LETTER

doi:10.1038/nature12786

Modelling the effects of subjective and objective decision making in scientific peer review

In-Uck Park[1,2], Mike W. Peacey[1,3] & Marcus R. Munafò[4,5,6]

The objective of science is to advance knowledge, primarily in two interlinked ways: circulating ideas, and defending or criticizing the ideas of others. Peer review acts as the gatekeeper to these mechanisms. Given the increasing concern surrounding the reproducibility of much published research[1], it is critical to understand whether peer review is intrinsically susceptible to failure, or whether other extrinsic factors are responsible that distort scientists' decisions. Here we show that even when scientists are motivated to promote the truth, their behaviour may be influenced, and even dominated, by information gleaned from their peers' behaviour, rather than by their personal dispositions. This phenomenon, known as herding, subjects the scientific community to an inherent risk of converging on an incorrect answer and raises the possibility that, under certain conditions, science may not be self-correcting. We further demonstrate that exercising some subjectivity in reviewer decisions, which serves to curb the herding process, can be beneficial for the scientific community in processing available information to estimate truth more accurately. By examining the impact of different models of reviewer decisions on the dynamic process of publication, and thereby on eventual aggregation of knowledge, we provide a new perspective on the ongoing discussion of how the peer-review process may be improved.

Current incentive structures in science promote attempts to publish in prestigious journals, which frequently prioritize new, exciting findings. One consequence of this may be the emergence of fads and fashions in the scientific literature (that is, 'hot topics')[1], leading to convergence on a particular paradigm or methodology. This may not matter if this convergence is on the truth—topics may simply cease to be hot topics as the problem becomes solved. However, there is increasing concern that many published research findings are in fact false[1]. It is common for early findings to be refuted by subsequent evidence, often leading to the formation of groups that interpret the same evidence in notably different ways[2], and this phenomenon is observed across many scientific disciplines[3,4]. There are a number of relatively recent examples of convergence on false hypotheses, such as the theory of stress causing gastric ulcer formation[5]. Once established, these can become surprisingly difficult to refute[6]—they may become "more 'vampirical' than 'empirical'—unable to be killed by mere evidence"[7]. Science may therefore not be as self-correcting as is commonly believed[8], and the selective reporting of results can produce literatures that "consist in substantial part of false conclusions"[9].

It is important to understand how convergence on false conclusions may come about. A number of possibilities present themselves. First, scientists may not in fact be rational individuals pursuing the truth after all—an argument made by some influential sociologists of science (the strong programme)[10]—or may be rational but stuck within a particular paradigm[11]. Second, some scientists may be biased or even immoral—a number of high profile cases of data fabrication and fraud have emerged in recent years[12]. Third, some scientists may care more about publication and careers than discovering the truth (that is, 'publish or perish'), a process which may be conscious or unconscious[13]. In competitive fields current incentive structures prioritize positive results, which may increase the likelihood of modification of data or conducting many statistical tests to achieve these; similarly, increased error rates may arise from multiple competing research groups testing the same hypotheses[14].

It has been shown that increased popularity of a particular research theme reduces the reliability of published results[14], and that findings published in prestigious journals are less reliable and more likely to be retracted[15]. Therefore, the convergence of research interest on a current hot topic may serve to undermine the reliability and veracity of subsequently published findings. In principle, peer review should eliminate or reduce these problems but, given empirical evidence for the unreliability of much published research, it may not in fact be conducted properly, or the process itself may be flawed. Empirical research and simulations have identified a number of factors which contribute to the likelihood that a published research finding is false[1,16]. However, the peer-review process itself has not been closely investigated as a possible influence, despite the fact that it acts as the ultimate gatekeeper of research publication. It is generally regarded as imperfect, although still the best model available to ensure both the quality and veracity of published scientific research, but there has been growing concern that it fails, at least in part, with respect to each of these two goals[1].

To understand the peer-review mechanism better, using a Bayesian approach in a model of the publication process, we analysed the behaviour of scientists who have developed their initial opinions independently as to which of the two opposing hypotheses, A and B, is more likely to be true. They know that on average their opinion is indeed correct with probability $\beta \in \left(\frac{1}{2}, 1\right)$, so they feel confident, but less than fully, about their opinion. The more controversial the issue, the lower the value of β. Upon receiving a manuscript that advocates one of the hypotheses, the editor of a hypothetical journal solicits a review from another scientist, who recommends acceptance or rejection. To focus on the influence of reviewer behaviour, rather than that of editor, we assume that the editor simply follows the reviewer's recommendation. Subsequently, the reviewer writes and submits their own manuscript to the journal, and the process repeats. The two decisions for each scientist are therefore: (1) whether or not to recommend acceptance of a manuscript that they are reviewing, and (2) which hypothesis to advocate in their own submission, which we term the 'theme' of their manuscript. As a publication history evolves (following cycles of submission, peer review and acceptance or rejection) a scientist revises their view on the likelihood of each hypothesis being true, in light of the relative probability of this particular history occurring when one hypothesis is true as opposed to the other. Being motivated to promote the truth, each scientist will advocate a theme that is more likely to be true, according to their revised view when they submit a manuscript.

