fig. 3 | Equilibrium sample size for individual scientists as a function of the start-up cost to single studies. Equilibrium sample size for individual scientists as a function of the start-up cost to single studies (500 generations and 50 repeats), plotted for two levels of decay: $d = 0.15$ (small cost of being scooped) and $d = 1.00$ (large cost of being scooped). Error bars represent 2 s.e. When $b_n = 0$, negative results have no value. When $b_n = 1$, negative results are as valuable as positive results. Equilibrium sample size increases as start-up costs increase. Equilibrium sample size decreases when there are more competitors, when the cost of being scooped (that is, decay) is large and when negative results are more valuable.

of questions with equal or more true than false results; (7) the total number of research questions with more true than false results; (8) the difference between the total number of true and false results; and (9) the average change in the absolute value of log-odds belief (see Supplementary Section 3).

PPV. PPV is the probability that a positive (that is, statistically significant) result corresponds to a true effect. We calculate PPV by dividing the number of true positive results by the total number of positive results (that is, both true and false positives). Figure 4 depicts PPV for start-up costs of 10 and 400. More competitors, larger benefits to negative results, smaller start-up costs and a larger cost to being scooped all decrease PPV. This occurs because these factors cause scientists to conduct studies with smaller sample sizes, which causes the average study to have lower statistical power. This decreases the true positive rate, while the false positive rate remains constant, which lowers the ratio of true to false positive results[37].

Proportion of research questions with more true results. Scientists sometimes assess evidence for research questions using heuristic tallies of positive and negative results[38]. As such, the proportion of questions with more true than false results is a useful metric for evaluating the proportion of questions for which scientists will acquire accurate beliefs. This metric is not equivalent to the proportion of questions with more significant versus non-significant results (that is, a tally[39]) because, in our model, approximately 22% of questions have null effects. Figure 5 depicts the proportion of research questions with more true than false results, for start-up costs of 10 and 400.

Competition decreases the proportion of questions with more true than false results. This effect occurs because competition lowers equilibrium sample sizes, which decreases the statistical power of the average study, thereby increasing the proportion of results that are negative on questions with true effects. The only region of parameter space in which the proportion of questions with more true than false results is approximately 50% or larger occurs when

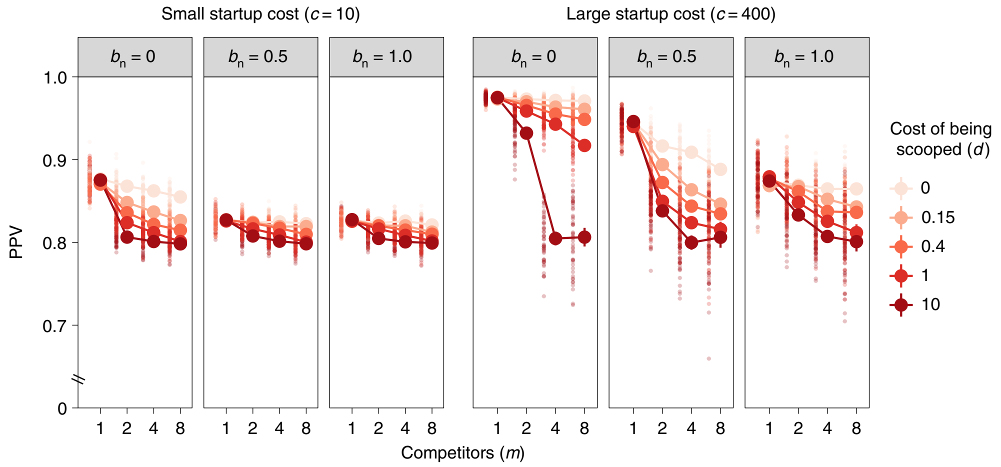

Fig. 4 | Positive predictive value. PPV (500 generations and 50 repeats), plotted for two levels of start-up cost, $c = 10$ (small start-up cost) and $c = 400$ (large start-up cost). Error bars represent 2 s.e. Note the axis break. When $b_n = 0$, negative results have no value. When $b_n = 1$, negative results are as valuable as positive results. PPV increases as start-up costs increase. PPV decreases when there are more competitors, when the cost of being scooped (that is, decay) is large and when negative results are more valuable.

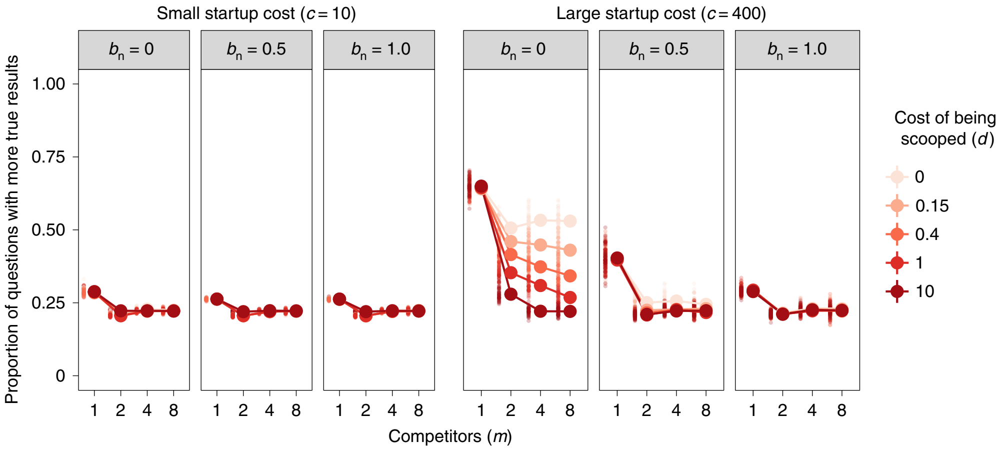

Fig. 5 | Proportion of questions with more true than false results. Proportion of questions with more true than false results (500 generations and 50 repeats), plotted for two levels of start-up cost, $c = 10$ (small start-up cost) and $c = 400$ (large start-up cost). Error bars represent 2 s.e. When $b_n = 0$, negative results have no value. When $b_n = 1$, negative results are as valuable as positive results. The proportion of questions with more true results increases as start-up costs increase. When start-up costs are low, when the cost of being scooped (that is, decay) is high and when the benefit to negative results is high, the proportion of questions with more true results always remains low.

positive results are worth much more than negative ones and the costs for being scooped are small.

