The natural selection of good science

Alexander J. Stewart[1,*] and Joshua B. Plotkin[2,*]

[1] School of Mathematics and Statistics, University of St Andrews, St Andrews, KY16 9SS, United Kingdom

[2] Department of Biology, University of Pennsylvania, Philadelphia, PA, USA

[*] E-mail: ajs50@st-andrews.ac.uk; jplotkin@sas.upenn.edu

Scientists in some fields are concerned that many published results are false. Recent models predict selection for false positives as the inevitable result of pressure to publish, even when scientists are penalized for publications that fail to replicate. We model the cultural evolution of research practices when labs are allowed to expend effort on theory – enabling them, at a cost, to identify hypotheses that are more likely to be true, prior to empirical testing. Theory can restore high effort in research practice and suppress false-positives to a technical minimum, even without replication. The mere ability to choose between two sets of hypotheses – one with greater prior chance of being correct – promotes better science than can be achieved with effortless access to the set of stronger hypotheses. Combining theory and replication can have synergistic effects. Based on our analysis we propose four simple recommendations to promote good science.

Scientists are concerned about the state of science[1]. There is ample evidence to suggest that in some fields a large portion of reported results may be false[2,3,4,5,6,7,8]. The quality and magnitude of empirical evidence for this concern varies across disciplines and is a matter of debate[9]. But there is widespread acceptance that "researcher degrees of freedom" – such as flexibility in study design, measurement, and reporting – can lead to a high rate of false-positive reports. The dominant view holds that, in some fields, a sizable portion of published findings are false – a viewpoint publicized so widely that the lay person may reasonably be suspicious of the scientific enterprise. To remedy this situation, there have been remarkable efforts to fund and undertake large-scale replication studies to help identify errors in the literature and understand how they arise from current scientific practice[7,10,11,12]. Among other approaches, such as pre-registration[8,9,13,14,15] and novel funding mechanisms[16,17], a balance between testing new hypotheses and replicating published studies is now prescribed as a matter of course[9].

At the same time, models for the cultural evolution of scientific practice have suggested that replication efforts will not suffice to arrest the inevitable trend towards increasing false-positive rates – the evolution of "bad science" – driven by incentives to publish positive results regardless of their veracity[18,19,20,21]. Several authors have instead called for increased attention to theory as key to restoring a healthy scientific practice[18,22,23,24]. Whether, or how, theory will actually promote good science – that is, reduce the rate of false-positive reports – has not been studied in a formal framework. Moreover, models of cultural evolution used to interrogate the value of replication have been investigated primarily by simulation, without systematic mathematical analysis. Here we work to address both of these outstanding issues in the meta-scientific literature.

There are two ways that theory can aid scientific inquiry. When a field of research is underpinned by a well developed body of theory, the community of scientists can focus on those hypotheses that are more important or have a greater prior chance of being correct. That is, theory can give all researchers easier access to "stronger" hypotheses. At the same time, even in fields where a theoretical framework is not yet well developed or widely accepted, an individual lab that expends effort to model the system they are researching will generate stronger hypotheses by clarifying and quantifying their intuitions and by weeding out unlikely or illogical hypotheses. We show that this latter process – individual labs expending effort to select stronger hypotheses – has profound consequences for the cultural evolution of scientific practice.

Our analysis generalizes earlier models for the evolution of scientific practice in response to pressure to publish positive results. In particular, we extend earlier work by analyzing the possibility that individual labs may expend effort on "theory" to improve the quality of the hypotheses they choose to test. We analyze our model both mathematically and by simulation, showing that the pressure to publish does *not* produce an inevitable decline in the quality of science provided effort can be expended on theory. Rather, the system becomes bi-stable: it can support either high-quality science (low rates of false-positive reports) aided by theory, or a decline towards low-quality science and minimal effort. We quantify the basins of attraction towards these two different outcomes. Then we show how interventions such as replication can facilitate the stability of good-science equilibria. Finally we offer four simple recommendations, arising from our analysis, to promote good science in the face of pressures to publish.

Results

Methods from cultural evolution can be applied to study the development of research practices in response to institutional incentives[17,18,19,25,26]. This approach rests on the idea that competing research groups vary in methodological traits that affect their success and that are "heritable" either by differential imitation[27] or by differential production of students who then form their own labs, adopting the practices of their mentors.

Model of Efficacy and Effort

In order to study the natural selection of good science we adapt the model of Smaldino *et al.*[18], which characterizes a research lab in terms of its "efficacy" and "effort". Efficacy and effort are treated as traits that can evolve in the population of labs via a process of natural selection. Together these traits determine the rate at which a given lab generates novel results for publication, which is a natural proxy for success (i.e. fitness) in the face of inter-lab competition and the pressure to publish positive results.

Efficacy in this context refers to the overall ability of a lab to generate positive results. Efficacy encompasses the entire process of obtaining funding, designing experiments, executing studies, and producing a publication. Increasing a lab's efficacy also increases its rate of false positives, unless effort is exerted[18]. Effort here is a measure of a lab's degree of conservatism and rigor, which reduces both the false positive rate and the true positive rate. Increasing effort decreases the productivity of a lab, because it takes longer to perform rigorous research. (Note that Smaldino *et al.*[18] referred to efficacy as "power," which we avoid because of potential confusion with the familiar concept of statistical power.)

Under this model the process of producing a research paper proceeds as follows: (i) A lab selects a hypothesis to test. (ii) If the hypothesis is in fact true, the lab identifies it as such with a probability $P(+|T)$ – the true positive rate – which depends on the efficacy of the lab's techniques and the effort exerted to test the hypothesis. However, if the hypothesis is in fact false the lab mis-identifies it as true with a probability $P(+|F)$ – the false positive rate – which again depends on the efficacy of the lab's techniques and on the effort they exert. (iii) If the hypothesis is labelled as true the work is published – that is, we assume that only positive results are published[18].

Note that the false *discovery* rate – i.e. the rate at which false hypotheses are published as true – is $P(F)P(+|F)$, that is the chance of first selecting a hypothesis that is false and then incorrectly labelling it as true. Similarly the true discovery rate is $P(T)P(+|T)$.

We assume that both true and false positive rates increase with a lab's efficacy, V, and decrease with the lab's effort, e. We also assume $V \in [0, 1]$ and $e \in [1, \infty)$[18]. We choose the following functional forms for the rates of true and false positives in terms of effort e and efficacy V,

$$\begin{aligned} P(+|T) &= \frac{V}{\gamma} \times \frac{\gamma e}{1+\gamma(e-1)} \\ P(+|F) &= \frac{V}{\theta} \times \frac{1+(\theta-1)e}{1+(\theta-V)(e-1)}. \end{aligned} \tag{1}$$

According to this formulation, the true positive rate increases linearly with efficacy, whereas it is a convex decreasing function of effort (see Supplementary Figure 1). The false positive rate is a convex increasing function of efficacy[18]; but this can be counterbalanced by increasing effort. Increasing efficacy always increases publication rate[18], namely the rate of positive findings, whereas increasing effort decreases the discovery rate as labs become more conservative and meticulous.

Our formulation for the rate of true and false positives generalizes the model of Smaldino *et al.*[18]. The two formulations are identical in the limit $\theta = \gamma = 1$ and $V = 1$. In general, however, our formulation differs in an important way: effort expended to reduce false positives also has the effect of reducing true positives (for $\gamma > 1$), whereas Smaldino *et al.*[18] assumed the true positive rate is independent of effort. This more general formulation avoids a pathology that was present in earlier work: the tautological limit of $P(+|T) \to 1$ and $P(+|F) \to 1$ occurs only when efficacy is maximized ($V \to 1$) *and* effort is minimized ($e \to 1$) under our model. This tautological limit corresponds to a situation where a lab simply labels all hypotheses as true, and so it should occur only when a lab expends minimal effort.

Our formulation, in which true and false positive rates are both convex decreasing functions of effort, also generalizes Smaldino *et al.*[18] in the limit of maximum effort, $e \to \infty$. This limit produces $P(+|T) \to 1/\gamma$ and $P(+|F) \to 1/\theta$, where the parameters $1/\theta$ and $1/\gamma$ define the technical limits on true and false positive rates in a given field of research. These parameters describe the current technical limits of scientific practice, including limitations of current measurement technology, as well as constraints such as available funding and feasible sample size, etc. The values of γ and θ describe the current state of technical development of a field, and we treat them as fixed for most of our analysis. In reality, however, technical limits change as technology develops and resources fluctuate. By treating γ and θ as fixed, we are assuming that the overall development of technology in a field is slow compared to the evolution of individual lab practices.

Model of Hypothesis Selection

The rate at which a lab discovers positive results depends on the true and false positive rates (Eq. 1) as well as the underlying probability that a hypothesis the lab selects to test is true, $P(T)$. One way to imagine science is as a "grab bag" of hypotheses, each of which is true with a fixed probability b. We might imagine scientists as reaching into the bag, eyes closed, and drawing a hypothesis which they then test[18].

For many labs though, hypothesis selection is itself a product of effort. This effort may consist of broad engagement with the prior literature, which highlights some hypotheses as more plausible than others based on consistency with established results across many fields of science. Alternatively, it may consist of a lab expending effort to produce models and theory, which enable the production of systematic and self-consistent predictions that can be tested as empirical hypotheses.

In order to describe the process of putting effort into hypothesis selection we assume

$$P(T) = \frac{b_0 + b_1(e-1)}{e}. \tag{2}$$

Under this formulation, we may think of there being two different "bags" of hypotheses. In the first bag, hypotheses are true with probability b_0, whereas in the second they are true with probability $b_1 > b_0$. Whether a lab selects a hypothesis from the first bag or second bag depends on its level on effort. In particular, the probability of drawing a hypothesis from the first bag is $1/e$ (see Methods). And so effort $e \geq 1$ expended on hypothesis selection increases the prior probability that a selected hypothesis is true from the baseline rate b_0, when $e = 1$, to a maximum value $b_1 > b_0$, achieved when a lab puts maximum effort ($e \to \infty$) into the development of theory and engagement with prior literature (Figure 1).

Model of Publication and Replication

To study the impact of replication on the cultural evolution of scientific practice, we assume that each lab can choose to replicate a published study, at rate r, rather than attempting to produce a novel study[18]. The outcome of each replication attempt depends on the standing body of published literature (see Methods). Replication outcomes can be analyzed concisely under the simplifying assumption that labs all experience replication of their work at the same rate. We analyze this case mathematically, and we later show via simulation that our analytic results are good approximations even when this assumption is relaxed.

We assume that a lab publishes novel results at an overall rate ρ,

$$\rho = (1 - \eta \log_{10}(e)) \times (1 - r) \times [P(T)P(+|T) + P(F)P(+|F)], \tag{3}$$

where the term $(1 - \eta \log_{10}(e))$ describes the time it takes to produce a piece of research using effort level e. This logarithmic form reflects the choice made by[18]. In the SI (Section 1.8) we consider other functional relationships between effort and time to produce research, and we show that our qualitative results are robust to this choice. The term $P(T)P(+|T) + P(F)P(+|F)$ gives the overall discovery rate for novel results. Similarly labs engage in replication studies at rate

$$\phi = (1 - \eta \log_{10}(e)) \times r \tag{4}$$

where we assume that all replications are publishable regardless of outcome[18].

Adaptive Dynamics of Science

We can analyze the natural selection of good science via the payoffs associated with publication of novel results and replication of previous results. We first analyse the evolution of scientific practice under the simplifying assumptions of adaptive dynamics. In this framework an infinite population of labs are assumed to use identical strategies, and the success of a new strategy i, which differs slightly from the norm, is tested against the current resident strategy[28,29]. The expected fitness of a lab with a novel strategy i, denoted $w(e_i, V_i, r_i)$, is approximated by (see Methods):

$$w(e_i, V_i, r_i) = \rho_i B_N + \phi_i B_r + \frac{1}{2}\frac{\rho_i \phi}{l}(p_i B_{O+} - q_i C_{O-}) \tag{5}$$

where B_N is the payoff for publishing a novel result, B_r is the payoff for publishing a replication study, B_{O+} the payoff for having another lab successfully reproduce your work, and C_{O-} the cost of having another lab fail to reproduce your work (Figure 1). Eq. 5 then simply describes the payoff received by a lab given their current practices and the practices of the field: the first term $\rho_i B_N$ describes the payoff from lab i publishing novel results; the second term $\phi_i B_r$ describes the payoff from lab i publishing replication studies; and the term $\frac{1}{2}\frac{\rho_i \phi}{l}$ approximates the rate at which lab i has their results replicated by other labs (see SI), while $p_i B_{O+}$ and $q_i C_{O-}$ describe the benefits and costs for those replications being successful or unsuccessful. Here l is the ratio of published material being considered for replication in the corpus of the field to the number of active labs. (Thus if $l = 10$, there are 10 times as many published works being considered for replication on a topic as there are active labs working on that topic.) Finally, p_i and q_i give the probability that a replication attempt by another lab on a study produced by lab i is successful or unsuccessful, respectively (see Methods).

We use the framework of adaptive dynamics[28,29] to determine the equilibria associated with the evolution of scientific practice, for a population of labs with fitness described by Eq. 5. Under this framework the equilibria of the system occur when the selection gradient is zero, i.e. when

$$\frac{\partial w}{\partial e_i}\Big|_{e_i=e} = 0$$
$$\frac{\partial w}{\partial r_i}\Big|_{r_i=r} = 0 \tag{6}$$

Note that $\frac{\partial w}{\partial V_i}\Big|_{V_i=V} > 0$ for all V, which means selection always favors increasing V, and so labs necessarily evolve to maximum methodological efficacy, $V = 1$ under all circumstances, as in previous work[18]. Although Eq. 6 cannot in general be solved analytically (see SI section 1), it can be systematically explored numerically to identify stable equilibria and their basins of attraction.