Our aim was to understand how different criteria of reviewing decisions influence the publication outcome, and how the resulting publication

[1]Department of Economics, University of Bristol, Bristol BS8 1TN, UK. [2]Department of Economics, Sungkyunkwan University, Seoul 110-745, South Korea. [3]Department of Economics, University of Bath, Bath BA2 7AY, UK. [4]MRC Integrative Epidemiology Unit (IEU), University of Bristol, Bristol BS8 1BN, UK. [5]UK Centre for Tobacco and Alcohol Studies, University of Bristol, Bristol BS8 1TU, UK. [6]School of Experimental Psychology, University of Bristol, Bristol BS8 1TU, UK.

histories and the information inherent in the relevant peer-review criterion influence the community's eventual understanding of the topic. To this end we modelled and compared two different ways that scientists approach the reviewing decision. In the first model (M1), the subjective criterion of how strongly the reviewer agrees with the conclusion of the research (that is, the theme of the manuscript) is reflected in the decision, in addition to other more objective criteria such as research design and methodology. In the second model (M2), the decision reflects objective criteria only. Our findings, therefore, may shed light on whether subjective assessment is desirable in the peer-review process and, if so, to what extent. As a benchmark, we also compared M1 and M2 with a default model (M3), in which all manuscripts are published without any filtering through peer review. As scientists will make inferences that take into account how reviewers arrive at their recommendations, the particular peer-review model in operation affects how they revise their views and, thereby, their decisions on which theme to advocate as an author, as well as their decisions as a reviewer.

The results of the three models (Fig. 1) indicate that: (1) almost certainly, some scientists will submit manuscripts on themes which disagree with their initial opinion (we term this 'herding'); (2) the extent to which the wider scientific community's perception of a literature is removed from the truth (we term this 'misperception') decreases with number of publications, but information transmission is greatly hampered once herding has occurred, to such an extent that no further improvement in understanding occurs except in M1 where a degree of subjectivity is allowed in the reviewing decision (that is, reviewers as well as authors act guided by Bayesian inference); and (3) the probability of another publication on a particular issue increases as the number of manuscripts published on that issue increases, owing to aggregation of information and herding reinforcing the scientific community's consensus.

The phenomenon known as herding is inherent in the behaviour of scientists operating under all of the models we consider. An individual is said to be herding if they choose a theme to advocate in their manuscript submission based entirely on what they have observed from others, independently of what they initially thought was true. The degree of herding depends on the peer-review model in operation, the number of manuscripts submitted so far, and how confident scientists feel about their initial opinion (β). Herding takes place relatively quickly (Fig. 1), and we observe discrete jumps in the measure of herding early on in the process, when each signal (that is, the information carried by a peer-review decision) carries a large weighting. Notably, the probability of herding and the speed with which it increases are eventually lower when a degree of subjectivity is allowed in the reviewing decision (M1), and only in this case can a fad be reversed following a sequence of publications on the same theme. As a fad persists, the total number of scientists required in order to reverse this fad increases—and at a faster rate.

We use 'misperception' to describe how incorrect the perception of the wider scientific community is after a history of publication outcomes. It is defined as the probability that an outsider assigns to a hypothesis being correct, based on Bayesian inference from the observed history, when it is actually incorrect. The level of expected misperception (Fig. 1) remains relatively stable for low and high values of β, but for intermediate values of β it declines with increasing numbers of submitted manuscripts. Critically, when a degree of subjectivity is allowed in the peer-review process (M1), this always eventually outperforms the other models, because in these models information completely fails to be transmitted after herding occurs.

In our models, manuscript submission decisions made by individual scientists are based in part on information inferred from others' actions, because individuals use information from the publication history within a particular field, as well as their personal opinions, to guide their decisions. This may have positive effects if the decisions cluster around a correct outcome, or have negative effects if they cluster around an incorrect outcome. A degree of subjectivity in the peer-review process will, on average, lead to lower misperception, because reviewer decisions (and subsequent editorial decisions) which go against the herding trend will continue to reveal new information. In addition, the process is dynamic, and we show that self-correction can eventually occur when a degree of subjectivity is allowed in the peer-review process; however, it may not when the reviewing decision is completely independent of the reviewer's subjective assessment of the theme of the manuscript, and is based only on other, largely objective characteristics of the manuscript, such as the quality of the research methodology. In this case the probability of herding reaches 1 within finite time for all values of β, and the level of misperception cannot go below a certain lower bound. The concept of herding has been discussed in the context of scientific research in the past[17], but ours is the first study, to our knowledge, to model the processes by which it may occur.

These results raise the question of whether a higher level of subjectivity in reviewer decisions will lead to more effective restraint of incorrect herding. We therefore decided to test generalized M1 models, in which we varied the degree to which the reviewer's recommendation is determined by their subjective assessment of the conclusion. Our results (Fig. 2) indicate that excessively subjective reviews are not effective in restraining incorrect herding. This is because, in this case, recommendations are sensitive to whether the conclusion agrees with the reviewer's viewpoint at that time, and this factor is predominantly determined by