When the start-up cost is low, scientists conduct studies with nearly the smallest possible sample size, even without competition. This means that most results are negative, despite the fact that 78% of research questions have a true effect. Overall, 22% of the time, the effect size of a question is 0, and the population ends up generating more negative than positive results on that question. When the start-up cost is high, competition has the largest effect on equilibrium sample size, which reduces the proportion of questions with more true results. When either the benefit to negative results or the cost of being scooped is large, equilibrium sample size is already small, so competition has a minimal effect.

Average per-study change in belief. Consider a scenario in which scientists are perfect Bayesians and use Bayes' rule to update their beliefs regarding the epistemic status of effects. Assume that scientists know (1) the results of all published studies; (2) the global false positive rate; (3) the average effect size; and (4) their studies' statistical power to detect the average effect size. The assumption that scientists know their statistical power to detect the average effect but are unaware of their exact statistical power on a specific research question is reasonable—knowledge of exact statistical power requires perfect information about each effect size, which is unrealistic and would mean that conducting a study is unnecessary in the first place.

Will such Bayesian scientists inevitably acquire accurate beliefs about whether research questions have a true or null effect? Or are there cases in which scientists will acquire false beliefs about the epistemic status of an effect? To address this question, we computed the mean change in belief per published result, for each effect size, across all model parameter combinations. The mean change in belief in the correct direction is one indicator of the information value of the average study. As in a recent model of scientists' expected change in beliefs[5], we use a log-odds scale. This is convenient because, unlike the probability scale, each published result increases or decreases the log-odds belief by a constant increment. Each published positive result increases the log-odds of belief by the following constant increment:

$$\ln\left[\frac{1-\beta}{\alpha}\right] > 0$$

Each published negative result decreases the log-odds of belief by the following constant decrement:

$$\ln\left[\frac{\beta}{1-\alpha}\right] < 0$$

where β is the false negative rate ($1-pwr$) and α is the false positive rate (0.05). For details and derivation, see ref. [5]. Figure 6 plots the mean change in log-odds belief as a function of the effect size, number of competitors and the cost of being scooped, when the benefit to negative results is 0.25. The same patterns hold for effect-size distributions with a larger proportion of null effects (see Supplementary Section 5).

For all effects (except $e=0.1$; see below) the average study shifts scientists' beliefs in the correct direction. That is, the average study makes scientists more confident that a true effect is indeed true and a null effect is indeed null. More competitors, larger costs for being scooped, smaller start-up costs and larger benefits to negative results all decrease the average change in log-odds belief by decreasing the average sample size of conducted studies. That is, when sample sizes are small, the average study provides less information. When effects are very small ($e=0.1$), the average study shifts scientists' beliefs in the wrong direction—a shift that occurs across all parameter values in our model. This occurs because, when effect sizes are small, scientists overestimate their statistical power and their beliefs are more strongly influenced by the large number of false negative results than they would be if scientists had perfect information about their statistical power.

Discussion

We developed an evolutionary agent-based model to test the effect of rewarding priority of discovery on the scientific research process. Our model incorporated critical aspects of real-world priority races, including the possibility of multiple competitors, the ability for scientists to abandon research problems upon being scooped, start-up costs to new problems, and differential payoffs for positive and negative results. We find that, across a broad range of parameters, rewarding priority causes populations to culturally evolve towards conducting research with smaller sample sizes. This reduces the reliability of published research and the information value of the average study. We identify two ways to attenuate the negative effects of competition for priority: increasing the start-up cost to setting up single studies and increasing the payoffs for scientists who are scooped. However, we find that the benefits of scoop protection are negligible when either negative results are highly rewarded or start-up costs are small. Our model also identifies conditions under which rewarding negative results incentivizes lower-quality research and conditions under which the average study causes scientists to develop false beliefs about the epistemic status of effects.

Scholars have had longstanding concerns that competition negatively affects the scientific process[13]. Such concerns have even inspired scoop protection reforms at several prominent journals[30–32]. Our model provides theoretical support for such reforms. Allowing scooped scientists to receive some payoff reduces the incentive to run small-sample-size studies in order to increase the probability of being the first to publish a result, which improves the average quality of conducted studies. However, scoop protection is no panacea, as it causes scientists to persist on research questions even after several results have been published, which leads the population to investigate fewer total questions (see Supplementary Section 3.9). Furthermore, when starting up a new study is cheap or when negative results are highly valued, scientists are incentivized to run small studies even with scoop protection. This reduces the positive predictive value (Fig. 4) and causes the majority of research questions to have more false than true findings (Fig. 5). Thus, although our model supports the logical coherence of scoop protection reforms, it also highlights that scoop protection is not sufficient to incentivize high-quality research or reliable published literature.

In our model, increased start-up costs allow populations to maintain higher sample sizes at equilibrium. Start-up costs are far from efficient: every researcher is forced to waste time on each investigation, resulting in fewer questions investigated and fewer completed studies (see Supplementary Sections 3.7 and 3.8). However, start-up costs disincentivize a quantity strategy wherein researchers conduct large numbers of underpowered studies[36]. This occurs because start-up costs place a time cost on a scientist every time they start a study, and scientists with smaller sample sizes pay this cost relatively more frequently. Our results point to start-up costs as one potentially important solution to the problem of scientific unreliability. Coincidentally, existing reforms have inadvertently introduced such costs. For example, pre-registration and registered reports make researchers spend more time thinking about and designing protocols before running investigations[40,41]. The time cost inherent in these practices is often conceptualized as an inconvenience. However, our model implies that such costs have an important function: they incentivize scientists to conduct higher-quality research than they would otherwise.

Note that the mechanism by which start-up costs incentivize higher-quality research does not necessarily depend on the timing of such costs. Any costs disproportionately paid by scientists who attempt to conduct quick, low-quality research will serve the same function[10]. These might include wrap-up costs, such as long peer review times, or costs at other points in the research pipeline (for example, an obligation to peer review n other papers for each submitted paper). In fact, start-up costs are a specific instantiation of a more general class of phenomena, wherein certain costs (for example, search costs and costs to beginning new relationships) incentivize individuals to invest more in a current endeavour instead of abandoning it in search of potentially better alternatives[42,43]. Other examples include the lengthy courtship rituals of some bird species (for example, albatrosses) and costly gift giving in interpersonal relationships[44,45]. The more general lesson is that it is wrong to conceptualize inefficiencies in the scientific process as necessarily harmful. For example, models demonstrate that inefficiencies in academic publishing, such as submission costs or long waiting times, can disincentivize scientists from submitting low-quality work to high-impact journals[10,46–50] or from submitting low-quality grant proposals to funding competitions[51]. The relevant question is thus not 'how can we reduce scientific inefficiency?', but rather 'what are the costs and benefits of inefficiency and when are the costs large enough to worry about?'.