Theory produces good science

When a lab cannot improve hypothesis selection by effort, then science will evolve to a state where labs simply label all novel hypotheses as true – that is, the evolution of bad science[18]. As we show below, however, the mere act of expending effort to find stronger hypotheses is sufficient to stabilize good science. We define good science as an equilibrium in the cultural evolution of lab practice that maintains a false positive rate close to the technical minimum, $P(+|F) \sim V/\theta$. Under our model this can occur only when effort is high (Eq. 1).

The act of expending effort to find stronger hypotheses is described by Eq. 2, where minimum effort ($e = 1$) results in selection of a hypothesis with prior probability $P(T) = b_0$ and maximum effort ($e \to \infty$) results in a hypothesis with prior probability $P(T) = b_1 > b_0$. That is to say, a lab can expend effort to identify stronger hypotheses. In practice, this type of effort typically involves theoretical work – by either formal modeling or leveraging an informal, conceptual framework – in order to identify hypotheses that have a greater prior chance of being correct.

Expending effort to find stronger hypotheses produces good science, whereas simply having effortless access to stronger hypotheses does not (Figure 2). The figure shows the results of simulations in three different regimes: (i) only weak hypotheses available ($b_0 = b_1 = 0.01$) (ii) only strong hypotheses available ($b_0 = b_1 = 0.25$) and (iii) choice, via effort, between weak and strong hypotheses ($b_0 = 0.01$ and $b_1 = 0.25$). In the first two cases bad science evolves, with effort declining to its minimum and true and false positive rates increasing to unity, which replicates the simulation results of Smaldino *et al.*[18]. However in the third case, when effort can be expended to select stronger hypotheses, we find something quite different. As labs evolve, effort *increases* from its initial value to a level that maintains a high true-positive rate and a low false-positive rate – the evolution of good science.

Notably, expending effort to select strong hypotheses produces a good-science equilibrium even when effortless access to equally strong hypotheses would lead to bad science (Figure 2c versus Figure 2b).

How does expending effort on hypothesis selection promote good science? Analysis of our model by adaptive dynamics (Eqs. 5-6 and SI section 1) shows that when effort can be expended to find stronger hypotheses the system becomes bi-stable (Supplementary Supplementary Figure 2-5). The bad-science equilibrium identified by[18] always exists, but once a tipping point is reached, another equilibrium emerges that features high effort and a low false positive rate. For a broad range of parameters the basin of attraction towards this good-science equilibrium is much larger than the basin of attraction towards the bad-science equilibrium (see Supplementary Figure 2-5). The reason why increasing effort can be advantageous is that greater effort results in a greater probability of selecting a true hypothesis to test in the first place, $P(T)$. Once the good-science equilibrium is reached, decreasing effort tends to reduce the overall rate of publication, because it makes hypotheses less likely to be true a priori; and the lab still puts effort into assessing the veracity of each hypothesis, so that they end up identifying more hypotheses as false, thereby reducing publication rate. This phenomenon, which opposes reduction in effort, is sufficient to stabilize good science.

Replication can facilitate good science

We have seen that effort expended at hypothesis selection – that is, theoretical work in advance of any empirical experimentation – can lead to the evolution of good scientific practices that ensure low false positive rates. Now we consider the additional effects of replication on the evolution of scientific practice. Unlike effort and efficacy, which evolve endogenously in response to incentive structures, the rate of replication can be increased or decreased exogenously by introducing institutional incentives or policies that require replication[9]. And so much of the debate over how to promote good science has been focused on encouraging replication and similar interventions[9,15].

Replication can help weed out bad science by re-testing published results and flagging the false positives. By imposing a cost on labs who publish false positives, replication reduces the incentive for labs to lazily label novel results as true without expending the effort to rigorously test them. However, previous models for the evolution of scientific practice have found that replication cannot prevent the natural selection of bad science[18]. We too find that, in the absence of theory to enable hypothesis selection, replication alone does not produce good science (Figure 4). However we also find (Figures 4 and S7-S10) that, in the presence of theory, replication can both increase the basin of attraction of good science and interact synergistically with stronger hypotheses and better methodology to stabilize good science. Figure 4 shows examples where the introduction of replication ($r > 0$) can make the difference between the evolution of bad versus good science.

Instead of fixing the replication rate, held in place by an external policy, we can alternatively study the case when labs choose their own degree of replication effort. To do this we analyze the co-evolution of effort and replication rate. Using the framework of adaptive dynamics (Eqs. 5-6) we find that, when the cost for studies that fail to replicate is large ($C_{O-} \gg 1$), both good- and bad-science equilibria persist, but replication is always lost (see SI Section 1.2 and Supplementary Figure 2). Individual-based simulations produce similar results: in combination with theory, replication rates evolve to low positive values and good science is maintained whereas without theory, replication alone cannot help to prevent the natural selection of bad science (SI Section 2.4). On the other hand, when replication occurs at a fixed rate by an external policy, it can dramatically expand the basin of attraction of good science (SI Section 1.4 and Figure 3).

Attention-grabbing hypotheses

Our model of hypothesis selection (Eq. 2) assumes that hypotheses are drawn from two pools that differ only in their prior probability of being true, i.e. $b_0 < b_1$. In reality, however, different types of novel results may generate different benefits, often depending on the effort spent on generating them. In particular, low-effort attention-grabbing hypotheses that seem surprising may be expected to generate more "hype" and therefore more benefit for the lab if successfully published, than carefully constructed high-effort hypotheses that build on prior work and have a great *a priori* chance of being true. To capture this effect we now assume that a positive report for a low-effort hypothesis, with prior probability of being true b_0, generates benefit B_N^0 whereas as positive result for a high-effort hypothesis with prior probability of being true $b_1 > b_0$, generates a smaller benefit $B_N^1 \leq B_N^0$. The probability of choosing a particular hypothesis to test is given by Eq. 2 as previously (see SI Section 1.6).

A scientific culture that rewards publication of a low-effort, attention-grabbing hypothesis more than publication of a high-effort hypothesis ($B_N^0 > B_N^1$) threatens to undermine the evolution of good scientific practice. Indeed, we find that setting $B_N^0 > B_N^1$ reduces the size of basin of attraction towards the good-science equilibrium. Nonetheless, a stable, good-science equilibrium persists even when the reward for publishing an attention-grabbing hypothesis is roughly twice as large as the reward for publishing a high-effort conservative hypotheses (Supplementary Figure 6). And so evolution can still promote labs that expend effort at hypothesis selection, provided the rewards for publication are not too heavily biased towards low-effort findings.

Good science across fields

The emergence of good science as a stable response to the pressure to publish depends on the extent to which a field has developed and on the costs and benefits associated with publication in that field. In terms of methodology, stable good science depends on a field's development along three major axes. (i) A field must have achieved a sufficient degree of technological advancement ($1/\theta$ sufficiently small, Supplementary Figure 2-3). That is, if low rates of false positives cannot possibly be achieved even through high effort, then good science cannot be maintained. (ii) Labs must have sufficient ability to discriminate between strong and weak hypotheses (b_1 sufficiently larger than b_0, Supplementary Figure 2-3). That is, good science cannot be maintained if a field does not yet have sufficient theory to enable the selection of stronger hypotheses through effort. (iii) Good science can be stabilized when labs undertake replication ($r > 0$, Figures 3 and 4), which can help make up for less technical advancement (θ) or less theory (b_1), but this is not always guaranteed (Supplementary Figure 3-5 and SI section 1.4) and the efficacy of replication depends on the relative size of the corpus of literature to the number of active labs (Supplementary Figure 7-10).

Methodological advancement and the ability to identify strong hypotheses varies widely across fields, as do norms regarding the costs and benefits of publication. We can assess the likely impact of interventions, such as increasing the frequency of replications, by calculating the likelihood that a good-science equilibrium is supported across a wide range of methodologies. In Figure 4 a-b we systematically vary parameters associated with a field's norms and technical limits (b_0, b_1, θ, γ, B_N^0, B_N^1) and shows the associated likelihood that a stable good-science equilibrium exists, across a range of different replication rates, and benefits and costs of replication. As this parameter sweep reveals, replication is indeed an effective tool for promoting the viability of good science when paired with high costs to a lab, incurred when a publication fails to replicate.

We also assessed whether better methods can make up for mediocre theory (Figure 4 c-d), in which we again systematically varied parameters associated with a field's norms and technical limits (b_0, b_1, r, γ, B_N^0, B_N^1) and calculated the likelihood of a good science equilibrium for different values of θ, which we use as a proxy for a field's level of technical advancement (since higher θ leads to lower technical limits on the rate of false positives). We find that increasing θ leads to greater viability of good science, all other things being equal.

Discussion

Scientific practice is amenable to scientific study. We have developed models of cultural evolution to study how theory influences research effort and methodological efficacy for labs under pressure to publish. The ability to expend theoretical effort on hypothesis selection produces bi-stable dynamics: evolution will lead either to high-effort labs that publish reports with few false positives (good science), or alternatively to minimal-effort labs that try to get ahead by publishing results replete with false positives, (bad science). Our mathematical analysis delineates when the good science equilibrium will arise, in terms of the payoffs for publication, the field's technical limits on true- and false-positive rates, the payoffs associated with replication efforts, and the extent to which theory can improve hypothesis selection in the field.

Our results highlight the role of theoretical effort in shaping scientific practice. Theory is construed broadly in this analysis, to include any activity that identifies hypotheses with a greater chance of being true, prior to empirical investigation. In some fields of science theory is pursued using a formal mathematical or computational framework[30]; whereas in other fields theory is an informal conceptual framework used for systematic, logical synthesis of the literature. In all of its various forms, theoretical effort has the effect of winnowing down the set of hypotheses that are likely to be correct, prior to empirical testing.

The history of science provides many illustrative examples of the value of theoretical effort. Physics in particular has produced many striking cases, such as the development of quantum electrodynamics (QED). QED is a prototype for the role of formal, mathematical theory in refining hypotheses before further experimentation. Here purely mathematical developments were required to produce internally consistent predictions for, e.g., the magnetic moment of an electron – predictions that were later verified to 11 significant digits by experiment[31]. But theory has been central to the development of many other fields beyond physics. In bio-medical science, Hodgkin and Huxley's[32] quantitative model for action potentials predicted the gating structure of ion channels, later verified by MacKinnon[33]. More importantly, the theoretical framework of Hodgkin and Huxley structured a productive feedback loop between hypothesis selection and experimentation throughout the development of electrophysiology[34]. In the social sciences, the development of prospect theory[35] has shaped our understanding of imperfect rationality in decision making under risk. First inspired by empirical observations that violated rational choice, prospect theory was developed into a broad conceptual framework that has advanced specific predictions for controlled experiments in behavioral economics, as well as explanations for field data[36]. These three paradigmatic examples illustrate the general conclusion of our analysis, on the productive role of theoretical effort across a diverse range of disciplines.

Four recommendations to promote good science

The question remains what lessons can be drawn from our analysis to guide the evolution of scientific practice in fields where pessimism about replication failures and competitive publication practices dominate the conversation. Our analysis suggests four simple recommendations to promote the evolution of good science, which offer both optimism and caution for researchers concerned about the publication of false results.

1. **Put resources into developing a robust theoretical framework.** A theoretical framework that enables labs to distinguish strong hypotheses from weak ones, even at a cost, is sufficient to preserve good science (Figure 2). Providing resources targeted at theoretical work, especially in fields where formal theory is lacking, should be a priority. Crucially, the impact of stronger hypotheses on the evolution of scientific practice is non-linear. Theory-driven hypothesis selection reaches a tipping point: before the tipping point only bad science is possible, after the tipping point good science can be sustained (Supplementary Figure 3).

2. **Replicate, but don't rely on replication.** Replication alone, absent theory, does not produce good science (Figure 3 and Supplementary Figure 11-12), but it can interact synergistically with theory to stabilize good science across fields. However this may require substantial penalties when a study fails to replicate (Figure 4), and imposing such costs (100 or 1000 times the benefit for successful novel publication) would distort incentives and may produce unintended consequences for lab behavior not well captured by our model.

3. **Better methods *can* make up for mediocre theory.** There is a trade-off between the methodological efficacy, theoretical sophistication, and the rate of replication required to sustain good science (Supplementary Figure 3-5). A field that is more developed in one area can afford to be less developed in another (Figure 4), meaning better methods can make up for mediocre theory, to some extent[23]. Where theory reaches a dead end, focusing on developing better methods can still help a field reach the tipping point when good science becomes viable.

4. **Bad science is always a danger.** Even when a good-science equilibrium is available, a bad-science equilibrium remains an option. Low-effort, attention-grabbing publication of

any and all hypotheses is a stable equilibrium in all fields (SI Section 1.2-1.3), and this outcome is increasingly likely when scientific culture, including journal policies, excessively rewards attention-grabbing or gratuitously novel findings (Supplementary Figure 6 and 12). All fields, no matter their theoretical and technical sophistication, are at risk of succumbing to bad science.

Models of models

In our model for the evolution of scientific practice, increased effort makes almost everything harder: research takes longer and a lab is more conservative when labelling a hypothesis as true, both of which reduce the overall rate of publication. The only direct benefit of effort lies in selecting stronger hypotheses. And yet this effect is often sufficient to induce a qualitative change in the equilibrium outcome – namely, to stabilize good science in the face of pressure to publish.

Like all models, ours is a simplification and abstraction of what is, in reality, an incredibly complex process. The purpose of the model is to cut through the complexity whilst retaining the most salient forces at play when scientists make decisions about what to study and by what methodology and effort. The value of a mathematical or computational model over a verbal hypothesis is that it allows systematic exploration of how these fundamental forces play out, without relying on intuition alone.