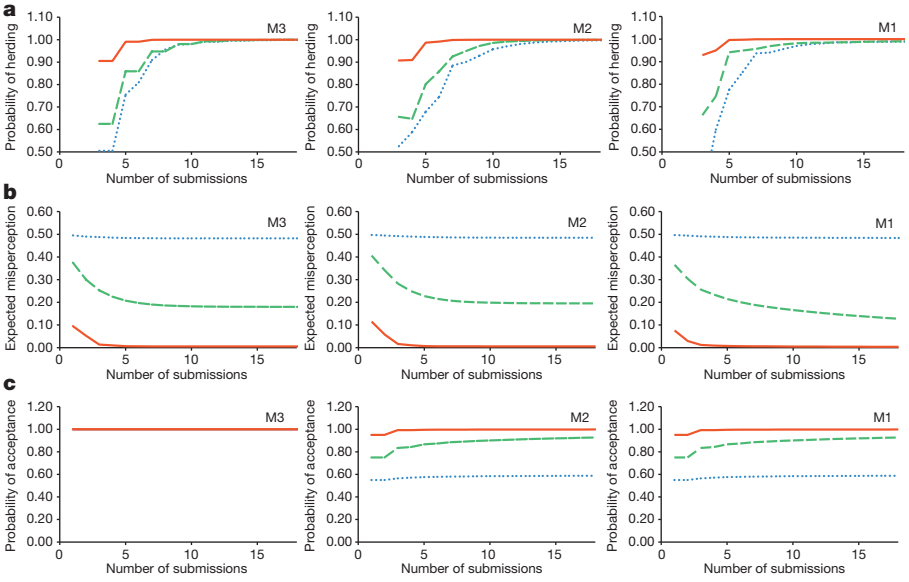

Figure 1 | Three models of peer-review behaviour. We show three models, M1 (right), M2 (middle) and M3 (left), which differ in the extent to which the peer-review decision depends on whether the reviewer agrees with the conclusion. Three outcomes are presented: (1) probability of herding (top), (2) average misperception generated (middle), and (3) probability of acceptance (bottom). The probability that the initial opinion is correct is reflected by β, and each outcome is presented for three values of β: (1) 0.55 (blue, dotted line), (2), 0.75 (green, dashed line), and (3) 0.95 (red, solid line), reflecting high, intermediate, and low uncertainty, respectively.

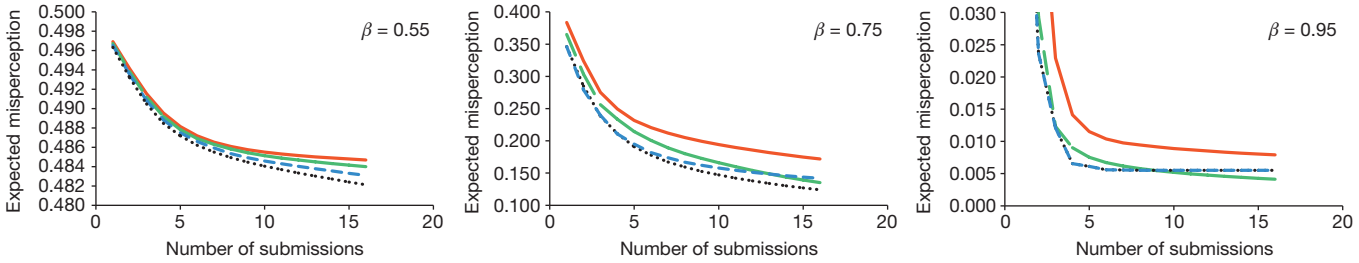

Figure 2 | Expected misperception in a generalized version of the M1 model. We show the expected misperception for three values of the probability that the initial opinion is correct (β): (1) 0.55 (left), (2) 0.75 (middle), and (3) 0.95 (right), reflecting high, intermediate, and low uncertainty. Results are shown for differing degrees to which the reviewer's subjective assessment determines their recommendation (v): (1) 0.75 (red, solid line), (2) 1.00 (green, long dashed line), (3) 1.25 (blue, short dashed line), and (4) 1.50 (black, dotted line). In the original M1 model $v = 1$, while lower values reflect a more objective reviewer, and higher values a more subjective reviewer. Excessively subjective reviews are not effective in restraining incorrect herding (this is not yet visible for $\beta = 0.55$, but would become apparent with more submissions).

the accumulated information, rather than their original opinion, as publication history lengthens. In other words, in this case even the reviewers' recommendations are subject to herding. It appears that a moderate degree of subjectivity (as depicted in M1) is near-optimal.

Two empirical examples show that herding occurs in the scientific literature. First, belief in a specific scientific claim can be (and is) distorted through preferential citations of studies which support a particular point of view rather than those which do not[17]. This phenomenon can be attributed to herding caused by preferential citations, potentially creating a spurious and unfounded sense of authority for specific claims. Second, using a meta-analytic review of a recent literature[18], we compared claims made in the abstracts of the contributing studies with support for those claims by the data reported therein. Meta-analysis imposes a standard analysis to maximize comparability, and thereby minimizes the extent to which the presentation of results can be influenced by flexible analytical options[19]. These results (Fig. 3) show a mismatch between the claims made in the abstracts, and the strength of evidence for those claims based on a neutral analysis of the data, consistent with the occurrence of herding.

We next consider whether scientists can decide on their conclusion before conducting an experiment. We suggest that herding leads to one outcome being preferable over another, and that flexible analysis and selective reporting allows data that do not conform to either be transformed[19] or relegated to the file drawer[20]. Mendel famously appears to have dropped observations from his data so that his results conformed to his expectations[21], but because his theory was ultimately proved correct this is now generally overlooked. There is in fact clear evidence that the reporting and interpretation of findings is often inconsistent with the actual results[22], and this appears to be particularly pronounced in abstracts of research articles (often the only part that is read)[23].

Scientists may be motivated by a number of factors, such as the desire to be the first to advocate an idea, and the natural tendency to side with others of a similar opinion. Herding is therefore expected when agents care only about being published and recognize some topics as 'hot' (and therefore publishable). If scientists are motivated in this way in our model, then in an equilibrium of the game they can simply follow the first author's claim to maximize the probability of being published (see Supplementary Information). However, our results indicate that we can expect herding, including convergence on false conclusions, even when scientists—both as authors and reviewers—are rational and motivated by the pursuit of truth. The emergence of fads and fashions in the scientific literature (that is, hot topics)[1] is therefore unsurprising.

The first herding model in economics modelled individuals' investment choices[24]. Herding may have positive consequences, by driving rapid convergence on a correct decision. Rational individuals process all the information available to them before making decisions, and herding therefore arises from natural motives—a rational individual in pursuit of truth can and should be influenced by what others think. That humans are influenced in this way has been shown by experiments in social psychology[25]. It is rational because humans are aware of their own fallibility, and so their opinions may be strengthened or weakened by the views of others. In other words, being aware of the wisdom of the crowd, humans are (rationally) influenced by the crowd; in order to update our beliefs in the light of new evidence, we should be guided by Bayes' theorem. However, herding may also have negative consequences, by driving convergence on an incorrect decision. This is particularly problematic if an outsider to the process is unaware that it is taking place, as it gives a spurious sense of certainty to the observed convergence.