In practice, will adding costs to the research pipeline produce desirable outcomes? This depends on several factors. The extent to which start-up costs incentivize higher-quality research depends on sampling costs. If sampling costs are large, the time required to conduct a single study is determined primarily by sampling costs, and vice versa. Start-up costs lose their effectiveness as sampling costs become relatively large (see Supplementary Section 5.9). Thus, a complementary way to increase research quality is to reduce data collection inefficiencies without altering start-up costs. Other issues concern scientists' strategic responses to costs. If scientists can circumvent costs (for example, avoiding wrap-up costs by file-drawering studies with undesirable results), then costs may not effectively incentivize higher-quality research. If start-up costs are too high, scientists may be more willing to engage in question-

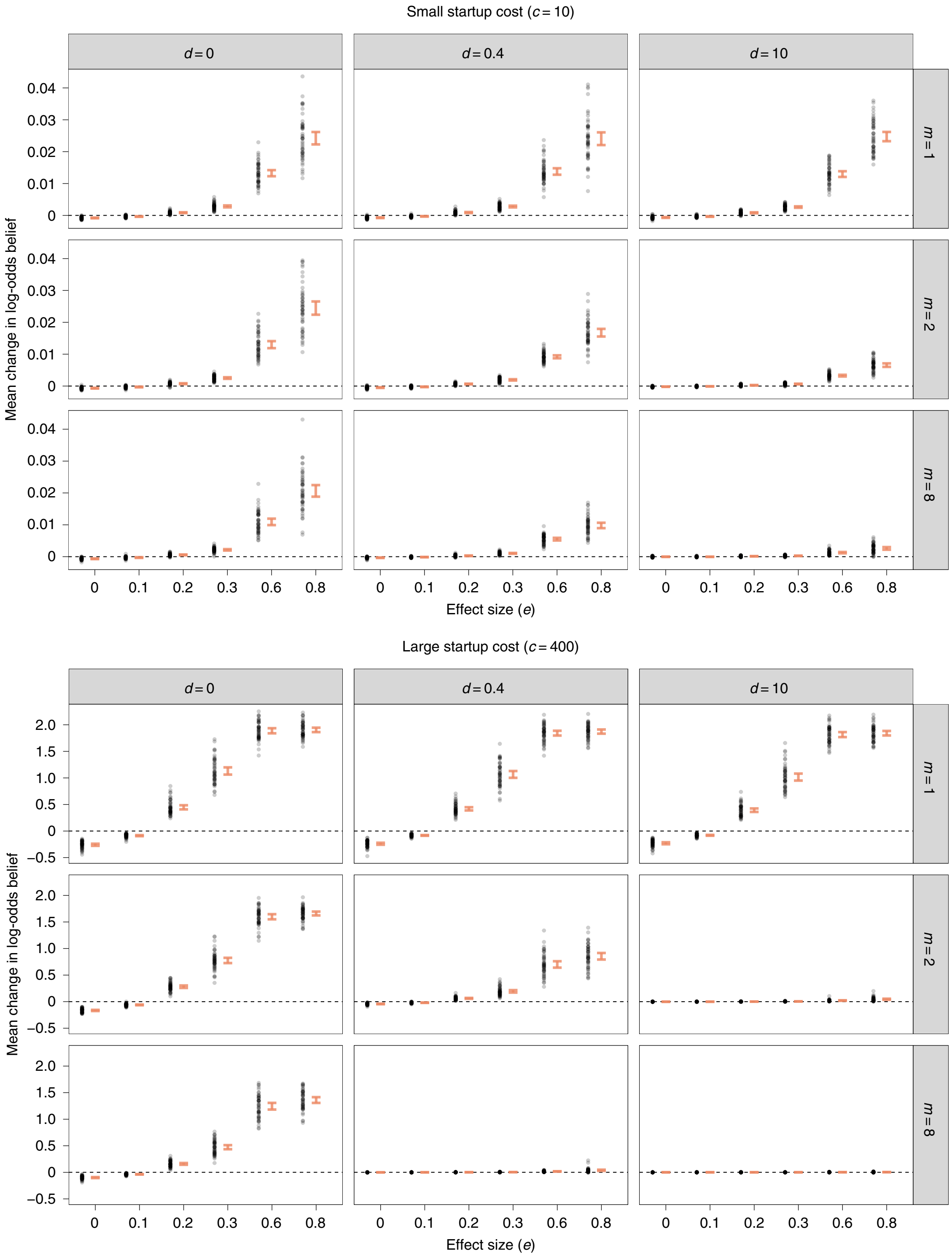

Fig. 6 | Mean change in log-odds belief per study. Mean change in log-odds belief per study (500 generations and 50 repeats), plotted for two levels of start-up cost, $c=10$ (small start-up cost) and $c=400$ (large start-up cost), and one level of benefit to negative results, $b_n=0.25$. Note the different axes in the two sets of plots. The y axis represents the natural log. Error bars represent 2 s.e. Dashed lines indicate no shift in belief. d, decay; m, number of competitors. For most effect sizes, the average study shifts scientists' beliefs in the correct direction. When effects are small (that is, $e=0.1$), the average study shifts scientists' beliefs in the direction of no effect, despite the fact that a true effect exists. This pattern indicates that a population of scientists using Bayesian updating would be expected to shift their beliefs towards 100% confidence that true effects existed and that null effects did not exist. The exception is that, when true effects are small, scientists would be expected to shift their beliefs towards 100% confidence that there was no effect, despite the fact that a true effect existed.

able research practices[52] to obtain publication-worthy findings from existing projects. Finally, it is worth noting that other solutions to the problem of scientific reliability (for example, stricter quality controls and minimum statistical power requirements) may be more desirable because they impose less of a burden on the scientific process or less of a cost on scientists' well-being.