To be truly useful, a model should tell us something that we did not know before we built it. In the context of scientific practice, we have seen that theory must provide new information about what constitutes a strong versus a weak hypothesis, in order to promote good science. A theoretical model whose output simply recapitulates the assumptions that went into building it is tautological, and it does not grant us any additional ability to distinguish between strong and weak hypotheses; it is wasted effort that does not help promote good science.

Our findings reinforce and justify calls made by several authors for more theoretical effort, particularly in the social sciences[18,22,23]. Our analysis makes no assumptions or prescriptions about what type of theory should be developed (e.g. statistical models, agent-based simulations, mathematical models etc.). Rather, we have considered any theoretical technique that improves hypothesis selection, and analyzed its influence on the cultural evolution of scientific practice.

The availability of theory that improves hypothesis selection will vary across fields, depending on the field's age and topic matter. While physicists have used mathematical models for centuries, biologists actively debated their utility in the mid-twentieth century[37,38] although that debate is now largely resolved[39,40]. Contemporary discussions of theory in other scientific fields are not dissimilar to the historical developments in physics and biology. What our analysis highlights, regardless of the discipline, is the tremendous potential for theoretical effort to alter the culture of scientific practice.

We also offer some optimistic results for those who lament the pressure to publish as corroding good science[41,42,43,44]. Such concerns have a long history[45] and an exponentially expanding scientific literature[46] poses profound challenges for researchers, even if the rate of false-positive reports is low. Yet our results show that pressure to publish, and competition between labs in general, can stimulate effort and produce excellent science provided the theoretical and empirical tools in a field are sufficiently well developed. It is only when theoretical tools are not yet developed, or go unused, that pressure to publish creates perverse incentives that lead to the evolution of bad science.

Methods

We analyze a model for the natural selection of scientific publication strategy under the framework of adaptive dynamics[28,29]. Within this framework we follow the basic assumptions of Smaldino *et al.*[18]: a lab's success is measured in terms of the number of publications and (un)successful replications of their work by other labs. We assume that labs "reproduce" by adopting the research strategies of other labs, chosen based on their past success. Under this framework we assume an infinite population of labs, each using the same resident publication strategy, and we perform an invasion analysis to determine which resident strategies are stable in the face of local "mutations" that perturb the resident research strategy. While the assumptions of adaptive dynamics are unrealistic in several important ways, they nonetheless allow us to systematically explore the qualitative behavior of the system, and our key finding from the analysis – that competition to publish can produce good science when accounting for the role of theory in selecting hypotheses – holds when relaxation these simplifying assumptions in individual-based simulations.

Lab life cycle

We consider a population of labs whose life cycle proceeds via a phase of publication followed by a phase of selection and reproduction, in which the current population is replaced with a population of new labs. This simplifying assumption allows us to assume that all labs are the same age during the selection phase, ignoring effects that arise due to older labs appearing more successful due to having had more time to publish. We relax this assumption in our simulations and show that it does not qualitatively alter our results.

As described in the Results section, a lab i produces novel results at a rate

$$\rho_i = (1 - \eta \log_{10}(e_i)) \times (1 - r_i) \times (P_i(T)P_i(+|T) + P_i(F)P_i(+|F)). \tag{7}$$

The probability that the hypothesis being tested is true is given by

$$P_i(T) = \frac{b_0 + b_1(e_i - 1)}{e_i}, \tag{8}$$

and $P_i(F) = 1 - P_i(T)$. Eq. 8 can be understood as a generalization of the "grab bag" model[18] in which a selected hypothesis is true with probability b_0. Eq. 8 describes a scenario in which there are two "types" of hypotheses. The weaker hypotheses are true with probability b_0 and make up a proportion $1/e$ of all hypotheses; whereas the stronger hypotheses are true with probability $b_1 > b_0$ and make up the remaining $(1 - 1/e)$ of all hypotheses. Thus by expending greater effort e a lab can alter the space of hypotheses to which they have access. Once a hypothesis is selected, it is tested and the chance of a positive finding is described by the following equations:

$$
\begin{aligned}
P_i(+|T) &= \frac{V_i}{\gamma} \times \frac{\gamma e_i}{1+\gamma(e_i-1)} \\
P_i(+|F) &= \frac{V_i}{\theta} \times \frac{1+(\theta-1)e_i}{1+(\theta-V)(e_i-1)}.
\end{aligned} \tag{9}
$$

The behavior of Eq. 9 as a function of effort is shown in Supplementary Figure 1.

Labs produce replication studies at rate

$$\phi_i = (1-\eta\log_{10}(e_i)) \times r_i \tag{10}$$

where $(1-\eta\log_{10}(e_i))$ describes the time it takes to complete a study. A lab carrying out a replication study of an original report produced by another lab i, successfully reproduces the original finding with probability

$$p_{ij} = \frac{P_i(T)P_i(+|T)P_j(+|T) + P_i(F)P_i(+|F)P_j(+|F)}{P_i(T)P_i(+|T) + P_i(F)P_i(+|F)}, \tag{11}$$

while they produce a different finding to lab i with probability

$$q_{ij} = \frac{P_i(T)P_i(+|T)(1-P_j(+|T)) + P_i(F)P_i(+|F)(1-P_j(+|F))}{P_i(T)P_i(+|T) + P_i(F)P_i(+|F)}. \tag{12}$$

From Eqs. 11-12 we define $p_i = \frac{1}{N-1}\sum_{j\neq i} p_{ij}$ and $q_i = \frac{1}{N-1}\sum_{j\neq i} q_{ij}$, the probability of successful and unsuccessful replication attempts for lab i by the rest of the population.

To model the production of publications during a lab's life we use a system of ordinary differential equations, where $x_n^i(t)$ denotes the number of novel results that have been produced at time t by lab i and $x_r^i(t)$ the number or replication studies published by lab i. We also define $z^i(t)$ as the number of novel studies produced by lab i that have been replicated by other labs at time t. Under these assumptions the dynamics of publication are as follows

$$
\begin{aligned}
\frac{dx_n^i}{dt} &= \rho_i \\
\frac{dx_r^i}{dt} &= \phi_i \\
\frac{dz^i}{dt} &= \sum_{j \neq i} \frac{x_n^i - z^i}{L} \phi_j
\end{aligned}
\tag{13}
$$

where L is the size of the corpus of published materials available for replication, which is assumed for simplicity to be fixed. Following the standard assumptions of adaptive dynamics[28,29], we consider the fitness of a lab i in a monomorphic population such that $\phi_j = \phi$. If we set the number of publications at time $t = 0$ to zero, the distribution of publications for a lab i at time t is given by

$$
\begin{aligned}
x_n^i(t) &= \rho_i t \\
x_r^i(t) &= \phi_i t \\
z^i(t) &= (N-1)\left(\mathrm{e}^{-\phi t/L} - 1 + \frac{\phi}{L} t\right) \frac{L\rho_i}{\phi}
\end{aligned}
\tag{14}
$$

We assume that the lifespan of each lab is one time unit, such that the integral must be evaluated at $t = 1$. This corresponds to a scenario in which there are many more publications in the corpus of literature for a field than can be replicated in the lifetime of a lab, i.e. $L \gg 1$. By Taylor expansion of $z^i(t)$ in terms of L^{-1} and neglecting terms $O\left(L^{-2}\right)$ and higher we recover

$$
\begin{aligned}
x_n^i &= \rho_i \\
x_r^i &= \phi_i \\
z^i &= (N-1)\frac{1}{2}\frac{\phi\rho_i}{L} + O\left(L^{-2}\right).
\end{aligned}
\tag{15}
$$

Taking the limit $N \to \infty$, $L \to \infty$ and $L/N \to l$ we recover

$$\begin{aligned} x_n^i &= \rho_i \\ x_r^i &= \phi_i \\ z^i &= \frac{1}{2}\frac{\phi\rho_i}{l} \end{aligned} \tag{16}$$

which gives us the expression for fitness used in the Results section (Eq. 5). Further details of the invasion analysis for this model are given in the SI.

Individual-based Simulations

In addition to mathematical analysis by adaptive dynamics, we also perform Monte Carlo simulations in polymorphic, finite populations of size N, where lab strategies replicate according to a copying process[27]. We assume that science is produced according to Eqs. 7-10 and that replication can occur once for any study present in the corpus, which has absolute size L. Labs are assumed to become inactive when they adopt a new strategy, which may be thought of as retirement of a senior professor and replacement by a new hire. When a new lab is formed we assume that mutations perturb effort e, efficacy V, and replication rate r. Mutational perturbations are drawn uniformly from $[-0.01, 0.01]$, and mutations occur at rate μ_e, μ_V and μ_r respectively (see SI for full details).

In the limit $\gamma = \theta = 1$, where our model coincides with Smaldino *et al.*[18], simulations reproduce the finding[18] that bad science evolves in the absence of theory (Supplementary Figure 5).

Data availability

All scripts data to reproduce the results are available at 10.5281/zenodo.4616768

Code availability

All scripts necessary to reproduce the results are available at 10.5281/zenodo.4616768

Acknowledgements

We thank Paul Smaldino and four anonymous referees for constructive feedback. The authors received no specific funding for this work.

Author contributions

A.J.S. and J.B.P. conceived the project and developed the model. A.J.S. ran the simulations and analysed the model with input from J.B.P. A.J.S. and J.B.P. wrote the paper.

Competing interests

The authors declare no competing interests.

References

1. Nissen, S. B., Magidson, T., Gross, K. & Bergstrom, C. T. Publication bias and the canonization of false facts. *Elife* **5**, e21451 (2016).

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

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

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

5. John, L. K., Loewenstein, G. & Prelec, D. Measuring the prevalence of questionable research practices with incentives for truth telling. *Psychological science* **23**, 524–532 (2012).

6. Simonsohn, U., Nelson, L. D. & Simmons, J. P. P-curve: a key to the file-drawer. *Journal of experimental psychology: General* **143**, 534 (2014).

7. Rahal, R., Collaboration, O. S. *et al.* Estimating the reproducibility of psychological science. *Science* **349**, aac4716 (2015).

8. Begley, C. G. & Ioannidis, J. P. Reproducibility in science: improving the standard for basic and preclinical research. *Circulation research* **116**, 116–126 (2015).

9. Munafò, M. R. *et al.* A manifesto for reproducible science. *Nature Human Behaviour* **1**, 0021 (2017).

10. Klein, R. A. *et al.* Many labs 2: Investigating variation in replicability across samples and settings. *Advances in Methods and Practices in Psychological Science* **1**, 443–490 (2018).

11. Ebersole, C. R. *et al.* Many labs 3: Evaluating participant pool quality across the academic semester via replication. *Journal of Experimental Social Psychology* **67**, 68–82 (2016).

12. Camerer, C. F. *et al.* Evaluating the replicability of social science experiments in nature and science between 2010 and 2015. *Nature Human Behaviour* **2**, 637 (2018).

13. Nosek, B. A. *et al.* Promoting an open research culture. *Science* **348**, 1422–1425 (2015).

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

15. Munafò, M. R. & Davey Smith, G. Robust research needs many lines of evidence. *Nature* **553**, 399–401 (2018).

16. Gross, K. & Bergstrom, C. T. Contest models highlight inherent inefficiencies of scientific funding competitions. *PLoS biology* **17** (2019).

17. Smaldino, P. E., Turner, M. A. & Contreras Kallens, P. A. Open science and modified funding lotteries can impede the natural selection of bad science. *Royal Society open science* **6**, 190194 (2019).

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

19. Grimes, D. R., Bauch, C. T. & Ioannidis, J. P. A. Modelling science trustworthiness under publish or perish pressure. *R Soc Open Sci* **5**, 171511 (2018).

20. Devezer, B., Nardin, L. G., Baumgaertner, B. & Buzbas, E. O. Scientific discovery in a model-centric framework: Reproducibility, innovation, and epistemic diversity. *PloS one* **14**, e0216125–e0216125 (2019).

21. Szollosi, A. *et al.* Is preregistration worthwhile? *Trends Cogn Sci* **24**, 94–95 (2020).

22. Muthukrishna, M. & Henrich, J. A problem in theory. *Nature Human Behaviour* **3**, 221–229 (2019).

23. Smaldino, P. Better methods can't make up for mediocre theory. *Nature* **575**, 9 (2019).

24. van Rooij, I. & Baggio, G. Theory before the test: How to build high-verisimilitude explanatory theories in psychological science. *Perspect Psychol Sci* 1745691620970604 (2021).

25. McElreath, R. & Smaldino, P. E. Replication, communication, and the population dynamics of scientific discovery. *PLoS One* **10**, e0136088 (2015).

26. O'Connor, C. The natural selection of conservative science. *Stud Hist Philos Sci* **76**, 24–29 (2019).

27. Traulsen, A., Nowak, M. A. & Pacheco, J. M. Stochastic dynamics of invasion and fixation. *Phys Rev E Stat Nonlin Soft Matter Phys* **74**, 011909 (2006).

28. Mullon, C., Keller, L. & Lehmann, L. Evolutionary stability of jointly evolving traits in subdivided populations. *Am Nat* **188**, 175–95 (2016).

29. Leimar, O. Multidimensional convergence stability. *Evolutionary Ecology Research* **11**, 191–208 (2009).

30. Gray, C. T. & Marwick, B. Truth, proof, and reproducibility: There's no counter-attack for the codeless. In Nguyen, H. (ed.) *Statistics and Data Science*, 111–129 (Springer Singapore, Singapore, 2019).