Free, open and global access to research reports has been proposed as an alternative to peer review (http://am.ascb.org/dora/), but, as we have shown, peer review can reveal more information relative to free and complete sequential publication. Reviewer recommendations, and resulting editor decisions, contain information, and thus prevent herding from completely blocking new information flow. However, this depends on specific parameters such as the popularity of the subject (for example, how many people are writing about this issue, or how long it is discussed) and how strongly scientists feel about their initial dispositions (that is, the level of β). In particular, if reviewers (and editors) are explicitly encouraged to be as objective as possible they will not be guided by Bayes' theorem when making their recommendations—it is only when reviewers are allowed a degree of subjectivity that this is done. Our results indicate that peer review performs best when the reviewers exercise their subjectivity at an intermediate level; higher levels enhance the risk of complete herding in reviewer decisions, whereas lower levels curb the information flow from reviewer decisions.

The peer-review process is therefore in principle self-correcting over a sufficiently extended period (although distortions may occur in the shorter term), in that de-herding can also occur. In reality, de-herding will not always occur, because publication histories within a topic may not

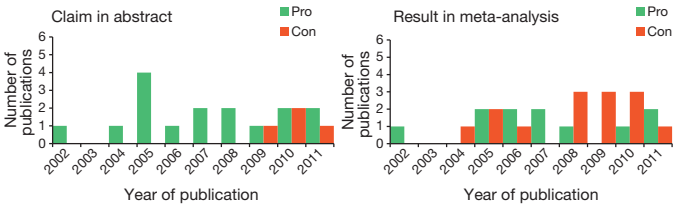

Figure 3 | Empirical evidence of discrepancy between claims and results. We show claims made in the abstracts of studies, and the results of those studies derived from a standardized analysis. Abstracts were coded as pro or con depending on whether an association was claimed, based on the judgement of an independent rater. Results were coded as pro or con depending on whether the overall effect size for the full sample in the study was statistically significant at $P < 0.05$. Five abstracts could not be coded as either pro or con. The proportion of pro versus con classifications differed for claims (80% pro) and results (44% pro), suggesting herding around the first published claim (McNemar test: $P = 0.016$, two-tailed test). Treating abstracts that could not be coded as pro or con did not alter these results substantially (84% versus 64% pro).

persist for sufficiently long. Science may therefore not be as self-correcting as is commonly assumed[8], and peer-review models which encourage objectivity over subjectivity may reduce the ability of science to self-correct. Although herding among agents is well understood in cases where the incentives directly reward acting in accord with the crowd (for example, financial markets), it is instructive to see that it can occur when agents (that is, scientists) are motivated by the pursuit of truth, and when gatekeepers (that is, reviewers and editors) exist with the same motivation. In such cases, it is important that individuals put weight on their private signals, in order to be able to escape from herding. Behavioural economic experiments indicate that prediction markets, which aggregate private signals across market participants, might provide information advantages[26]. Knowledge in scientific research is often highly diffuse, across individuals and groups[26], and publishing and peer-review models should attempt to capture this. We have discussed the importance of allowing reviewers to express subjective opinions in their recommendations, but other approaches, such as the use of post-publication peer review, may achieve the same end.

METHODS SUMMARY

Model. A number of scientists, indexed as $i = 1,2,\cdots$, deliberate over two opposing hypotheses perceived *ex ante* to be equally likely to be true. Initially each scientist i receives an independent private signal regarding the true hypothesis, which is correct with probability $\beta \in \left(\frac{1}{2},1\right)$. Sequentially, scientist i submits a manuscript defending one of the two hypotheses, termed its theme, which is reviewed by the next scientist $i+1$ who decides whether to accept or reject the manuscript. This decision, and the theme if accepted, becomes common knowledge. Each scientist submits a manuscript defending a theme that is more likely to be the true hypothesis according to their posterior belief, formed by Bayes' rule based on all the information available at that time. We consider three models of reviewer decision. In M1, the reviewer accepts a manuscript with a probability proportional to the likelihood of its theme being true according to their posterior belief. In M2, they accept it irrespective of its theme with the *ex ante* probability they would accept a manuscript after the same publication history in M1. In M3, they simply accept it.
Concepts. A scientist is herding if their posterior belief attaches a probability greater than 0.5 to a particular hypothesis regardless of their own signal when they submit. Their probability of herding is the *ex ante* probability that they will be herding. The misperception after a publication history is the expected probability attached to the hypothesis, which is in reality incorrect, by outside observers who form their posterior beliefs on true hypothesis by Bayes' rule based on the history. The expected misperception after n submissions is the probability-weighted sum of misperceptions over all possible histories that may occur with n submissions.
Analysis. We wrote a computer program to recursively calculate numerical values of algebraic formulae for various concepts reported, and algebraically derived asymptotic properties for large numbers of submissions.

Online Content Any additional Methods, Extended Data display items and Source Data are available in the online version of the paper; references unique to these sections appear only in the online paper.

Received 1 August; accepted 16 October 2013.
Published online 4 December 2013.