In our model, rewarding negative results reduces the equilibrium sample size and harms scientific reliability. This pattern occurs because we assume that the payoff for publication is independent of sample size or effect size. When negative and positive results are equally valuable, scientists have no incentive to conduct large studies in order to increase the probability of positive results—a large study costs more time but generates the same payoff as an underpowered study that can quickly produce a negative result. How does this finding fit into ongoing discussions about whether researchers should publish all of their results[2,53,54], or whether some types of publication bias[5,39,55,56] or publication restrictions[10,57–59] are desirable? Our model points out that, when study quality is not sufficiently rewarded, a bias against negative results incentivizes scientists to conduct larger studies. However, such an outcome comes at the cost of reducing the amount of information in the published literature[5] and biasing estimates of true underlying effect sizes[60]. A better solution would thus be to supplement reforms to publish negative results with reforms that disincentivize underpowered studies. It is promising that several emerging reforms, such as changes to norms for reporting statistics (for example, effect sizes and confidence intervals[61]), alternative statistical approaches (for example, Bayes factors and equivalence tests[62,63]) and new publishing formats that require high levels of statistical power (for example, Registered Reports[41]), all plausibly increase the rewards for null results from high-quality studies.

Our model has implications for the extent to which science is self-correcting. A longstanding notion is that the normative structure of science prevents the proliferation of false claims[64]. However, it is becoming increasingly clear that scientific self-correction is not guaranteed[65] and that many factors can cause scientists to converge on false beliefs. Known mechanisms that impede self-correction include publication bias against null results[5], lack of replication research[39,66], excessive conformity[67] and others[68–70]. Our model demonstrates another mechanism by which scientists may converge on false beliefs. When effect sizes are small, the average study is so underpowered that most results are false negatives. Because scientists update their beliefs based on the average effect size and their studies' statistical power to detect the average effect, they overestimate their statistical power (that is, underestimate their false negative rate) on research questions where effect sizes are smaller than average. As a consequence, each published negative result decreases scientists' belief that there is a true effect more strongly than it would if scientists had perfect information about their statistical power. This causes scientists to falsely believe that there is no effect when a true effect indeed exists (Fig. 6). This finding suggests that low statistical power combined with imperfect information about statistical power are sufficient to cause scientists to converge on false beliefs, and provides yet another reason why increasing the statistical power of empirical research is essential[60].

Our model suggests several avenues for empirical research (Table 2). These include evaluating how research quality varies across fields that vary in the cost of being scooped and whether reforms that increase the start-up cost to single studies (for example, pre-registration) incentivize fewer, higher-quality studies. Our model has several limitations, which could be addressed in future work. For example, our assumption that payoffs are independent of study quality or effect size could be modified such that larger studies and effects generate higher payoffs. Similar assumptions are made in some models of priority races in which mature ideas receive larger payoffs[33]. Our assumption that scientists pay a cost for each study could be modified such that scientists strategically decide whether to pay costs. Another assumption—that scientists can only respond to competition by modifying their sample size and probability of abandonment—ignores other potential responses to competition (for example, increasing research effort[21,22,71], but see refs. [23,72]). Other extensions might allow questionable research practices[52] in response to competition. Finally, our model assumes that all results are published, in contrast with several existing models[5,29,35,39]. We do not think that modifying this assumption would qualitatively affect our results. Our model varied the relative benefit of negative results while keeping publication probability constant. This is equivalent, in terms of expected value, to varying the probability of publication without varying the relative benefit to negative results. Furthermore, a simpler version of our model in which only positive results were published produced the same qualitative patterns as the current model[73].

Table 2 | Hypotheses for future empirical research

Larger rewards for scientists who are first to provide a solution to a research problem (for example, a treatment for COVID-19) lead to lower-quality research as scientists attempt to increase their chances of coming first.
Across fields, the extent to which novel results receive a larger payoff than secondary (that is, scooped) results is associated with lower statistical power and a higher rate of errors in published studies.
Reforms to increase the publication of negative results, without corresponding controls on the quality of research, lead scientists to conduct lower-quality studies.
Field-wide reforms that increase the start-up cost to single studies (for example, mandating rigorous pre-registration of each conducted study) cause to scientists to conduct fewer (but higher-quality) studies.
Fields in which individual data points are cheap relative to the start-up costs of single studies will be characterized by higher-quality studies than fields in which individual data points are relatively costlier (for example, studies of captive versus wild animal populations).

Effective interventions to improve scientific practice require a causal understanding of the forces that shape scientists' behaviours. Our model takes one step towards this goal. We encourage more formal modelling of ideas for scientific reform, as a complement to verbal arguments and empirical tests. Such models are useful for evaluating ideas in theory instead of wading directly into the empirical morass[74]. This improves scientific efficiency by weeding out logically incoherent ideas, determining the conditions under which an idea applies, and making transparent which observations must be made to test an idea's empirical validity[75]. After all, science walks forward on two feet—theory and experiment—and continuous progress depends on maintaining an intimate connection between the two[76].

Methods

Consider a population of $n = 120$ scientists. Each scientist is characterized by two parameters representing their characteristic methods: the sample size of their conducted research studies, s, and their probability of abandoning a research question when another scientist publishes a result on that question, a. Scientists transmit their methods to trainees and trainees select mentors according to mentors' success, so the distributions of these parameters can evolve across generations. Each population is initialized by sampling n integer values of s from a uniform distribution [2–1,000] and n real-numbered values of a from a uniform distribution [0–1]. Sensitivity checks indicate that the simulation results are robust to initializing populations from distributions of high or low s and a values (see Supplementary Section 5.2 and 5.3) and that the long-run stable distributions of sample sizes (that is, equilibrium sample sizes) to which populations evolve are robust to running the simulation with a larger population size (see Supplementary Section 5.5). Equilibrium abandonment probabilities are affected by population size; however, the abandonment strategy that evolves at large population sizes is highly artificial and does not qualitatively affect equilibrium sample sizes (for full abandonment analyses, see Supplementary Sections 2, 5.5 and 5.6).

On any given question, a scientist's statistical power, pwr, can take on any real-numbered value in the range 0.05–1. pwr is a function of three parameters: sample size, s, the false positive rate, α, and the size of the effect being studied, e. pwr is calculated using a two-sample t-test, implemented with the pwr.t.test() function in the pwr package in R[77,78]. This effectively assumes that all research is of the form where scientists collect s independent data points from each of two populations and test for a difference between the two. Examples of such a research question may be whether some drug (for example, lithium) effectively treats some disease (for example, bipolar disorder) or whether *P* values are more difficult to understand than Bayes factors.