31. Feynman, R. P. *QED: the strange theory of light and matter* (Princeton University Press, Princeton, N.J., 1985).

32. Hodgkin, A. L. & Huxley, A. F. A quantitative description of membrane current and its application to conduction and excitation in nerve. *J Physiol* **117**, 500–44 (1952).

33. MacKinnon, R. Nobel lecture. potassium channels and the atomic basis of selective ion conduction. *Biosci Rep* **24**, 75–100 (2004).

34. Schwiening, C. J. A brief historical perspective: Hodgkin and huxley. *The Journal of physiology* **590**, 2571–2575 (2012).

35. Kahneman, D. & Tversky, A. Prospect theory: An analysis of decision under risk. *Econometrica* **47**, 263–291 (1979).

36. Barberis, N. C. Thirty years of prospect theory in economics: A review and assessment. *Journal of Economic Perspectives* **27**, 173–96 (2013).

37. Mayr, E. Where are we? *Cold Spring Harbor Symp Quant Biol* **24**, 1–14 (1959).

38. Haldane, J. B. S. A defence of beanbag genetics. *Perspectives in Biology and Medicine* **7**, 343–359 (1964).

39. Ewens, W. J. Commentary: On haldane's 'defense of beanbag genetics'. *Int J Epidemiol* **37**, 447–51 (2008).

40. Crow, J. F. Mayr, mathematics and the study of evolution. *Journal of Biology* **8** (2009).

41. Sarewitz, D. The pressure to publish pushes down quality. *Nature* **533**, 147 (2016).

42. Rawat, S. & Meena, S. Publish or perish: Where are we heading? *Journal of research in medical sciences : the official journal of Isfahan University of Medical Sciences* **19**, 87–89 (2014).

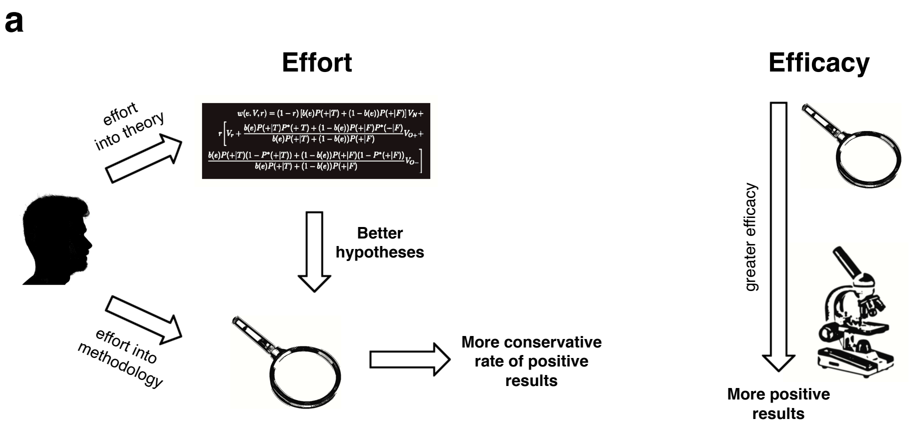
43. Dinis-Oliveira, R. J. & Magalhães, T. The inherent drawbacks of the pressure to publish in health sciences: Good or bad science. *F1000Research* **4**, 419–419 (2015).

44. Kurt, S. Why do authors publish in predatory journals? *Learned Publishing* **31**, 141–147 (2018).

45. Price, D. J. D. S. *Little Science, Big Science* (New York: Columbia University Press, 1963).

46. Bornmann, L. & Mutz, R. Growth rates of modern science: A bibliometric analysis based on the number of publications and cited references. *Journal of the Association for Information Science and Technology* **66**, 2215–2222 (2015).

Figure Captions



b

Parameter	Meaning	Default value
b_0	Probability of a weak hypothesis being true	0.01
b_1	Probability of a strong hypothesis being true	0.25
B_N	Benefit for novel publication	1
B_r	Benefit for publishing replication	0.2
B_{O+}	Benefit for being successfully replicated	0.1
C_{O-}	Cost for failure to successfully replicate	100
$1/\theta$	Technical limit on false positive rate	1/25
$1/\gamma$	Technical limit on true positive rate	10/11

Figure 1: **How can a lab do better science?** a) Science can be made better in two basic ways: 1) A lab can expend more effort, which means (all other things equal) that the lab selects a hypothesis with a higher prior probability of being correct and that, at the same time, the lab is more conservative about testing the hypothesis. Increased effort is associated with theoretical work to select hypotheses that have greater prior likelihood of being correct, as well as more conservative procedures for testing these hypotheses. 2) A lab can develop more effective methods, which means (all other things equal) that the rate of positive results increases. Increased efficacy is associated with greater measurement precision, larger sample sizes, or simply more funding, for example. b) Our model includes eight parameters that describe the technical state of a field and the costs and benefits associated with publication and replication. The technical limits $1/\theta$ and $1/\gamma$ describe the false and true positive rates that are achieved by a lab using methods of maximum available efficacy, $V = 1$, and maximum effort $e \to \infty$.

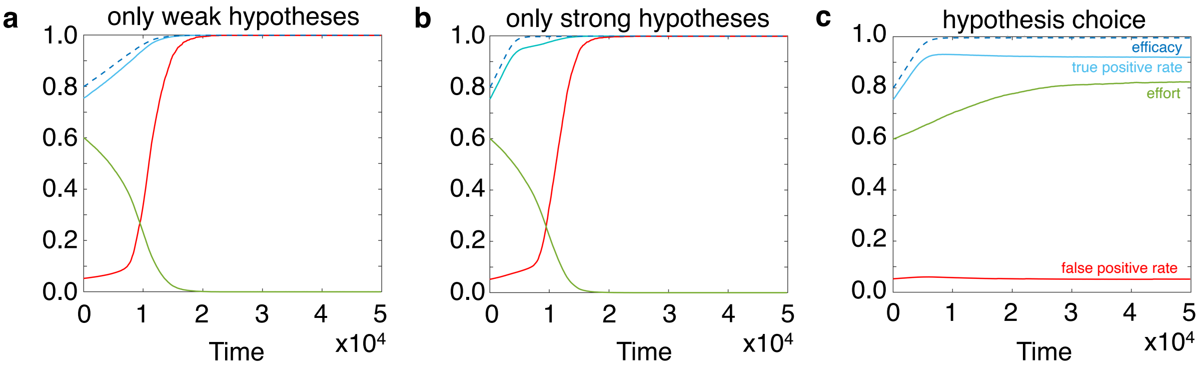


Figure 2: **The evolution of good science:** We ran individual-based simulations in which $N = 100$ labs compete to publish positive results, in the absence of replication. In each panel we plot the trajectories of efficacy V (dashed blue line), true positive rate $P(+|T)$ (solid blue line), false-positive rate $P(+|F)$ (red line), and effort, re-scaled as $(e-1)/e$ so that values lie $[0, 1]$ (green line). a) When only weak hypotheses are available ($b_0 = b_1 = 0.01$) efficacy increases over time, but effort declines, so that the population evolves to a bad-science equilibrium in which the true and false positive rates both evolve to 1 – that is, all hypotheses are labelled as true. b) The same is true when only strong hypotheses are available ($b_0 = b_1 = 0.25$). c) When effort can be put into choosing between weak and strong hypotheses ($b_0 = 0.01$ and $b_1 = 0.25$) a stable, good-science equilibrium emerges, and effort and efficacy both increase, leaving the false positive rate close to the technical minimum $P(+|F) \sim 1/\theta$. The figures show the mean trajectories over an ensemble of 10^3 replicate simulations. The rate of publication for each lab was determined by Eqs. 1-3; mutations occurred to effort e and efficacy V at rate $\mu_e = \mu_V = 0.01$. Mutational perturbations to efficacy were drawn uniformly from the range $[-0.01, 0.01]$, and effort was assumed to change by ±1 upon mutation. Cultural evolution occurred via a copying process (see SI), payoffs were set at $B_N = 1$, with $\gamma = 1.1$ and $\theta = 25$, with no replication ($r = 0$)

.

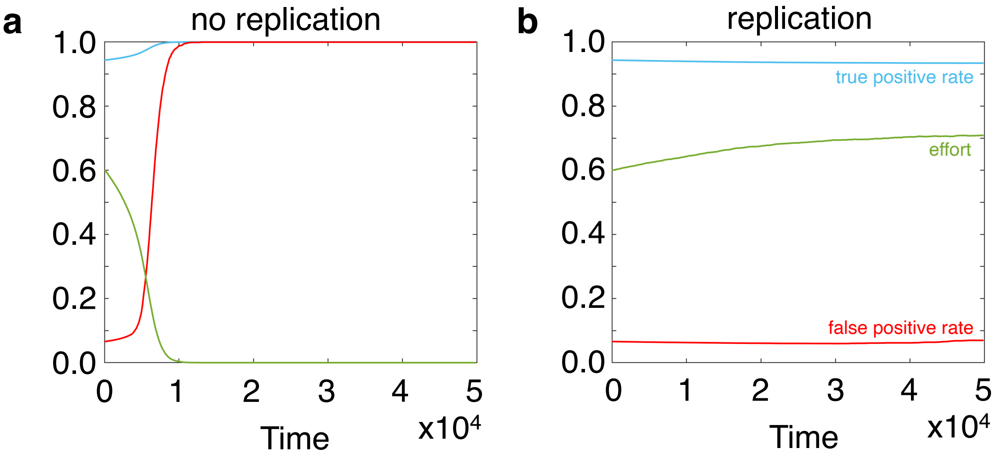


Figure 3: **Synergy between replication and theory.** The figure shows results of individual-based simulations for the evolution of scientific practice with and without replication. In the regime $1/\theta = 0.04$, shown here, theory and replication are both required to produce good science, as predicted by mathematical analysis by adaptive dynamics (Figure 3a). (a) In the absence of replication, both true (blue) and false (red) positive rates increase to unity, and effort declines to a minimum $(e-1)/e = 0$, i.e. $e = 1$ (green). b) However, when replication occurs at a rate $r = 0.1$, effort increases over time towards a good-science equilibrium in which false positives are rare. All parameters are the same as in Figure 2c, except for θ. Replications are chosen from a corpus of $L = 10^5$ novel studies, and each study is allowed to be replicated only once (see SI). Payoffs are $B_N = 1$, $B_r = 0.2$, $B_{O+} = 0.1$ and $C_{O-} = 100$.

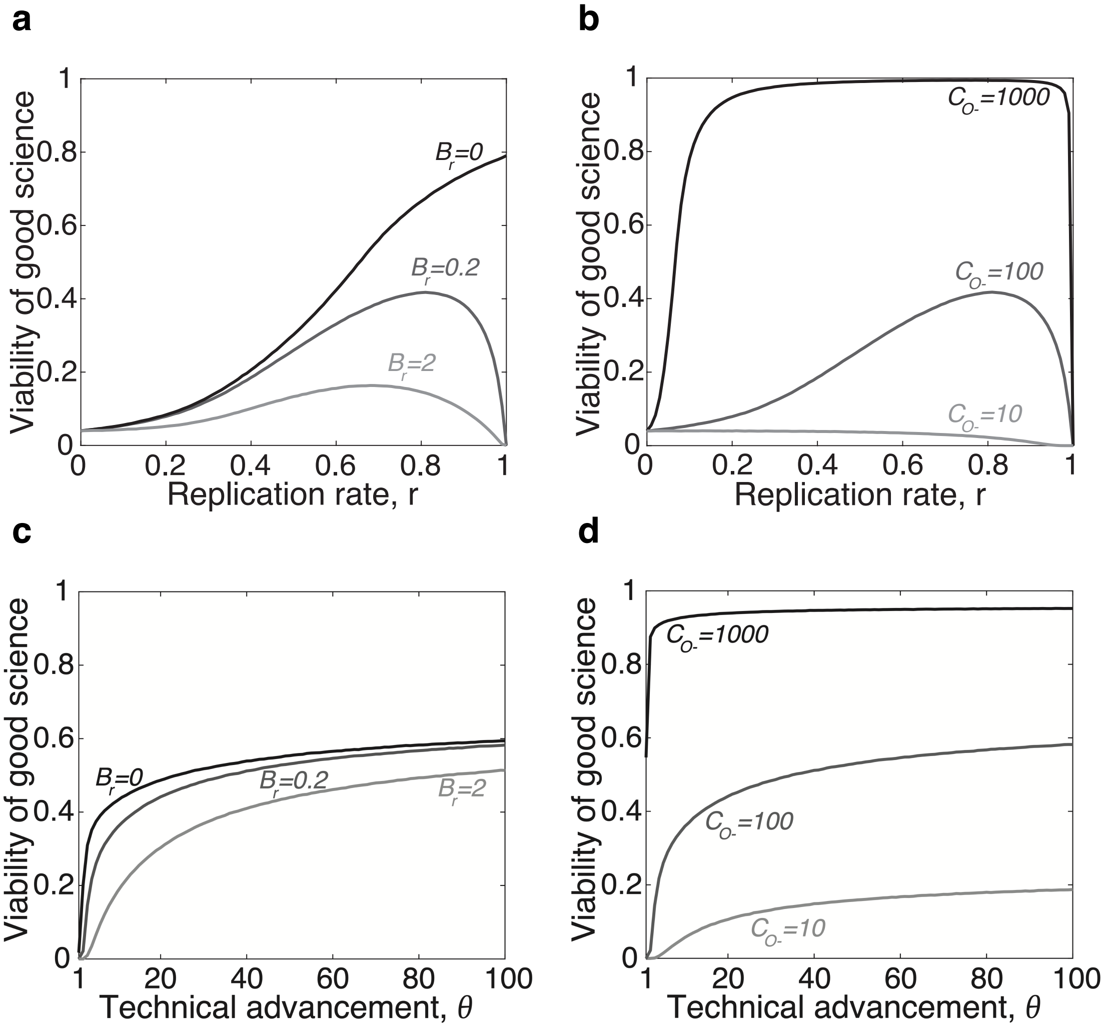


Figure 4: **Viability of good science across fields.** The figure shows the proportion of parameter sets which support a stable good-science equilibrium, as a function of the replication rate r (a and b) and level of technical advancement, θ (c and d), for different costs and benefits of publication. In all cases studied, we see that introducing replication $r > 0$ initially increases the viability of good science. But this only occurs up to a point: when replication rates are very high, and replication studies are beneficial, there is comparatively less reward for effort spent at hypothesis selection and novel research. On the other hand, we see that increasing the technical advancement of a field $\theta > 1$ always increases the viability of good science. (a and c) Larger benefits for replication B_r tend to reduce the viability of good science. This is because the benefit for performing a replication study is awarded independent of effort, which reduces the relative benefit of effort spent at hypothesis selection. (b and d) Increased costs to a lab of failure to have their study replicated C_{O-} increase the viability of good science. This is because false discoveries, although initially beneficial when published, become extremely costly if they are later flagged in a replication study. For each curve, we drew 10^7 parameter sets for every value of r at increments of 0.01 between 0 and 1. We chose parameters from the following ranges: $b_0 \in [0, 1]$, $b_1 \in [b_0, 1]$, $\theta \in [2, \infty)$, $\gamma \in [1, 2]$, $B_N^0 \in [1, 2]$, $r \in [0, 1]$. Unless otherwise indicated we fixed $B_N^1 = 1$, $B_r = 0.2$, $B_{O+} = 0.1$, $C_{O+} = 100$, $\eta = 0.2$ and $l = 10$. The effects of varying l, B_{O+} and η are shown in Figure S7-S8 and are qualitatively similar to panel b).