1. Ioannidis, J. P. A. Why most published research findings are false. *PLoS Med.* **2,** e124 (2005).
2. Ioannidis, J. P. A. Scientific inbreeding and same-team replication: Type D personality as an example. *J. Psychosom. Res.* **73,** 408–410 (2012).
3. Ioannidis, J. P. A. Contradicted and initially stronger effects in highly cited clinical research. *J. Am. Med. Assoc.* **294,** 218–228 (2005).
4. Ioannidis, J. P. & Trikalinos, T. A. Early extreme contradictory estimates may appear in published research: the Proteus phenomenon in molecular genetics research and randomized trials. *J. Clin. Epidemiol.* **58,** 543–549 (2005).
5. Davey Smith, G. in *Biopsychosocial Medicine: An Integrated Approach to Understanding Illness* (ed. White, P.) 77–102 (Oxford Univ. Press, 2005).
6. Tatsioni, A., Bonitsis, N. G. & Ioannidis, J. P. A. Persistence of contradicted claims in the literature. *J. Am. Med. Assoc.* **298,** 2517–2526 (2007).
7. Freese, J. in *Intergenerational Caregiving* (eds Crouter A. C., Booth A., Bianchi S. M. & Seltzer J. A.) 145–177 (Urban Institute Press, 2008).
8. Ioannidis, J. P. A. Why science is not necessarily self-correcting. *Perspect. Psychol. Sci.* **7,** 645–654 (2012).
9. Sterling, T. D. Publication decisions and their possible effects on inferences drawn from tests of significance—or vice versa. *J. Am. Stat. Assoc.* **54,** 30–34 (1959).
10. Barnes, B., Bloor, D. & Henry, J. *Scientific Knowledge: A Sociological Analysis.* (Univ. Chicago Press, 1996).
11. Kuhn, T. S. *The Structure of Scientific Revolutions.* (Univ. Chicago Press, 1962).
12. Yong, E. & Simonsohn, U. The data detective. *Nature* **487,** 18–19 (2012).
13. Martinson, B. C., Anderson, M. S. & de Vries, R. Scientists behaving badly. *Nature* **435,** 737–738 (2005).
14. Pfeiffer, T. & Hoffmann, R. Large-scale assessment of the effect of popularity on the reliability of research. *PLoS ONE* **4,** e5996 (2009).
15. Brembs, B., Button, K. & Munafò, M. R. Deep impact: unintended consequences of journal rank. *Front. Hum. Neurosci.* **7,** 291 (2013).
16. Button, K. S. *et al.* Power failure: why small sample size undermines the reliability of neuroscience. *Nature Rev. Neurosci.* **14,** 365–376 (2013).
17. Greenberg, S. A. How citation distortions create unfounded authority: analysis of a citation network. *Br. Med. J.* **339,** b2680 (2009).
18. Murphy, S. E. *et al.* The effect of the serotonin transporter polymorphism (5-HTTLPR) on amygdala function: a meta-analysis. *Mol. Psychiatry* **18,** 512–520 (2013).
19. Simmons, J. P., Nelson, L. D. & Simonsohn, U. False-positive psychology: undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychol. Sci.* **22,** 1359–1366 (2011).
20. Rosenthal, R. The file drawer problem and tolerance for null results. *Psychol. Bull.* **86,** 638–641 (1979).
21. Edwards, A. W. F. More on the too-good-to-be-true paradox and Gregor Mendel. *J. Hered.* **77,** 138 (1986).
22. Boutron, I., Dutton, S., Ravaud, P. & Altman, D. G. Reporting and interpretation of randomized controlled trials with statistically nonsignificant results for primary outcomes. *J. Am. Med. Assoc.* **303,** 2058–2064 (2010).
23. Gøtzsche, P. C. Believability of relative risks and odds ratios in abstracts: cross sectional study. *BMJ* **333,** 231–234 (2006).
24. Banerjee, A. V. A simple model of herd behavior. *Q. J. Econ.* **107,** 797–817 (1992).
25. Asch, S. E. Studies of independence and conformity. *Psychol. Monogr.* **70,** 1–70 (1956).
26. Almenberg, J., Kittlitz, K. & Pfeiffer, T. An experiment on prediction markets in science. *PLoS ONE* **4,** e8500 (2009).

Supplementary Information is available in the online version of the paper.

Acknowledgements This research was supported by an Economics and Social Research Council UK PhD studentship to M.W.P. M.R.M. is a member of the UK Centre for Tobacco and Alcohol Studies, a UKCRC Public Health Research Centre of Excellence. Funding from the British Heart Foundation, Cancer Research UK, Economic and Social Research Council, Medical Research Council, and the National Institute for Health Research, under the auspices of the UK Clinical Research Collaboration, is gratefully acknowledged. The authors are grateful to S. Murphy for her assistance in coding the meta-analysis study abstracts, and to A. Bird and G. Huxley for their comments on earlier drafts of this manuscript.

Author Contributions All authors contributed equally to the design and analysis of the models and the writing of the manuscript. The project was conceived by I.-U.P. and M.R.M., and the computer program was written by M.W.P.

Author Information Reprints and permissions information is available at www.nature.com/reprints. The authors declare no competing financial interests. Readers are welcome to comment on the online version of the paper. Correspondence and requests for materials should be addressed to M.R.M. (marcus.munafo@bristol.ac.uk).

METHODS

Model of the peer review process. We analyse a model in which $n+1$ *ex ante* identical scientists deliberate over two opposing hypotheses, labelled A and B. It is known that only one of these hypotheses is correct, and that *ex ante* both are equally likely to be correct. Denoting the correct hypothesis by τ, this is expressed as $P(\tau=A)=P(\tau=B)=\frac{1}{2}$. Before the game starts, each scientist i receives a private signal, $s_i\in\{A,B\}$, regarding which is the true hypothesis. These signals are independent random variables that assume a value equal to the correct hypothesis with probability β. The signals are informative but not perfect, that is, $\beta\in\left(\frac{1}{2},1\right)$. Lower values of β can be interpreted as reflecting a more controversial nature of the issue under question, when the signals tend to be less accurate.