Following convention, the level of statistical significance required for a positive result, α, remains fixed at 0.05. We assume that there are an infinite number of research questions, each of which is characterized by an effect size *e*, where *e* represents a standardized mean difference between two populations. Given that effect sizes in several fields are known to be distributed exponentially[79,80], we assume that the *e* value of each question is drawn from an exponential distribution, with a rate parameter (λ) of 5 and rounded to the nearest 0.1. This corresponds to a distribution of Cohen's *d* effect sizes with a mean of 0.20 and a median of 0.10, and where roughly one in five research questions has an effect size of 0. Alternative distributions of exponentially distributed effects do not qualitatively affect our results (see Supplementary Section 5 and ref. [73]).

Each research question has a unique ID (for example 1, 2, 3, ….) and a maximum of *m* scientists can work on any given question. A scientist begins their career on the smallest-numbered open research question (that is, the smallest-numbered question occupied by fewer than *m* other scientists). We do this to avoid unrealistic outcomes (for example, all scientists working on a single question or all scientists working on different questions) and to control the intensity of competition by manipulating the number of scientists allowed to work on a single question. Each scientist's career lasts $t = 15,000$ time steps. In one specific case of low start-up costs ($c = 10$; see below), career length was reduced to 5,000 time steps for computational efficiency, without affecting the simulation results (see Supplementary Section 5.1).

Once their career has started, a scientist collects data until they reach their desired sample size as dictated by their respective *s* value. The number of time steps required to do this, *t*, is:

$$t = sc_s + c,$$

where c_s represents the sample cost (the number of time steps needed to acquire one data point (fixed at 1)) and *c* represents the start-up cost (the number of time steps needed to set up a study). Thus, as *c* increases, variations in *s* have a smaller effect on a scientist's time cost per study. We assume that *c* is independent of *s* (for example, scientists may need to obtain Institutional Review Board approval or pre-register their research plan before conducting a study); such actions cost time independent of the number of participants that a scientist ultimately recruits. Once a scientist has completed a study, they perform a significance test. For questions with a true effect ($e > 0$), a scientist obtains a statistically significant result with probability *pwr*. For questions with no true effect ($e = 0$), a scientist obtains a statistically significant result with probability α.

We assume that the results of all completed studies are published but that there may be bias against negative results (see below). Once a scientist publishes a result, the scientist's payoff is determined by the novelty of the result, *v*, and whether the result is positive (that is, significant) or negative (that is, non-significant). The novelty of a result is calculated as:

$$v = \left(\frac{1}{1 + \text{number of previous results on question}}\right)^d$$

where *d* (the decay) determines the severity of the cost of being scooped. When *d* is small (for example, <0.5), *v* decays slowly, whereas when *d* is large (for example, >2), *v* decays rapidly. Supplementary Fig. 1 illustrates the relationship between *d* and *v* as a function of the number of published results. For positive results, scientists receive payoff *v*. For negative results, scientists receive payoff vb_n, where $0 \le b_n \le 1$. In the extreme case of $b_n = 0$, there is no reward for publishing null results. This payoff function reflects the assumption that statistically significant results may be valued more than non-significant ones and is mathematically equivalent, in terms of expected value, to assuming that non-significant results have a smaller probability of being published.

After publishing, the scientist moves to the next open research question (that is, one with fewer than *m* other scientists working on it) for which no results have yet been published. The scientist then starts a new study from scratch. All other scientists working on the question corresponding to the newly published result (and who are not themselves publishing a result during that time period) abandon that question with a probability determined by their individual *a* value.

Those who abandon move on to the next open research question. To prevent scientists from getting stuck on the same questions as the scientist who just scooped them, we assume that scientists who abandon move to a different question than the one assigned to their scooper (see Supplementary Section 6). This process repeats until scientists reach the end of their careers, at which point all scientists retire. Figure 1 provides a visualization of scientists' behaviour within our model.

When one generation of scientists retires, a new (non-overlapping) generation is created. Each new trainee scientist inherits their *s* and *a* values from mentor scientists in the previous generation. For each trainee, mentors are chosen with a probability equal to the prospective mentors' accumulated payoffs divided by the accumulated payoffs accrued by all prospective mentors. Mentors for each value are chosen independently, so most scientists have two mentors. This evolutionary component of our model corresponds to the assumption that scientists who are more successful (for example, have more publications) are more likely to pass on their research strategies to the subsequent generation of scientists. This is plausible if younger scientists preferentially imitate the behaviours of successful, well-established scientists (that is, payoff-biased social learning[81,82]) or if scientists who are more successful are more likely to remain in academia and are thus disproportionately available as cultural models for other scientists[6,83,84]. We assume that inheritance is noisy: a trainee's *s* value is drawn from a normal distribution centred on their mentor's value with a standard deviation of 2. The resulting *s* values are rounded to the nearest integer and truncated to remain in the range 2–1,000. Values of $s < 2$ are set to 2 because two-sample *t*-tests require at least two samples per group. Similarly, a trainee's *a* value is drawn from a normal distribution centred on their mentor's value with a standard deviation of 0.01 and truncated to remain in the range 0–1. Table 1 summarizes all of the parameter values used in our model.

To ensure convergence to equilibrium sample sizes (see Supplementary Section 7), the evolutionary process proceeds for 500 generations, at which point the simulation stops.

Reporting Summary. Further information on research design is available in the Nature Research Reporting Summary linked to this article.

Data availability

No empirical data were generated during the current study. Simulation results are available on the Open Science Framework repository (https://osf.io/cbftz/).

Code availability

The R code for the agent-based model, figures and supplementary analyses, as well as a code-review report, are available on the Open Science Framework repository (https://osf.io/cbftz/).