The natural selection of good science: Supplementary Information

Alexander J. Stewart[1,*] and Joshua B. Plotkin[2,*]

[1] School of Mathematics and Statistics, University of St Andrews, St Andrews, KY16 9SS, United Kingdom
[2] Department of Biology and Department of Mathematics, University of Pennsylvania, Philadelphia, PA, USA
[*] E-mail: ajs50@st-andrews.ac.uk; jplotkin@sas.upenn.edu

Contents

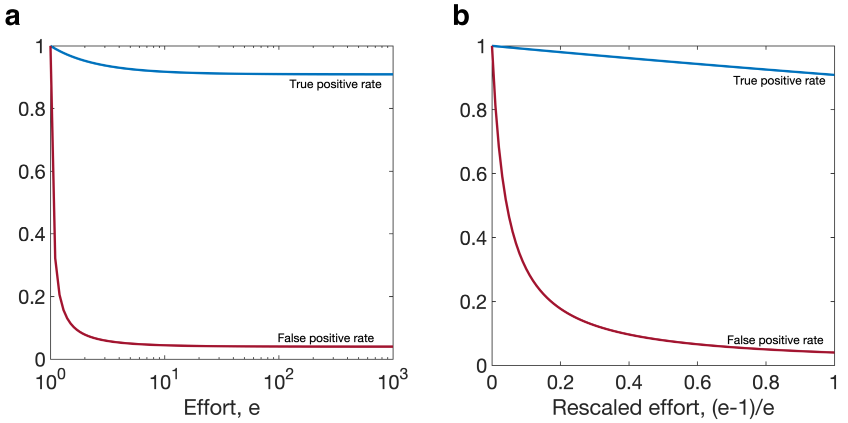
In this supplement we provide derivations for the equations in the main text, along with additional analysis and simulation results to demonstrate the robustness of our findings to relaxation of model assumptions. Note that equation numbers continue from the main text.

1 Adaptive dynamics model of publication

We consider a population of labs whose life cycle proceeds via a phase of publication followed by a phase of selection and reproduction, in which the current population is replaced with a population of new labs as described in the Methods section.

1.1 Re-scaled effort

Throughout we use "re-scaled effort" in our figures, where the re-scaled effort is simply $e^* = (e-1)/e$. Plotting e^* allows us to visulaize effort on the interval $[0, 1]$ rather than $[1, \infty)$. A comparison of the True and False positive rates (Eq. 9) as a function of effort e and re-scaled effort e^* is shown in Supplementary Figure 1 below.



Supplementary Figure 1: Effort and rate of positive findings. Shown are the true (blue) and false (red) positive rates under the default parameters used in simulations, with efficacy at it's equilibrium level, $V = 1$. a) Effort e of the scale $[1, \infty)$, both true and false positive rates approach their technical limit, $1/\gamma$ and $1/\theta$ for $e \approx 10$. b) When effort is re-scaled to lie in $[0, 1]$ we see that these technical limits correspond to a high degree of effort

1.2 Invasion analysis

We now perform an invasion analysis for the model described in the Methods section of the main text.

Taking Eq. 16 as the publication distribution at the end of the publication cycle, we can then describe the fitness of a lab i against a monomorphic background of competing labs, following publication as

$$w(e_i, V_i, r_i) = \rho_i B_N + \phi_i B_r + \frac{1}{2}\frac{\rho_i \phi}{l}(p_i B_{O+} - q_i C_{O-}) \tag{17}$$

which we can write as

$$\begin{aligned}
&w(e_i, V_i, r_i) = \\
&(1-\eta\log_{10}(e_i)) \times (1-r_i)\Bigg[P_i(T)P_i(+|T)\left(B_N + \frac{\phi}{2l}B_{O+}P(+|T) - \frac{\phi}{2l}C_{O-}(1-P(+|T))\right) + \\
&P_i(F)P_i(+|F)\left(B_N + \frac{\phi}{2l}B_{O+}P(+|F) - \frac{\phi}{2l}C_{O-}(1-P(+|F))\right)\Bigg] + (1-\eta\log_{10}(e_i)) \times r_iB_r
\end{aligned} \tag{18}$$

We can now compute the selection gradient for the system. From Eq. 9 we immediately see that fitness is monotonically increasing in V_i thus we need only evaluate the gradient at $w(e_i, 1, r_i)$. This gives us

$$\begin{aligned}
s_e = \left.\frac{\partial w}{\partial e_i}\right|_{e_i=e, r_i=r} &= -\tfrac{\eta}{e\log[10]}\Bigg[(1-r)P(T)P(+|T)\alpha + (1-r)P(F)P(+|F)\beta + rB_r\Bigg] + \\
&\quad (1-\eta\log_{10}(e)) \times (1-r)\left[\tfrac{d(P_i(T)P_i(+|T))}{de_i}\alpha + \tfrac{d(P_i(F)P_i(+|F))}{de_i}\beta\right] \\
s_r = \left.\frac{\partial w}{\partial r_i}\right|_{e_i=e, r_i=r} &= -(1-\eta\log_{10}(e))\Bigg[P(T)P(+|T)\alpha + P(F)P(+|F)\beta\Bigg] + (1-\eta\log_{10}(e))B_r
\end{aligned} \tag{19}$$

where

$$\begin{aligned}
\alpha &= \left(B_N + \frac{\phi}{2l}B_{O+}P(+|T) - \frac{\phi}{2l}C_{O-}(1-P(+|T))\right) \\
\beta &= \left(B_N + \frac{\phi}{2l}B_{O+}P(+|F) - \frac{\phi}{2l}C_{O-}(1-P(+|F))\right)
\end{aligned} \tag{20}$$

with

$$\frac{d(P_i(T)P_i(+|T))}{de_i} = \frac{b_1 - \gamma b_0}{(1+\gamma(e_i-1))^2} \tag{21}$$

and

$$\frac{d(P_i(F)P_i(+|F))}{de_i} = -\frac{1}{1+(\theta-1)(e_i-1)}\frac{1}{\theta e_i} \times \left[(1+(\theta-1)e_i)\frac{(b_1-b_0)}{e_i} + (e_i(1-b_1)+(b_1-b_0))\left(\frac{\theta-1}{1+(\theta-1)(e_i-1)}\right)\right]. \tag{22}$$

From Eqs. 13-17 we can calculate the points at which the selection gradient vanishes, $(\hat{e}, \hat{r})$, which satisfy:

$$\frac{\eta}{\hat{e}\log[10]}\left[\left(\frac{b_0+b_1(\hat{e}-1)}{\hat{e}}\right)\left(\frac{\hat{e}}{1+\gamma(\hat{e}-1)}\right)\alpha + \left(1-\frac{b_0+b_1(\hat{e}-1)}{\hat{e}}\right)\left(\frac{1+(\theta-1)\hat{e}}{\theta(1+(\theta-V)(\hat{e}-1))}\right)\beta + \frac{\hat{r}}{1-\hat{r}}B_r\right] = (1-\eta\log_{10}(\hat{e})) \times \left[\left(\frac{b_1-\gamma b_0}{(1+\gamma(\hat{e}-1))^2}\right)\alpha - \frac{1}{1+(\theta-1)(\hat{e}-1)}\frac{1}{\theta\hat{e}} \times \left[(1+(\theta-1)\hat{e})\frac{(b_1-b_0)}{\hat{e}} + (\hat{e}(1-b_1)+(b_1-b_0))\left(\frac{\theta-1}{1+(\theta-1)(\hat{e}-1)}\right)\right]\beta\right] \tag{23}$$

where

$$\hat{r} = \frac{B_r - B_N(\hat{P}(T)\hat{P}(+|T) - \hat{P}(F)\hat{P}(+|F))}{(B_{O+}\hat{P}(+|T) - C_{O-}(1-\hat{P}(+|T)))\hat{P}(T)\hat{P}(+|T) + (B_{O+}\hat{P}(+|F) - C_{O-}(1-\hat{P}(+|F)))\hat{P}(F)\hat{P}(+|F)} \times \frac{2l}{(1-\eta\log_{10}(\hat{e}))}. \tag{24}$$

Eqs. 23-24 cannot be solved analytically in general and in particular Eq. 23 can produce multiple solutions in the physically relevant range. However we observe that the condition for any equilibrium to be convergent stable under all mutation matrices is that the 2×2 Jacobian matrix $\mathbf{J}$ for the system must have negative eigenvalues or, equivalently, be negative definite (Leimar, 2009) which in turn implies that $(\mathbf{J})_{rr} = \frac{\partial s_r}{\partial r} < 0$ must hold. This condition is satisfied only if

$$\begin{aligned} &P(T)P(+|T)(B_{O+}P(+|T) - C_{O-}(1-P(+|T))) + \\ &P(F)P(+|F)(B_{O+}P(+|F) - C_{O-}(1-P(+|F))) > 0. \end{aligned} \tag{25}$$

If we assume $C_{O-} \gg B_{O+}$, i.e. the penalties for publishing false results are very large, then Eq. 25 is only satisfied in the limit $P(+|T) \to 1$ and $P(+|F) \to 1$ which is the bad-science equilibrium. Thus under our model assumptions there are no points of zero selection gradient that are convergent stable except close to the bad-science equilibrium. This is consistent with our numerical analysis of the system (Supplementary Figure 2), under which we find only unstable singular points. However this result does not exclude the possibility that stable equilibria can arise at the boundaries of phase space.

1.3 Boundary behavior

Stable equilibria can arise at the boundary if the selection gradient perpendicular to the boundary points towards it, and the selection gradient parallel to the boundary is zero. We now explore the behavior of the system at the boundaries $r = 0$, $r = 1$, $(e-1)/e = 0$ and $(e-1)/e \to 1$ beginning with tje resident bad-science equilibrium of Smaldino and McElreath (2016) which corresponds to $(e-1)/e = r = 0$.

Bad science $(e = 1, r = 0)$**:** The bad-science equilibrium of Smaldino and McElreath (2016) arises at $(e = 1, r = 0)$. From Eqs. 19-20 the selection gradient at this point is

$$\begin{aligned} s_e(1,0) &= -\frac{\eta}{e\log[10]}B_N - (1-\eta\log_{10}(e)) \times \left[b_0(\gamma-1) + (1-b_0)(\theta-1)^2/\theta\right]B_N \\ s_r(1,0) &= -(1-\eta\log_{10}(e))(B_N - B_r) \end{aligned} \tag{26}$$

which is always negative, indicating that the bad-science equilibrium is always a stable state of the system provided the benefit of publishing a novel result is greater than that for publishing a replication, $B_N \geq B_r$.