Sequentially, and motivated to publish what is true, different scientists submit a manuscript, each defending a particular hypothesis. The 'theme' of scientist i's manuscript, $t_i\in\{A,B\}$, denotes the hypothesis that is defended. We postulate that, upon receiving a manuscript, the editor elicits peer review from a scientist whose stance on the topic is unknown to the editor, which eliminates the editor's influence on the editorial decision through reviewer selection. This is done to focus our analysis on reviewer behaviour, and means that in our model each manuscript is assigned to a scientist who has neither submitted their own manuscript nor acted as a reviewer at that point (because otherwise the editor would have inference on their stance from the theme of their submission or their previous decision as a reviewer). The editor follows the reviewer's recommendation in deciding whether to accept or reject the manuscript. If it is accepted, its theme becomes common knowledge; if it is rejected, the theme is not disclosed, but the rejection becomes common knowledge. Then, a new submission is made by a scientist who has not submitted before. In particular, our analysis is focused on the case that the next scientist who submits a manuscript is the one who reviewed the previous manuscript.

Thus, labelling the scientist who writes the i-th submission as i, each scientist $i\in\{1,2,\ldots,n\}$ sequentially submits a manuscript advocating a theme $t_i\in\{A,B\}$, which is reviewed by the next scientist $j=i+1$, who subsequently writes and submits their own manuscript. Scientist $n+1$, who also receives a signal s_{n+1}, only reviews. Scientists observe the history of publication outcomes as they arise. Let $h^i\in\{A,B,\varnothing\}^i$ denote a history of the first i publication outcomes, where each published manuscript is recorded by its theme, A or B, and each unpublished manuscript by Ø. Then, there are three items of information available to each scientist j when they make decisions: (1) their own private signal $s_j\in\{A,B\}$; (2) a manuscript to be reviewed with a theme $t_{j-1}\in\{A,B\}$ if $j>1$; and (3) a history $h^{j-2}\in\{A,B,\varnothing\}^{j-2}$ if $j>2$. The two decisions to make are whether or not to recommend acceptance of a manuscript that they are reviewing, and the theme of the manuscript they subsequently submit.

We made a few modelling choices that simplify real practices, namely that: (1) only one reviewer is consulted for each submission; (2) the current reviewer is the next author; (3) rejections become common knowledge; and, (4) authors conform to the rationality assumption that they are Bayesian updaters. Choices 1 and 2 maximize the number of submissions that can be reviewed by a given number of scientists, subject to the editor not soliciting a review from someone with a known stance. Choice 3 spares scientists from having to make probabilistic inferences as to what other submissions might have been made but rejected, which would have been necessary to determine the optimal choices when they act. These features enable us to examine the largest possible number of submissions with the available computing power, and thus allow us to generate more meaningful outputs without changing the essential processes operating. We believe that our main message will remain valid when these assumptions are relaxed (see Supplementary Information for a further discussion of choice 3). However, the complexity of the computer program needed to analyse such cases, and the corresponding computing power required, will increase exponentially. Choice 4 assumes authors use all of the information available to them, in accordance with Bayes' theorem[27], to determine the relative likelihood (called a posterior belief) that each of the two alternative hypotheses is correct. Then, being motivated to publish what is true, each scientist will submit a manuscript advocating the hypothesis that is more likely to be correct according to their posterior belief, augmented by a standard tie-breaking rule of following their own signal when both are equally likely[24]. This is one of the rationality assumptions that economists place on humans.

Models of reviewer behaviour. In the first model, M1, scientist $j=i+1$ recommends acceptance of scientist i's manuscript with the same probability, denoted by $P(\tau=t_i|\beta,h^{j-2},t_i,s_j)$, that they infer the theme of the manuscript to be the correct hypothesis, by Bayes' rule based on all the information available to them at that point. Therefore, reviewers as well as authors act guided by Bayesian inference in this model. The acceptance probabilities are endogenous and evolve differently depending on how the publication history unfolds.

In the second model, M2, the acceptance decision is completely independent of the reviewer's subjective assessment of the theme of the manuscript, and rather is based on other, largely objective characteristics of the manuscript, such as the quality of the research methodology. Presuming that these traits are statistically independent of the manuscript's conclusion, the acceptance probabilities in M2 are independent of both the theme of the manuscript and the assigned reviewer (insofar as the only feature that distinguishes reviewers is their assessment of which hypothesis is correct). Thus, the acceptance probabilities can be thought of as the likelihood that the methodological quality of the manuscript is sufficient to warrant publication, and not a reflection of whether or not the reviewer agrees with the conclusions. However, our model does not specify what those probabilities should be. To aid comparison between the models, we considered two cases. In one, scientist j, irrespective of their own signal, recommends acceptance of i's manuscript with a probability equal to the *ex ante* probability that they would recommend acceptance of i's manuscript in M1 after the same history (this results in the same expected number of publications in both M1 and M2). In the other, the acceptance probability remains the same throughout, at the initial expected acceptance probability of the M1 model, which is β. To verify this, note that scientist 2 would recommend acceptance of scientist 1's manuscript with probability $\frac{\beta^2}{\beta^2+(1-\beta)^2}$ when s_2 agrees with t_1 (which happens with probability $\beta^2+(1-\beta)^2=1-2\beta+2\beta^2$) but with probability 0.5 otherwise. Hence, the expected probability of acceptance is $\beta^2+\frac{1}{2}(2\beta-2\beta^2)=\beta$. As the results are similar in the two cases of M2, here we report only on the former.

In the third (benchmark) model, M3, all manuscripts are published without any filtering through peer review. This model is identical to M2 but with the acceptance probability equal to 1 throughout the process. This is a simple model of herd behaviour[24,28] that has become standard in economics when modelling self-motivated, rational individuals who sequentially take actions. A consequence of this model is that each scientist will have access to all previous submissions when forming their decision (because everything is published in this model). Note that this differs from a full information case (that is, where every scientist has access to all private signals, as well as public actions).