Received: 22 May 2020; Accepted: 18 December 2020;
Published online: 28 January 2021

References

1. Ioannidis, J. P. How to make more published research true. *PLoS Med.* **11**, e1001747 (2014).
2. Munafò, M. R. et al. A manifesto for reproducible science. *Nat. Hum. Behav.* **1**, 0021 (2017).
3. Nosek, B. A., Spies, J. R. & Motyl, M. Scientific utopia: II. Restructuring incentives and practices to promote truth over publishability. *Perspect. Psychol. Sci.* **7**, 615–631 (2012).
4. Franco, A., Malhotra, N. & Simonovits, G. Publication bias in the social sciences: unlocking the file drawer. *Science* **345**, 1502–1505 (2014).
5. Nissen, S. B., Magidson, T., Gross, K. & Bergstrom, C. T. Publication bias and the canonization of false facts. *eLife* **5**, e21451 (2016).
6. Smaldino, P. E. & McElreath, R. The natural selection of bad science. *R. Soc. Open Sci.* **3**, 160384 (2016).
7. Jellison, S. et al. Evaluation of spin in abstracts of papers in psychiatry and psychology journals. *BMJ Evid. Based Med.* https://doi.org/10.1136/bmjebm-2019-111176 (2019).
8. McKiernan, E. C. et al. Use of the Journal Impact Factor in academic review, promotion, and tenure evaluations. *eLife* **8**, e47338 (2019).
9. Quan, W., Chen, B. & Shu, F. Publish or impoverish: an investigation of the monetary reward system of science in China (1999–2016). *Aslib J. Inform. Manage.* **69**, 486–502 (2017).
10. Tiokhin, L. et al. Honest signaling in academic publishing. Preprint at *OSF* https://doi.org/10.31219/osf.io/gyeh8 (2019).
11. Vazire, S.Quality uncertainty erodes trust in science. *Collabra Psychol.* **3**, 1 (2017).
12. Vinkers, C. H., Tijdink, J. K. & Otte, W. M. Use of positive and negative words in scientific PubMed abstracts between 1974 and 2014: retrospective analysis. *Br. Med. J.* **351**, h6467 (2015).
13. Merton, R. K. Priorities in scientific discovery: a chapter in the sociology of science. *Am. Sociol. Rev.* **22**, 635–659 (1957).
14. Strevens, M. The role of the priority rule in science. *J. Philos.* **100**, 55–79 (2003).
15. Darwin, C. *To Charles Lyell. 3 May [1856]* (Darwin Correspondence Project, 1856); https://www.darwinproject.ac.uk/letter/DCP-LETT-1866.xml
16. Fang, F. C. & Casadevall, A. Competitive science: is competition ruining science? *Infect. Immun.* **83**, 1229–1233 (2015).
17. Makel, M. C., Plucker, J. A. & Hegarty, B. Replications in psychology research: how often do they really occur? *Perspect. Psychol. Sci.* **7**, 537–542 (2012).
18. Hagstrom, W. O.Competition in science. *Am. Sociol. Rev.* **39**, 1–18 (1974).
19. Hill, R. & Stein, C. *Scooped! Estimating Rewards for Priority in Science* Working Paper (Massachusetts Institute of Technology, 2019).