Maximum replication $(r = 1)$**:** When replication rate is at its maximum, $r = 1$, the selection gradient parallel to the boundary, calculated from Eq. 19, is given by

$$s_e(e,1) = -\frac{\eta}{e\log[10]}B_r \tag{27}$$

which is always negative. Thus we need only evaluate the selection gradient perpendicular to the boundary at $(r = 1, e = 1)$ which, from Eq. 19 gives

$$s_r(1,1) \quad = \quad -\left[B_N + B_{O+}/l - B_r\right]$$

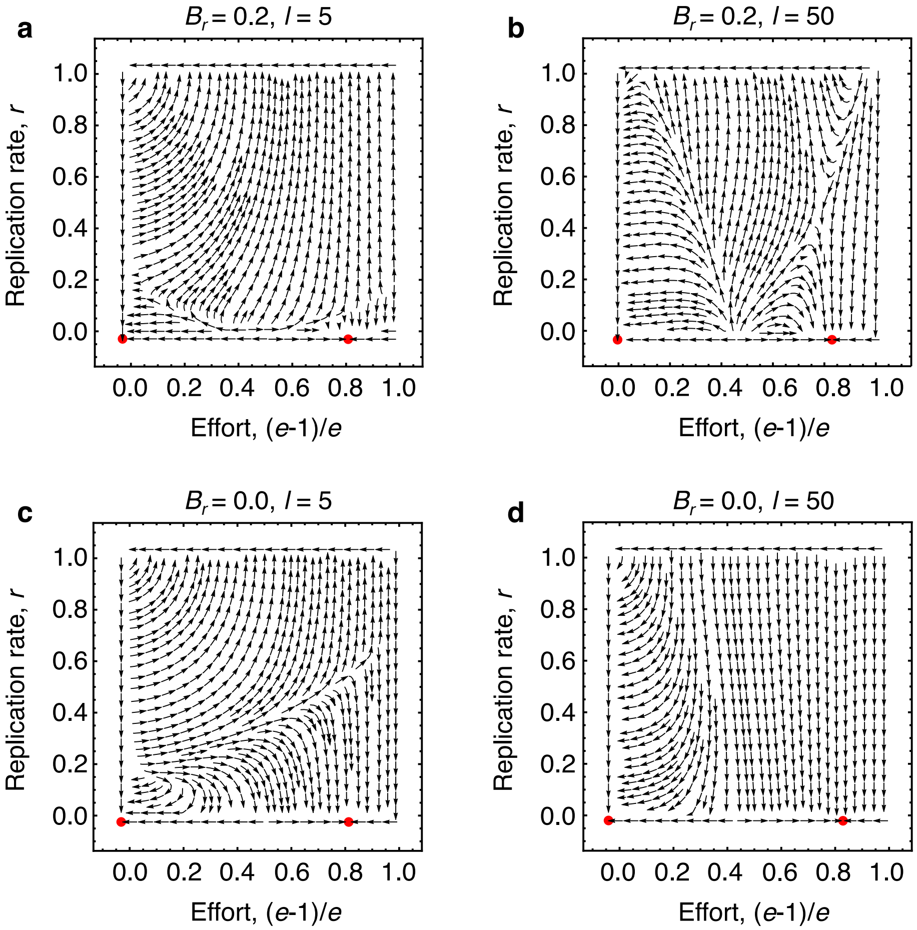
which, under our assumption $B_N \geq B_r$, is always negative. Thus there is no stable equilibrium with maximum replication.

Minimum replication $(r = 0)$**:** Finally we consider the behavior of the system when replication rate is minimized, $r = 0$. For Eqs. 19-20 we find selection gradient

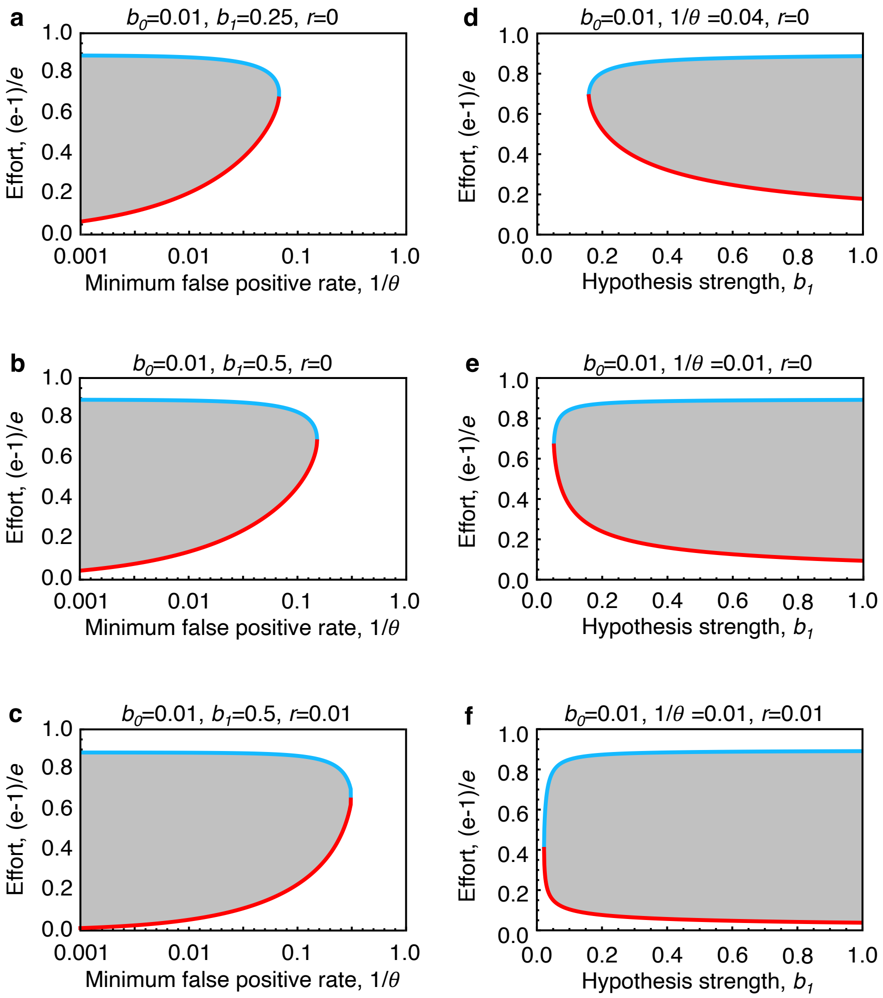
$$
\begin{aligned}
s_e(e,0) &= -\tfrac{\eta}{e\log[10]}\Bigg[(P(T)P(+|T)+P(F)P(+|F)\Bigg]B_N + \\
(1-\eta\log_{10}(e)) &\times \Bigg[\tfrac{d(P_i(T)P_i(+|T))}{de_i}+\tfrac{d(P_i(F)P_i(+|F))}{de_i}\Bigg]B_N \\
s_r(e,0) &= -(1-\eta\log_{10}(e))\Bigg[(P(T)P(+|T)+P(F)P(+|F))B_N - B_r\Bigg]
\end{aligned}
\tag{28}
$$

at the boundary, where Eq. 30 must be treated numerically as above. Eq. 30 is negative provided $(P(T)P(+|T)+P(F)P(+|F))B_N > B_r$. The term $(P(T)P(+|T)+P(F)P(+|F))$ is non-monotonic in e and thus, depending on the solution to Eq. 30 and the choice of B_r the boundary may be either stable or unstable. Crucially this means that the addition of replication to the evolutionary dynamics of the system may cause a stable, high-effort equilibrium to become unstable.

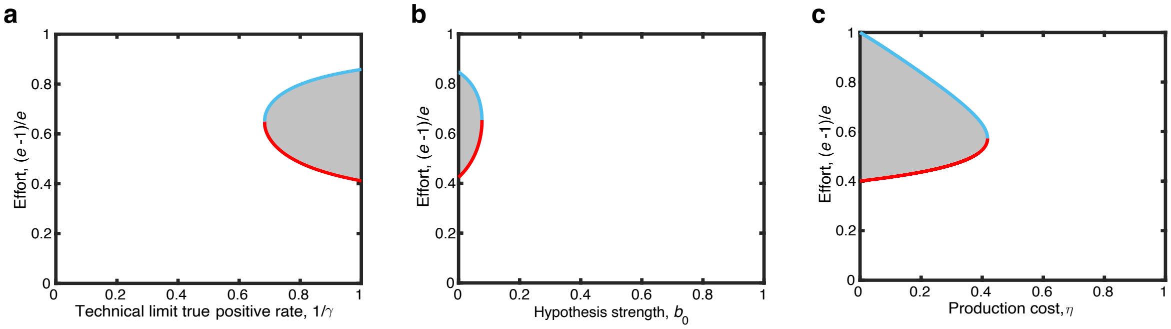
The resulting evolutionary trajectories of the system across a range of parameter are shown in Supplementary Figure 2 and the basin of attraction for the good- and bad-science equilibria in both the presence and absence of enforced replication are shown for different model parameters in Supplementary Figure 3-S5. Note that for each parameter varied in Supplementary Figure 3-S5 there is a “tipping point” at which good science becomes stable. Nonetheless the system remains bi-stable, with good and bad science equilibria coexisting. Thus two things are required for good science to emerge where it is absent: 1) The tipping point must be reached and 2) a perturbation must occur to push the system from the bad science to the good-science equilibrium.



Supplementary Figure 2: Co-evolution of e and r Phase portraits in the regime of adaptive dynamics for a) high benefits for replication, $B_r = 0.2$ and a small corpus of literature $l = 5$ b) high benefits for replication, $B_r = 0.2$ and a large corpus of literature $l = 50$ c) no benefit for replication, $B_r = 0.0$ and a small corpus of literature $l = 5$ b) no benefit for replication, $B_r = 0.0$ and a large corpus of literature $l = 50$. All other parameters are chosen as in Figure 3. The good-science equilibrium consisting of high effort and zero replication rates, always exists alongside the bad-science equilibrium at minimum effort and zero replication rate. We see that high levels of replication can undermine good science and pull the system back to the bad-science equilibrium . Both equilibria are marked with red dots. In all cases we assume that the costs for failed replication is high, $C_{O-} = 100$



Supplementary Figure 3: Analysis of equillibria by adaptive dynamics. The figure shows equilibrium publication strategies in a large population of labs, as a function of model parameters. Plotted in each panel are the locations of the stable (blue) and unstable (red) equilibria as a function of either the technical minimum false positive rate $1/\theta$ (left column) or the maximum achievable hypothesis strength b_1 (right column). For many parameter choices the system is bi-stable, with a good-science equilibrium indicated by the blue line and a bad-science equilibrium at minimum effort $(e-1)/e = 0$. In the gray regions selection favors increasing effort towards the good-science equilibrium; whereas in the white regions selection favors ever decreasing effort towards to bad-science equilibrium. a) For $b_0 = 0.01$ and $b_1 = 0.25$ and without replication ($r = 0$), stable good science requires a technical minimum true positive rate no greater than $1/\theta = 0.08$. b) With better theory, meaning the possibility of stronger hypotheses $b_1 = 0.5$, good science is stable with even lower methodological efficacy (e.g. $1/\theta > 0.1$). c) Adding replication at a low rate ($r = 0.01$) enables good science to be maintained for even larger values of $1/\theta$. Similar patterns occur when we fix $1/\theta$ and vary b_1 (right column): increasing methodological efficacy allows good science to emerge even with weaker hypotheses (panels d-e), and replication decreases the need for strong theory even further (panel f). Payoffs are set at $B_N = 1$, $B_r = 0.2$, $B_{O+} = 0.1$ and $C_{O-} = 100$ and $l = 5$.



Supplementary Figure 4: Analysis of equillibria by adaptive dynamics. The figure shows equilibrium publication strategies in a large population of labs, as a function of model parameters. Plotted in each panel are the locations of the stable (blue) and unstable (red) equilibria as a function of all five parameters of the system without replication. For many parameter choices the system is bi-stable, with a good-science equilibrium indicated by the blue line and a bad-science equilibrium at minimum effort $(e-1)/e = 0$. In the gray regions selection favors increasing effort towards the good-science equilibrium; whereas in the white regions selection favors ever decreasing effort towards to bad-science equilibrium. a) Impact of the technical limit true-positive rate $1/\gamma$ of the basin of attraction for good science. b) Impact of hypothesis strength b_0. c) Impact of the production cost of science η. Payoffs are set at $B_N = 1$, and $l = 1$. All other parameters are as in Supplementary Figure 3 unless otherwise specified in the panel.

1.4 Tipping points and changing technical limits

An important feature of our analysis is that a good-science equilibrium does not emerge gradually from a bad-science one, as conditions improve. Rather there is a “tipping point” at which the system becomes bi-stable, with a good- and bad-science equilibria coexisting under certain parameter regimes. For example, in Supplementary Figure 3 we may consider the x-axis as describing the overall level of methodological development of a field. As the minimum rate of false positives, $1/\theta$, declines, or the strength of theory-driven hypotheses, b_1, increases, a tipping point is reached beyond which a good-science equilibrium (blue line) exists. For the parameters in Supplementary Figure 3a, for example, this tipping point occurs when $1/\theta \approx 0.08$. This illustrates how changes in the technical limits or theoretical development of a field over time can lead to sudden improvement and the emergence of good science.

1.5 Replication as a policy

So far we have studied replication as an evolving trait, which labs can choose to engage in as a way to improve their success through publication. However replication of published research can, in principle at least, be implemented as policy, in which a proportion r of all published studies are replicated by an outside agency. To study replication as policy it is sufficient to set $B_r = 0$ and $r_i = 0$ in Eq. 18. We then retrieve selection gradient

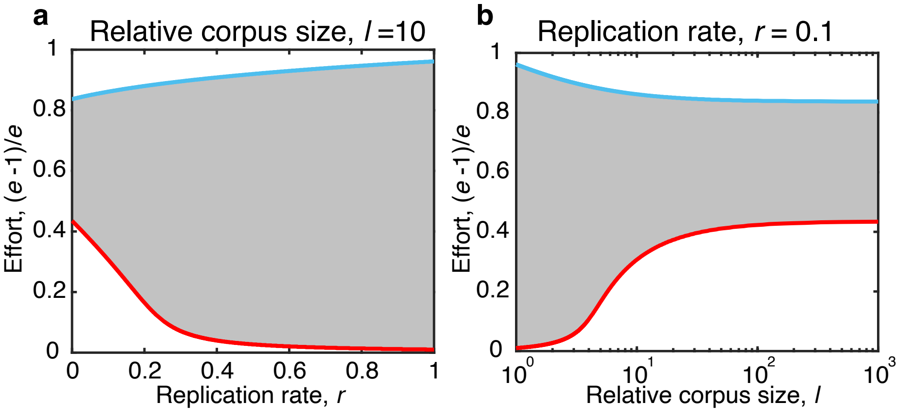
$$\begin{aligned} s_e = \left.\frac{\partial w}{\partial e_i}\right|_{e_i=e} = & \quad -\tfrac{\eta}{e\log[10]}\left[P(T)P(+|T)\alpha + P(F)P(+|F)\beta\right] + \\ & (1-\eta\log_{10}(e)) \times \left[\tfrac{d(P_i(T)P_i(+|T))}{de_i}\alpha + \tfrac{d(P_i(F)P_i(+|F))}{de_i}\beta\right] \end{aligned} \tag{29}$$

where α and β given by Eq. 21 account for the amount of enforced replication under the policy. As in previous examples, Eq. 29 must be solved numerically. Supplementary Figure 5 shows the basin of attraction for good and bad science as a function of replication rate r and literature size l. Supplementary Figure 3c and 3f shows the effect of introducing replication on the basin of attraction as a function of θ and b_1. We see that more stringent replication (arising from either higher rates of enforced replication, or a lower ratio of literature to labs) results in a larger basin of attraction for good science.