In the generalized M1 models, scientist j recommends acceptance with probability $\min\left\{1,\frac{1}{2}+v\cdot\left(P(\tau=t_i|\beta,h^{j-2},t_i,s_j)-\frac{1}{2}\right)\right\}$ if $P(\tau=t_i|\beta,h^{j-2},t_i,s_j)\geq\frac{1}{2}$, and with probability $\max\left\{0,\frac{1}{2}+v\cdot\left(P(\tau=t_i|\beta,h^{j-2},t_i,s_j)-\frac{1}{2}\right)\right\}$ if $P(\tau=t_i|\beta,h^{j-2},t_i,s_j)<\frac{1}{2}$, where $v>0$. The case $v=1$ corresponds to the original M1 model, with higher values of v indicating that the recommendation is more heavily influenced by the reviewer's subjective assessment on the advocated theme, and lower v meaning that it is less so.

Definitions and algebraic formulae. The misperception is defined from the perspective of outsiders who observe the publication history. Using all the information available to them from the observed history, $h^n\in\{A,B,\varnothing\}^n$, outside observers will form via Bayes' rule a posterior belief that attaches probability $P(\tau|h^n)=\frac{P(h^n|\tau)}{P(h^n|A)+P(h^n|B)}$ to hypothesis τ being true for $\tau\in\{A,B\}$, where $P(h^n|\tau)$ is the probability that the history h^n realizes under hypothesis $\tau\in\{A,B\}$. We define the misperception, after history h^n, as the expected posterior probability attached to the hypothesis which is in reality incorrect: since $P(\tau=A)=P(\tau=B)=\frac{1}{2}$, it is:

$$\frac{\frac{1}{2}\sum_{\tau=A,B}[1-P(\tau|h^n)]\cdot P(h^n|\tau)}{\frac{1}{2}\sum_{\tau=A,B}P(h^n|\tau)} \qquad (1)$$

The expected misperception after n submissions is defined as a probability-weighted sum of misperceptions over all possible histories of length n that may occur:

$$E[\text{misperception}]=\frac{1}{2}\sum_{\tau=A,B}\sum_{h^n\in\{A,B,\varnothing\}^n}[1-P(\tau|h^n)]P(h^n|\tau) \qquad (2)$$

Note that these calculations are done for an underlying value of β.

Focusing on h^1 (for which we need two scientists), there are three possible histories, namely $h^1\in\{A,B,\varnothing\}$. Equation (2), above, which gives us the expected misperception, will have 6 terms when $n=1$, because each of the three histories can occur from either hypothesis $\tau\in\{A,B\}$. Note that $P(\tau|h^1)$ is symmetric in the sense that its value remains the same when A and B (as values of τ and elements of h^1) are permuted. A consequence of this symmetry is that we only need to consider the case when one hypothesis (for example, A) is correct, and the sum of 6 terms

will be equal to twice of the sum of the three items relevant for $\tau = A$. For $n = 2$ because there are $3^2 = 9$ possible histories, there will be 9 terms to calculate (after taking into account the symmetry). Similarly, the expected misperception after n submissions can be obtained by calculating 3^n terms:

$$E[\text{misperception}] = \sum_{h^n \in \{A,B,\varnothing\}^n} [1 - P(A|h^n)]P(h^n|A) \quad (3)$$

Herding is defined for scientists who are submitting papers. A scientist, say j, is said to be herding if they would choose the same theme to advocate regardless of their private signal as their posterior belief would attach a probability more than one half to a particular hypothesis regardless of their own signal, that is, if:

$$\text{For some } \tau \in \{A,B\},\ \min\left\{P\left(\tau|\beta,h^{j-2},t_{j-1},s_j=A\right),P\left(\tau|\beta,h^{j-2},t_{j-1},s_j=B\right)\right\} > \frac{1}{2} \quad (4)$$

The probability of herding, for a scientist j, can easily be calculated by the following probability-weighted sum:

$$\text{Probability of herding} = \sum_{\{\forall h^{j-2},t_{j-1}\}} 1_H\left(h^{j-2},t_{j-1}\right)\cdot P(h^{j-2},t_{j-1}) \quad (5)$$

where $P(h^{j-2}, t_{j-1})$is the probability that (h^{j-2}, t_{j-1}) realizes from either hypothesis and 1_H is the indicator function that assumes a value of 1 if (4) holds, and 0 otherwise.

When herding occurs, some histories and information profiles will occur with probability zero. This means that there will generally be a number of terms in (3) and (5) that will never occur, so the calculations required will generally be over a smaller number of terms than the theoretical upper bound. Nevertheless, the large number of terms that result from even a moderate n are impossible to simplify to obtain a closed-form algebraic expression for either the expected misperception or the probability of herding. We therefore wrote a computer program to numerically calculate the algebraic expressions within available computing power.

Computer program. The program (code provided in the Supplementary Information) worked by building and evaluating the algebraic formulae to obtain results that are accurate up to the level of precision the computer used in its calculations (52 dp), as explained through a number of key steps described below for various values of β. The information a reviewer j has, $\left(h^{j-2},t_{j-1},s_j\right) \in \{A,B,\varnothing\}^{j-2} \times \{A,B\}^2$, is referred to as their 'information profile'.

Step 1: For each of the two possible private signals of scientist 1, $s_1 \in \{A,B\}$, a probability is set for the occurrence of that signal conditional on each of the two hypothesis $\tau \in \{A,B\}$: $P(s_1|\tau) = \beta$ if $s_1 = \tau$ and $P(s_1|\tau) = 1 - \beta$ otherwise. Thus, the posterior on the true hypothesis is calculated as: $P(\tau|s_1) = \dfrac{P(s_1|\tau)}{P(s_1|A) + P(s_1|B)}$.