20. Balietti, S., Goldstone, R. L. & Helbing, D.Peer review and competition in the Art Exhibition Game. *Proc. Natl Acad. Sci. USA* **113**, 8414–8419 (2016).
21. Dechenaux, E., Kovenock, D. & Sheremeta, R. M. A survey of experimental research on contests, all-pay auctions and tournaments. *Exp. Econ.* **18**, 609–669 (2015).
22. Gneezy, U., Niederle, M. & Rustichini, A. Performance in competitive environments: gender differences. *Q. J. Econ.* **118**, 1049–1074 (2003).
23. Tiokhin, L. & Derex, M. Competition for novelty reduces information sampling in a research game—a registered report. *R. Soc. Open Sci.* **6**, 180934 (2019).
24. Yong, E. In science, there should be a prize for second place. *The Atlantic* (1 February 2018).
25. Romero, F. Novelty versus replicability: virtues and vices in the reward system of science. *Philos. Sci.* **84**, 1031–1043 (2017).
26. Cohen, B. A. Point of view: how should novelty be valued in science? *eLife* **6**, e28699 (2017).
27. Anderson, M. S., Ronning, E. A., De Vries, R. & Martinson, B. C. The perverse effects of competition on scientists' work and relationships. *Sci. Eng. Ethics* **13**, 437–461 (2007).
28. Phillips, N. D., Hertwig, R., Kareev, Y. & Avrahami, J. Rivals in the dark: how competition influences search in decisions under uncertainty. *Cognition* **133**, 104–119 (2014).
29. Higginson, A. D. & Munafò, M. R. Current incentives for scientists lead to underpowered studies with erroneous conclusions. *PLoS Biol.* **14**, e2000995 (2016).
30. Editors, T. P. B. S. The importance of being second. *PLoS Biol.* **16**, e2005203 (2018).
31. Marder, E. Scientific publishing: beyond scoops to best practices. *eLife* **6**, e30076 (2017).
32. Kiermer, V. & Heber, J. The importance of being second—PLOS-wide edition. *The Official PLOS Blog* https://theplosblog.plos.org/2020/04/the-importance-of-being-second-plos-wide-edition/ (2020).
33. Bobtcheff, C., Bolte, J. & Mariotti, T. Researcher's dilemma. *Rev. Econ. Stud.* **84**, 969–1014 (2017).
34. Heesen, R. Why the reward structure of science makes reproducibility problems inevitable. *J. Philos.* **115**, 661–674 (2018).
35. Smaldino, P. E., Turner, M. A. & Contreras Kallens, P. A. Open science and modified funding lotteries can impede the natural selection of bad science. *R. Soc. Open Sci.* **6**, 190194 (2019).
36. Bakker, M., van Dijk, A. & Wicherts, J. M. The rules of the game called psychological science. *Perspect. Psychol. Sci.* **7**, 543–554 (2012).
37. Ioannidis, J. P. Why most published research findings are false. *PLoS Med.* **2**, e124 (2005).
38. Van den Akker, O., Alvarez, L. D., Bakker, M., Wicherts, J. M. & van Assen, M. A. L. M. How do academics assess the results of multiple experiments? Preprint at *PsyArXiv* https://doi.org/10.31234/osf.io/xyks4 (2018).
39. McElreath, R. & Smaldino, P. E. Replication, communication, and the population dynamics of scientific discovery. *PLoS ONE* **10**, e0136088 (2015).
40. Nosek, B. A., Ebersole, C. R., DeHaven, A. C. & Mellor, D. T.The preregistration revolution. *Proc. Natl Acad. Sci. USA* **115**, 2600–2606 (2018).
41. Chambers, C. D. & Tzavella, L. Registered reports: past, present and future. Preprint at *MetaArXiv* https://doi.org/10.31222/osf.io/43298 (2020).
42. Bergstrom, C. T., Kerr, B. & Lachmann, M. in *Moral Markets: the Critical Role of Values in the Economy* 142–156 (Princeton Univ. Press, 2008).
43. Stephens, D. W. & Krebs, J. R. *Foraging Theory* (Princeton Univ. Press, 1986).
44. Camerer, C. Gifts as economic signals and social symbols. *Am. J. Sociol.* **94**, S180–S214 (1988).
45. Sozou, P. D. & Seymour, R. M. Costly but worthless gifts facilitate courtship. *Proc. R. Soc. B Biol. Sci.* **272**, 1877–1884 (2005).
46. Azar, O. H.The review process in economics: is it too fast? *South. Econ. J.* **72**, 482–491 (2005).
47. Azar, O. H. A model of the academic review process with informed authors. *B.E. J. Econ. Anal. Policy* **15**, 865–889 (2015).
48. Cotton, C. Submission fees and response times in academic publishing. *Am. Econ. Rev.* **103**, 501–509 (2013).
49. Heintzelman, M. & Nocetti, D.Where should we submit our manuscript? An analysis of journal submission strategies. *B.E. J. Econ. Anal. Policy* **9**, 1–28 (2009).
50. Leslie, D. Are delays in academic publishing necessary? *Am. Econ. Rev.* **95**, 407–413 (2005).
51. Gross, K. & Bergstrom, C. T. Contest models highlight inherent inefficiencies of scientific funding competitions. *PLoS Biol.* **17**, e3000065 (2019).
52. John, L. K., Loewenstein, G. & Prelec, D. Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychol. Sci.* **23**, 524–532 (2012).
53. Nosek, B. A. et al. Promoting an open research culture. *Science* **348**, 1422–1425 (2015).
54. Asendorpf, J. B. et al. Recommendations for increasing replicability in psychology. *Eur. J. Pers.* **27**, 108–119 (2013).
55. De Winter, J. & Happee, R.Why selective publication of statistically significant results can be effective. *PLoS ONE* **8**, e66463 (2013).
56. Van Assen, M. A., van Aert, R. C., Nuijten, M. B. & Wicherts, J. M. Why publishing everything is more effective than selective publishing of statistically significant results. *PLoS ONE* **9**, e84896 (2014).
57. Nelson, L. D., Simmons, J. P. & Simonsohn, U. Let's publish fewer papers. *Psychol. Inq.* **23**, 291–293 (2012).
58. Martinson, B. C. Give researchers a lifetime word limit. *Nature* **550**, 303 (2017).
59. Azar, O. H. The academic review process: how can we make it more efficient? *Am. Econ.* **50**, 37–50 (2006).
60. Button, K. S. et al. Power failure: why small sample size undermines the reliability of neuroscience. *Nat. Rev. Neurosci.* **14**, 365–376 (2013).
61. Calin-Jageman, R. J. & Cumming, G. The new statistics for better science: ask how much, how uncertain, and what else is known. *Am. Stat.* **73**, 271–280 (2019).
62. Lakens, D., Scheel, A. M. & Isager, P. M. Equivalence testing for psychological research: a tutorial. *Adv. Methods Pract. Psychol. Sci.* **1**, 259–269 (2018).
63. Wagenmakers, E.-J., Morey, R. D. & Lee, M. D. Bayesian benefits for the pragmatic researcher. *Curr. Dir. Psychol. Sci.* **25**, 169–176 (2016).
64. Merton, R. K. Science and technology in a democratic order. *J. Legal. Polit. Sociol.* **1**, 115–126 (1942).
65. Ioannidis, J. P. Why science is not necessarily self-correcting. *Perspect. Psychol. Sci.* **7**, 645–654 (2012).
66. Pashler, H. & Harris, C. R. Is the replicability crisis overblown? Three arguments examined. *Perspect. Psychol. Sci.* **7**, 531–536 (2012).
67. Weatherall, J. O. & O'Connor, C. Do as I say, not as I do, or, conformity in scientific networks. Preprint at *PhilSci-Archive* http://philsci-archive.pitt.edu/16035/ (2019).
68. Akerlof, G. A. & Michaillat, P.Persistence of false paradigms in low-power sciences. *Proc. Natl Acad. Sci. USA* **115**, 13228–13233 (2018).
69. Weatherall, J. O., O'Connor, C. & Bruner, J. How to beat science and influence people: policy makers and propaganda in epistemic networks. Preprint at *arXiv* https://arxiv.org/abs/1801.01239 (2018).
70. Zollman, K. J. The epistemic benefit of transient diversity. *Erkenntnis* **72**, 17–35 (2010).
71. Dohmen, T. & Falk, A. Performance pay and multidimensional sorting: productivity, preferences, and gender. *Am. Econ. Rev.* **101**, 556–590 (2011).
72. Lezzi, E., Fleming, P. & Zizzo, D. J. *Does it Matter Which Effort Task You Use? A Comparison of Four Effort Tasks When Agents Compete for a Prize.* Working Paper (Univ. East Anglia, 2015).
73. Tiokhin, L. *Improving the Reliability and Generalizability of Scientific Research.* Doctoral dissertation. (Arizona State Univ., 2018).
74. Borsboom, D. Theoretical amnesia. *Open Science Collaboration Blog* http://osc.centerforopenscience.org/2013/11/20/theoretical-amnesia/ (2013).
75. Scheel, A. M., Tiokhin, L., Isager, P. M. & Lakens, D. Why hypothesis testers should spend less time testing hypotheses. *Persp. Psychol. Sci.* https://doi.org/10.1177/1745691620966795 (2020).
76. Millikan, R. A. *The Electron and the Light-Quant from the Experimental Point of View* Nobel Lecture (Nobel Media, 1924); https://www.nobelprize.org/prizes/physics/1923/millikan/lecture/
77. Champely, S. et al. Package 'pwr'. R package version 1(2). ftp://www.r-project.org/pub/R/web/packages/pwr/pwr.pdf (2018).
78. R Core Development Team. *R: A Language and Environment for Statistical Computing* (R Foundation for Statistical Computing, 2017).
79. Park, J.-H. et al. Estimation of effect size distribution from genome-wide association studies and implications for future discoveries. *Nat. Genet.* **42**, 570–575 (2010).
80. Wilson, B. M. & Wixted, J. T. The prior odds of testing a true effect in cognitive and social psychology. *Adv. Methods Practices Psychol. Sci.* https://doi.org/10.1177/2515245918767122 (2018).
81. Morgan, T. J. H., Rendell, L. E., Ehn, M., Hoppitt, W. & Laland, K. N. The evolutionary basis of human social learning. *Proc. R. Soc. B Biol. Sci.* **279**, 653–662 (2011).
82. Rendell, L. et al. Cognitive culture: theoretical and empirical insights into social learning strategies. *Trends Cogn. Sci.* **15**, 68–76 (2011).
83. Brischoux, F. & Angelier, F. Academia's never-ending selection for productivity. *Scientometrics* **103**, 333–336 (2015).
84. Van Dijk, D., Manor, O. & Carey, L. B. Publication metrics and success on the academic job market. *Curr. Biol.* **24**, R516–R517 (2014).