1.6 Attention-grabbing hypotheses

Up until this point we have assumed that strong hypotheses, which are true with probability b_1 produce the same benefits on publication as weak hypotheses, which are true with probability $b_0 < b_1$. However we may also consider a scenario in which publication of different types of hypotheses produce different benefits, B_N^1 and B_N^0.

In particular, the case where $B_N^0 > B_N^1$ describes a scenario in which weak hypotheses are also attention-grabbing, due to their novelty and surprise relative to prior work. Scientific culture



Supplementary Figure 5: Replication as a policy under adaptive dynamics. The figure shows equilibrium publication strategies in a large population of labs, as a function of model parameters. Plotted in each panel are the locations of the stable (blue) and unstable (red) equilibria as a function of either the replication rate r or the size of the corpus of literature relative to the number of active labs, l. For many parameter choices the system is bi-stable, with a good-science equilibrium indicated by the blue line and a bad-science equilibrium at minimum effort $(e-1)/e = 0$. In the gray regions selection favors increasing effort towards the good-science equilibrium; whereas in the white regions selection favors ever decreasing effort towards to bad-science equilibrium. a) Increasing the replication rate r increases the basin of attraction for good science, for a corpus of relative size $l = 10$. Here a 20% replication rate is sufficient to produce a large basin of attraction for the good-science equilibrium. b) Impact of corpus size l on the basin of attraction for good science for a fixed replication rate $r = 0.1$. Here a larger corpus of $l > 10$ acts to minimize the size of the basin of attraction for good science. All other parameters are as in Supplementary Figure 3.

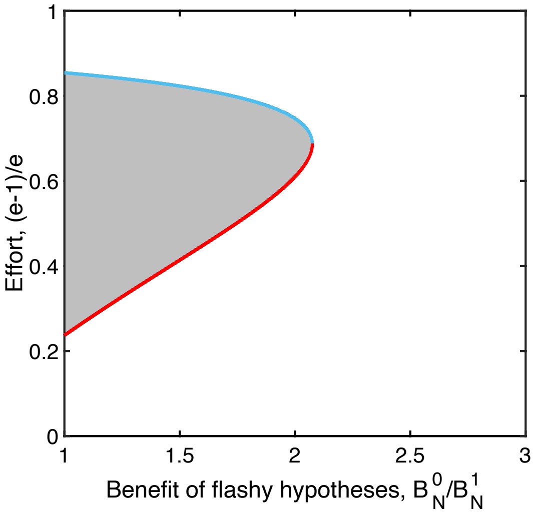
that provides greater rewards to publishing attention-grabbing hypotheses may undermine a good-science equilibrium. To assess the impact of attention-grabbing hypotheses on good science we looked at how the basin of attraction for good science changes as B_N^0 increases (Supplementary Figure 6). We see that when $B_N^0/B_N^1 < 2$ good science is still sustainable.

1.7 Good science viability

In the analysis and figures above we have explored how the basin of attraction of a good-science equilibrium changes as individual parameters are varied. We would also like to estimate the overall viability of good science in the full 9-dimensional parameter space of the model. In order to achieve this we performed a systematic numerical exploration of parameter space, randomly selecting 10^7 parameter sets for a given replication rate r, corpus size l and research time η, each of which was then varied systematically (Figure 4 and Figures S7-S9). For each set of 10^7 parameters we estimated the likelihood that good science was viable by calculating the proportion of parameter sets that could sustain a stable good-science equilibrium.

1.8 Time to produce good science

Following Smaldino and McElreath (2016) we typically assume that the time to produce a study is a convex function of the effort, e put into science, of the form $(1 - \eta \log_{10}(e))$. To assess the impact of this assumption on our results we also consider a concave function of the form $\eta \log_{10}(10^{1/\eta} + 1 - e)$

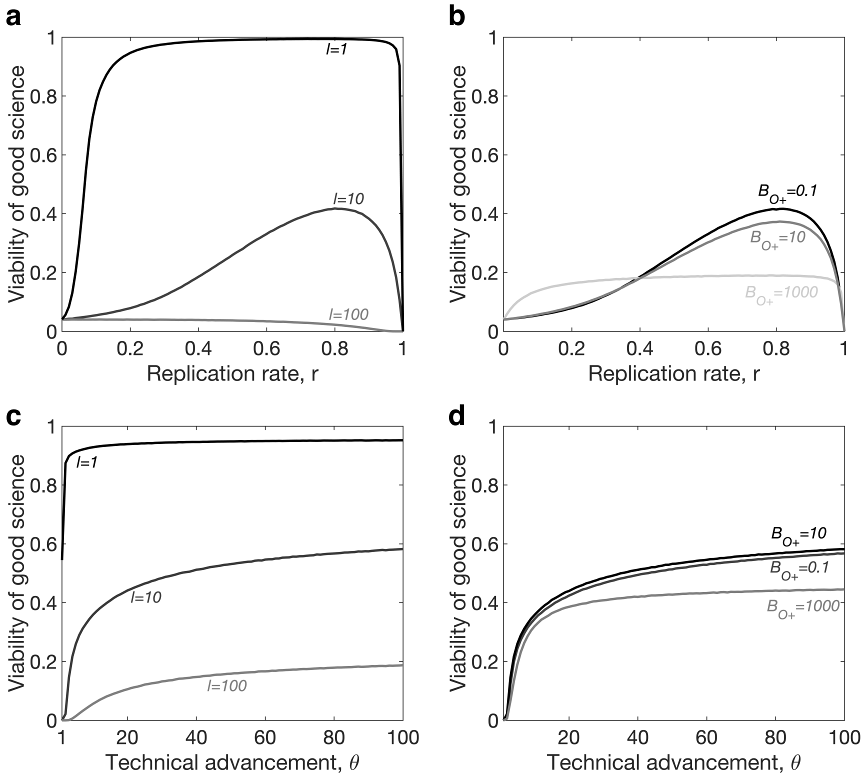


Supplementary Figure 6: Equillibria under adaptive dynamics with attention-grabbing hypotheses. The figure shows equilibrium publication strategies in a large population of labs, as a function of model parameters. Plotted in each panel are the locations of the stable (blue) and unstable (red) equilibria as a function of all five parameters of the system without replication. For many parameter choices the system is bi-stable, with a good-science equilibrium indicated by the blue line and a bad-science equilibrium at minimum effort $(e-1)/e = 0$. In the gray regions selection favors increasing effort towards the good-science equilibrium; whereas in the white regions selection favors ever decreasing effort towards to bad-science equilibrium. When $B^0_N/B^1_N < 2$ a good-science equilibrium remains viable for replication rate $r = 0.15$. All other parameters are as in Supplementary Figure 3 unless otherwise specified in the panel.

and a linear function of the form $1 - 10^{-1/\eta}(e-1)$ (Supplementary Figure 8c). As expected the convex function is the most conservative choice, in the sense that it produces a lower level of good science viability than either the linear or concave functions; but results are qualitatively similar for all these formulations of time as a function of effort.

1.9 Unequal distribution of effort

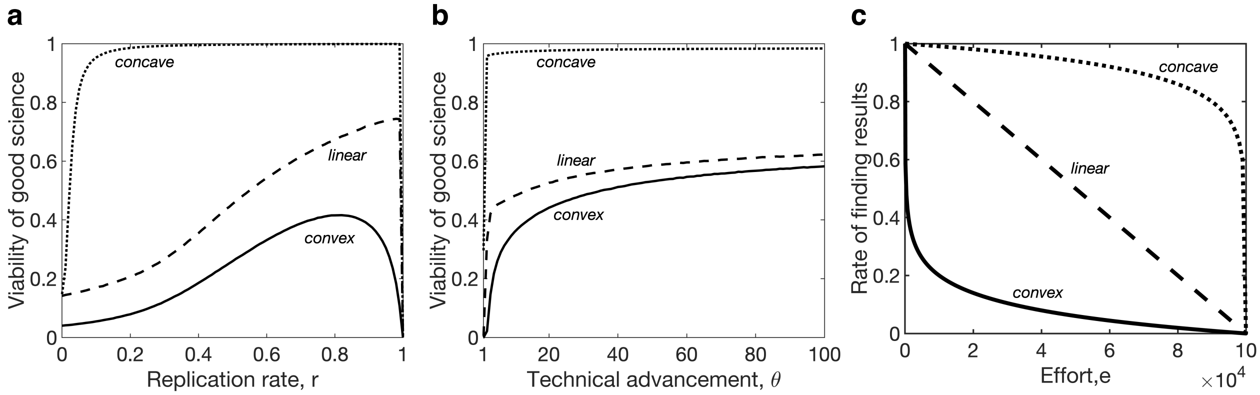
We have assumed that effort e impacts both hypothesis selection and hypothesis testing equally. However in reality, a lab may emphasize one or the other of these two aspects of the scientific process, and split effort unequally between them. To address this we introduce a parameter f which describes the distribution of effort between hypothesis selection and hypothesis testing. In particular, if the total level of effort expended by a lab is e, the effort spend on hypothesis selection is fe while the effort spent on hypothesis testing is $(1-f)e$. Note that when $f = 0$ we recover the model of Smaldino and McElreath (2016) in which good science cannot be sustained. We see that when $f = 0.33$ - i.e. when twice as much effort is put into testing than is put into selection, and when $f = 0.67$, i.e. when twice as much effort is put into hypothesis selection, good science remains highly viable.



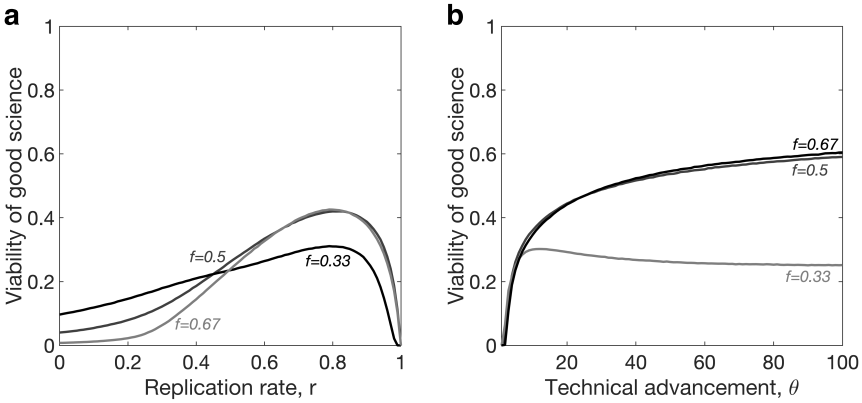
Supplementary Figure 7: Viability of good science across fields. The figure shows the proportion of parameter sets which support a stable good-science equilibrium, as a function of the replication rate r (a and b) and level of technical advancement, θ (c and d), for different costs and benefits of publication. In all cases studied, we see that introducing replication $r > 0$ initially increases the viability of good science. But this only occurs up to a point: when replication rates are very high, and replication studies are beneficial, there is comparatively less reward for effort spent at hypothesis selection a novel research. On the other hand, we see that increasing the technical advancement of a field $\theta > 1$ always increases the viability of good science. (a and c) Larger relative corpus sizes l tend to reduce the viability of good science. This is because the benefit for performing a replication study is awarded independent of effort, which reduces the marginal benefit of effort spend at hypothesis selection. (b and d) Increased benefits to a lab following successful replication of the study B_{O+} can increase or decrease the viability of good science depending on the replication rate. For each curve, we drew 10^7 parameter sets for every value of r at increments of 0.01 between 0 and 1. We chose parameters from the following ranges: $b_0 \in [0, 1]$, $b_1 \in [b_0, 1]$, $\theta \in [2, \infty)$, $\gamma \in [1, 2]$, $B_N^0 \in [1, 2]$, $r \in [0, 1]$. Unless otherwise indicated we fixed $B_N^1 = 1$, $B_r = 0.2$, $B_{O+} = 0.1$, $C_{O+} = 100$, $\eta = 0.2$ and $l = 10$.

1.10 Good science basin of attraction

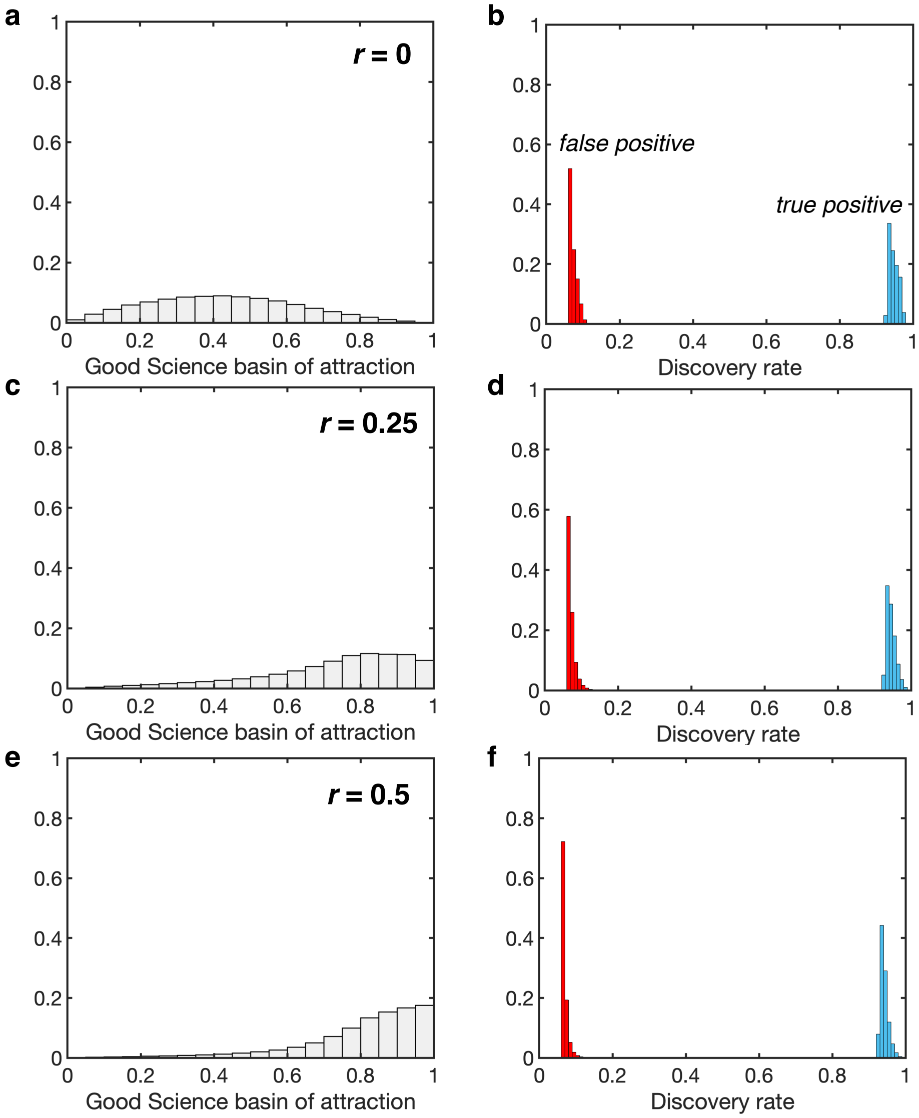
The results above show the viability of good science across a wide range of parameters. However the basin of attraction of good science when it exists also varies as we vary the replication rate. Supplementary Figure 10 shows the distribution of sizes for the basin of attraction of good science, as well as the true and false positive rates, as replication rate r is varied. We see that increasing the replication rate does not impact the rate of true and false positive but does make the size of the basin of attraction of good science bigger. This is consistent with our observation that replication is synergistic with theory, not a replacement for it.



Supplementary Figure 8: Viability of good science across fields. The figure shows the proportion of parameter sets which support a stable good-science equilibrium, as a function of the replication rate r (a) and level of technical advancement, θ (b), for different choices of functional form for the time to produce a study (convex, concave or linear as described in the text). (a-b) A convex function is always conservative (produces a lower viability of good science) than concave of linear unctions. c) Rate of finding results as a function of effort for all three functions. For each curve, we drew 10^7 parameter sets for every value of r at increments of 0.01 between 0 and 1. We chose parameters from the following ranges: $b_0 \in [0,1]$, $b_1 \in [b_0,1]$, $\theta \in [2,\infty)$, $\gamma \in [1,2]$, $B_N^0 \in [1,2]$, $r \in [0,1]$. Unless otherwise indicated we fixed $B_N^1 = 1$, $B_r = 0.2$, $B_{O+} = 0.1$, $C_{O+} = 100$, $\eta = 0.2$ and $l = 10$.



Supplementary Figure 9: Viability of good science across fields. The figure shows the proportion of parameter sets which support a stable good-science equilibrium, as a function of the replication rate r (a) and level of technical advancement, θ (b), for different distributions of effort between hypothesis selection and testing, f. (a) Depending on the rate of replication, putting more effort into selection or testing may improve the viability of good science. c) As we vary the level of technical advancement, putting more effort into theory is typically better for sustaining viable good science. For each curve, we drew 10^7 parameter sets for every value of r at increments of 0.01 between 0 and 1. We chose parameters from the following ranges: $b_0 \in [0,1]$, $b_1 \in [b_0,1]$, $\theta \in [2,\infty)$, $\gamma \in [1,2]$, $B_N^0 \in [1,2]$, $r \in [0,1]$. Unless otherwise indicated we fixed $B_N^1 = 1$, $B_r = 0.2$, $B_{O+} = 0.1$, $C_{O+} = 100$, $\eta = 0.2$ and $l = 10$.



Supplementary Figure 10: Basin of attraction of good science across fields. The figure shows the distribution of sizes for the basin of attraction of good science (left) and the distribution of true and false positive rates (right), conditional on good science being viable, for 10^7 randomly drawn parameter sets. (a-b) When replication rate is 0, false positive rates are low but there is wide variation in the basin of attraction of good science. (c-d) A 25% replication rate does not noticeably impact the true and false positive rate, but the size of the basin of attraction increases. (e-f) Increasing the replication rate further to 50% further grows the basin of attraction of good science. For each plot, we drew 10^7 parameter sets for every value of r indicated, and calculated basin of attraction as the maximum and minimum levels of effort such that, when initialized at that value, the system would evolve towards the good-science equilibrium under our adaptive dynamics analysis. We chose parameters from the following ranges: $b_0 \in [0,1]$, $b_1 \in [b_0, 1]$, $\theta \in [2, \infty)$, $\gamma \in [1,2]$, $B_N^0 \in [1,2]$, $r \in [0,1]$. Unless otherwise indicated we fixed $B_N^1 = 1$, $B_r = 0.2$, $B_{O+} = 0.1$, $C_{O+} = 100$, $\eta = 0.2$ and $l = 10$.

2 Individual-based simulations

We ran individual-based simulations, relaxing the assumptions of the adaptive dynamics model described above to account for (i) variation in lab age and (ii) heterogeneity in lab publication strategy (iii) a finite population of active labs. We treated effort e, efficacy V and replication rate r as heritable, evolving traits. We ran ensembles of 10^3 replicate simulations to produce each simulation figure and plotted the average trajectories over time. Further details of the simulation setup are provided below.

2.1 Lab aging

Under the assumptions of adaptive dynamics the population of labs is infinite and the lab life cycle ensures that all labs are the same age when natural selection occurs. These simplifying assumptions are made for mathematical convenience but do not describe a particularly realistic case: in any given field there is a wide range of labs of different ages, and the older a lab is, the more it has published. This has consequences for the rate at which the lab experiences replication attempts (as they have contributed more novel results to the corpus of results in their field) which in turn has consequences for their fitness.

We assume that labs "die" when they copy another lab's strategy (see below). Furthermore we assume that the fitness of a lab is determined by the average payoff received due to novel publication and replication over the lab lifetime.

2.2 Natural selection and the copying process

We assume that lab birth and death occurs via the copying process Traulsen et al. (2006) used to study a process of cultural evolution via imitation. Under this model, we assume that a pair of labs i and j are chosen at random, such that lab i chooses to adopt the strategy of lab j with a probability π_{ij} where

$$\pi_{ij} = \frac{1}{1 + e^{\sigma(\bar{w}_i - \bar{w}j)}} \tag{30}$$

where $\bar{w}_i$ is the average payoff to lab j during its lifetime. This birth-death process can be thought of as a fixed population of labs who update their strategies, described by their methodological efficacy V, effort e and replication rate r, when they see another lab doing better. This may be thought of as occurring whenever an old lab is disbanded and replaced with a new lab in a university or research institute. Alternatively it may be understood as occurring among a fixed population of competing labs trying to gain an edge over one another.

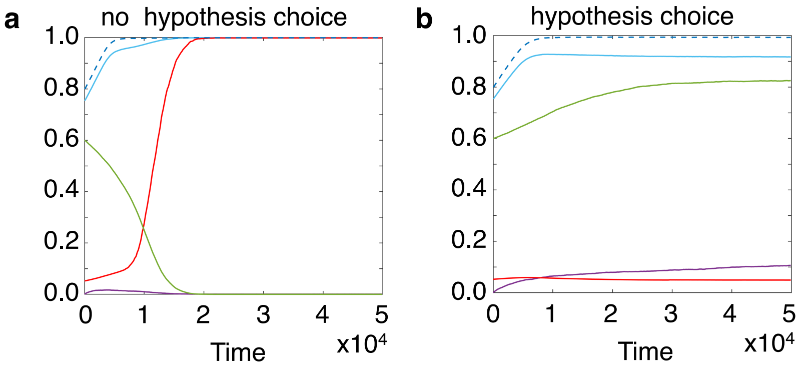
2.3 Replication

Populations of competing labs are assumed to contribute to a corpus of literature of size L. When choosing a study to replicate a lab chooses a study at random from the corpus. They attempt to reproduce the study using the same level of methodological efficacy V and effort e as for testing a novel hypothesis. After an attempt at reproduction the study is moved from the corpus of literature available for replication.

As a result, a lab that has produced n papers with novel results has a study reproduced with probability n/L when another lab decides to undertake a replication study. If the outcome of the replicating labs study is positive, the replication is successful otherwise it is not. The corpus of literature is always assumed to contain L novel papers available for replication - if all the papers by currently active labs have been replicated we assume that the labs can still reproduce older literature. Thus labs can in principle engage in replication at the maximum rate $r = 1$, although this pathological case is not observed in simulations or under the adaptive dynamics model, except transiently (Supplementary Figure 2).

2.4 Co-evolution of effort and replication

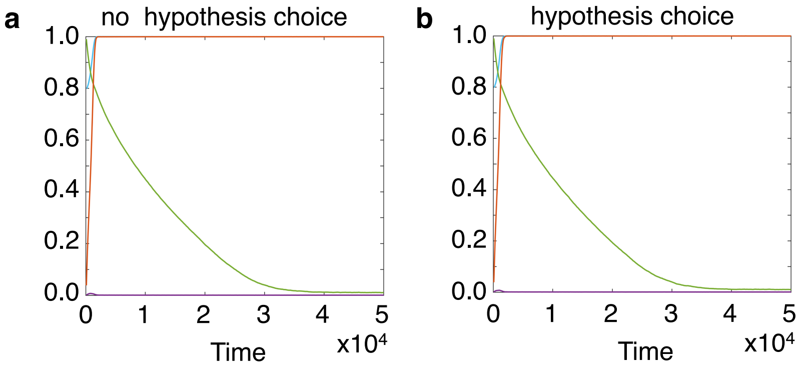
We explored the co-evolutionary dynamics of replication and effort via individual-based simulations (Supplementary Figure 11). In the absence of hypothesis choice only very low levels of replication emerged and, as in Figure 2 and Figure 4 of the main text, effort evolved to the bad-science minimum. In contrast, when hypothesis choice was allowed the good-science equilibrium was maintained and replication evolved steadily to around $r = 0.1$ (Supplementary Figure 11b).



Supplementary Figure 11: Co-evolution of replication and effort. The figure shows results of individual-based simulations for the co-evolution replication and effort. (a) In the absence of hypothesis choice, both true (blue) and false (red) positive rates increase to unity, and effort declines to a minimum $(e-1)/e = 0$ (green), while replication rate (purple) remains low. b) However, when hypothesis choice is allowed, effort increases over time towards a good-science equilibrium in which false positives are rare, and replication evolves to a modest rate. All parameters are the same as in Figure 2c. Replications are chosen from a corpus of $L = 10^5$ novel studies, and each study is allowed to be replicated only once (see SI). Payoffs are $B_N = 1$, $B_r = 0.2$, $B_{O+} = 0.1$ and $C_{O-} = 100$.

2.5 Limit of $\theta = \gamma = 1$

Our model reproduces that of Smaldino and McElreath (2016) in the limit $\gamma = \theta = 1$, and as such our simulations in this limit should produce the same qualitative results. We ran simulations in this limit without hypothesis choice and showed that, indeed, the bad-science equilibrium quickly emerged (Supplementary Figure 12a). When hypothesis choice was allowed (Supplementary Figure 12b) the bad-science equilibrium still evolved in this limit, since power $P(+|T)$ and false positive rate $P(+|F)$ are *both* independent of effort under this model, once efficacy evolves to its maximum $V = 1$. This latter result illustrates a pathology of the limit $\theta = \gamma = 1$, under which bad science (true- and false-positive rates equal to one) cannot be avoided, no matter how much effort a lab puts in, once methodological efficacy reaches its maximum – a state of affairs that does not reflect reality in any scientific field. However, when we separate out methodological efficacy from lab effort in identifying positive results, and allow for the possibility that a diligent lab can, in principle, expend effort to do good science (i.e. by setting $\gamma > 1$ and $\theta > 1$), the effects of theory on stabilizing good science become apparent, and both good- and bad-science equilibria emerge – a state of affairs that more accurately reflects what we see in scientific practice across fields.



Supplementary Figure 12: Simulations in the limit $\theta = \gamma = 1$. This figure is the same as Supplementary Figure 11 with the alteration that the technical limits of false- and true-positives are set to $\theta = \gamma = 1$. In this case both without (a) and with (b) hypothesis choice, bad science evolves.

References

Leimar, O. 2009. Multidimensional convergence stability. Evolutionary Ecology Research 11:191–208.

Smaldino, P. E., and R. McElreath. 2016. The natural selection of bad science. Royal Society open science 3:160384.

Traulsen, A., M. A. Nowak, and J. M. Pacheco. 2006. Stochastic dynamics of invasion and fixation. Phys Rev E Stat Nonlin Soft Matter Phys 74:011909.