Acknowledgements

We are grateful to N. Walasek, P. Smaldino, K. Panchanathan, W. Frankenhuis, D. Hruschka, R. Boyd, C. Bergstrom, D. Lakens, N. Coles, A. Scheel, N. van Dongen, F. Anvari, P. Isager, M. Maier, M. Macleod, P. Forscher, members of ABC&S at Arizona State University and members of the DEEP laboratory at Radboud University for feedback and comments on previous versions of this manuscript and model. We also thank E. Young for help with plotting and T. Lubiana for conducting a code review. This research was supported by VIDI grant 452-17-013 from the Netherlands Organisation for Scientific Research and (in part) by the Fetzer Franklin Fund of the John E. Fetzer Memorial Trust. The funders had no role in study design, data collection and analysis, decision to publish or preparation of the manuscript.

Author contributions

L.T. and M.Y. developed the initial idea. L.T. and T.J.H.M. developed the idea into its current form, programmed the model and analysed the results. L.T. and T.J.H.M. wrote the manuscript with feedback from M.Y.

Competing interests

The authors declare no competing interests.

Additional information

Supplementary information is available for this paper at https://doi.org/10.1038/s41562-020-01040-1.

Correspondence and requests for materials should be addressed to L.T.

Peer review information *Nature Human Behaviour* thanks Patrick Forscher and the other, anonymous, reviewer(s) for their contribution to the peer review of this work. Primary Handling Editor: Charlotte Payne.

Reprints and permissions information is available at www.nature.com/reprints.

Publisher's note Springer Nature remains neutral with regard to jurisdictional claims in published maps and institutional affiliations.

Corresponding author(s): Leonid Tiokhin

Last updated by author(s): Dec 17, 2020

Reporting Summary

Nature Research wishes to improve the reproducibility of the work that we publish. This form provides structure for consistency and transparency in reporting. For further information on Nature Research policies, see our Editorial Policies and the Editorial Policy Checklist.

Statistics

For all statistical analyses, confirm that the following items are present in the figure legend, table legend, main text, or Methods section.

n/a	Confirmed	
☒	□	The exact sample size (n) for each experimental group/condition, given as a discrete number and unit of measurement
☒	□	A statement on whether measurements were taken from distinct samples or whether the same sample was measured repeatedly
☒	□	The statistical test(s) used AND whether they are one- or two-sided *Only common tests should be described solely by name; describe more complex techniques in the Methods section.*
☒	□	A description of all covariates tested
☒	□	A description of any assumptions or corrections, such as tests of normality and adjustment for multiple comparisons
☒	□	A full description of the statistical parameters including central tendency (e.g. means) or other basic estimates (e.g. regression coefficient) AND variation (e.g. standard deviation) or associated estimates of uncertainty (e.g. confidence intervals)
☒	□	For null hypothesis testing, the test statistic (e.g. F, t, r) with confidence intervals, effect sizes, degrees of freedom and P value noted *Give P values as exact values whenever suitable.*
☒	□	For Bayesian analysis, information on the choice of priors and Markov chain Monte Carlo settings
☒	□	For hierarchical and complex designs, identification of the appropriate level for tests and full reporting of outcomes
☒	□	Estimates of effect sizes (e.g. Cohen's d, Pearson's r), indicating how they were calculated

Our web collection on statistics for biologists contains articles on many of the points above.

Software and code

Policy information about availability of computer code

Data collection	The agent-based model was programmed in R version 3.4.4.
Data analysis	R version 3.4.4 was used to generate the figures and conduct the supplementary analyses.

For manuscripts utilizing custom algorithms or software that are central to the research but not yet described in published literature, software must be made available to editors and reviewers. We strongly encourage code deposition in a community repository (e.g. GitHub). See the Nature Research guidelines for submitting code & software for further information.

Data

Policy information about availability of data

All manuscripts must include a data availability statement. This statement should provide the following information, where applicable:

- Accession codes, unique identifiers, or web links for publicly available datasets
- A list of figures that have associated raw data
- A description of any restrictions on data availability

No empirical data were generated during the current study. Simulation results are available on the Open Science Framework repository (https://osf.io/cbftz/).

Field-specific reporting

Please select the one below that is the best fit for your research. If you are not sure, read the appropriate sections before making your selection.

□ Life sciences ☒ Behavioural & social sciences □ Ecological, evolutionary & environmental sciences

For a reference copy of the document with all sections, see nature.com/documents/nr-reporting-summary-flat.pdf

Behavioural & social sciences study design

All studies must disclose on these points even when the disclosure is negative.

Study description	Agent-based model
Research sample	Not applicable
Sampling strategy	Not applicable
Data collection	Not applicable
Timing	Not applicable
Data exclusions	None
Non-participation	Not applicable
Randomization	Not applicable

Reporting for specific materials, systems and methods

We require information from authors about some types of materials, experimental systems and methods used in many studies. Here, indicate whether each material, system or method listed is relevant to your study. If you are not sure if a list item applies to your research, read the appropriate section before selecting a response.

Materials & experimental systems

n/a	Involved in the study
☒	□ Antibodies
☒	□ Eukaryotic cell lines
☒	□ Palaeontology and archaeology
☒	□ Animals and other organisms
☒	□ Human research participants
☒	□ Clinical data
☒	□ Dual use research of concern

Methods

n/a	Involved in the study
☒	□ ChIP-seq
☒	□ Flow cytometry
☒	□ MRI-based neuroimaging