Step 2: For each signal s_1 a submission decision of scientist 1 is prescribed. As $P(\tau = s_1|s_1) > 0.5$, for scientist 1 the theme of their submitted paper (t_1) will be identical to their signal (s_1). This determines the probability of $t_1 \in \{A,B\}$ conditional on $\tau \in \{A,B\}$.

Step 3: For each possible information profile $(t_1,s_2) \in \{A,B\}^2$ of scientist 2, the probability of acceptance (of scientist 1's submission with theme t_1) is determined in accordance with the adopted model. For M1 (and hence, M2), this involves calculating scientist 2's posterior beliefs as $P(\tau|t_1,s_2) = \dfrac{P(t_1,s_2|\tau)}{P(t_1,s_2|A) + P(t_1,s_2|B)}$ where $P(t_1,s_2|\tau) = P(t_1|\tau)P(s_2|\tau)$.

Step 4: If scientist 1's manuscript is rejected, a history $h^1 = \varnothing$ ensues. If accepted, a history $h^1 = t_1$ ensues. For each possible history h^1, the conditional probability $P(h^1|\tau)$ is obtained by aggregating the probabilities that it arises from different signal profiles (s_1, s_2) conditional on τ. The misperception is calculated for each history according to the formula (1), and then the expected misperception according to the formula (3).

Step 5: The submission decision of scientist 2, t_2, is equal to τ such that $P(\tau|t_1,s_2) > 0.5$ if such a τ exists; otherwise, that is, if $P(A|t_1,s_2) = P(B|t_1,s_2) = 0.5$, then $t_2 = s_2$. This determines the conditional probability $P(h^1,t_2|\tau)$. Herding (and other results) is calculated according to the relevant formulae given.

Step 6: Steps 3–5 are repeated for $j \in \{3, \ldots n+1\}$ for every possible information profile $\left(h^{j-2},t_{j-1},s_j\right)$ of scientist j with the following modifications: scientist j's posterior beliefs are $P(\tau|h^{j-2},t_{j-1},s_j) = \dfrac{P\left(h^{j-2},t_{j-1},s_j|\tau\right)}{P\left(h^{j-2},t_{j-1},s_j|A\right) + P(h^{j-2},t_{j-1},s_j|B)}$ where $P\left(h^{j-2},t_{j-1},s_j|\tau\right) = P\left(h^{j-2},t_{j-1}|\tau\right)P\left(s_j|\tau\right)$ in step 3; $h^{j-1} \in h^{j-2} \times \{A,B,\varnothing\}$ replaces h^1 and $P(h^{j-1}|\tau)$ is obtained by combining $P\left(h^{j-2},t_{j-1},s_j|\tau\right)$ and scientist j's acceptance probability given their information profile $\left(h^{j-2},t_{j-1},s_j\right)$ in step 4; and $P\left(\tau|h^{j-2},t_{j-1},s_j\right)$ and $P(h^{j-1},t_j|\tau)$ replace $P(\tau|t_1,s_2)$ and $P(h^1,t_2|\tau)$, respectively, in step 5.

Analytical results on asymptotic properties. Analytic comparison of different models is obtained asymptotically as the numbers of scientists tends to infinity. Consider M1. Let $H^n = \{A,B,\varnothing\}^n$ denote the set of all possible histories of length n, and $h^n \in H^n$ denote a history in H^n. Then, $F_n = \{\varnothing\} \cup H^1 \cup^{\cdots} \cup H^n$ for $n = 1,2,\cdots$, constitute an infinite sequence of σ-fields on H^∞.

For each h^n, let $P(h^n)$ be the *ex ante* probability that h^n will realize from either $\tau = A,B$. Let $X_n(h^n) = \dfrac{P(h^n|A)}{P(h^n|A) + P(h^n|B)}$ denote the Bayes-updated posterior belief that $\tau = A$ after h^n. Then, X_n is a random variable defined on (H^∞,F_n,P), and $\{(X_n,F_n)\}_{n=1,2,\cdots}$ constitutes a martingale. Let $Q(h^n) = P(h^n|A)$. Then, with X_n defined on (H^∞,F_n,Q), the sequence $\{(X_n,F_n)\}_{n=1,2,\cdots}$ constitutes a submartingale. By the Martingale Convergence Theorem[29], $E(X_n) \to E(X)$ almost surely where X is a random variable such that $X_n \to X$ with probability 1 and $E(\cdot)$ is taken relative to Q.

Consider a history h^n with the corresponding posterior $X_n = x < 1$. Then, there are three possible continuation histories of length $n + 1$: h^n followed by A, B, or Ø. As the manuscript of scientist $n + 1$ is accepted with a probability that is strictly between 0 and 1, (i) at least two of the three possible continuation histories realize with a strictly positive probability. Furthermore, (ii) the posteriors after these continuation histories differ, (iii) they depend on x but not on n, (iv) the distribution over these posteriors conditional on τ = A first-order stochastically dominates that conditional on $\tau = B$. Hence, $E(X_{n+1}|X_n = x) - x$ is a strictly positive constant that depends on x but not on n, and consequently, $E(X) < 1$ is not viable. As $E(X) \leq 1$, therefore, we conclude that $E(X) = 1$, that is, the posterior converges to true state with probability 1 when $\tau = A$. As a symmetric argument applies to the case that $Q(h^n) = P(h^n|B)$, that is, when $\tau = B$, the misperception converges to 0 as $n \to \infty$ under M1.

Next, consider the generalized M1 model with $\nu > 0$. As long as $0 < \nu < 1$, it is straightforward to verify that the deductions (i)–(iv) hold and, consequently, the same argument as above leads to the same conclusion that the misperception converges to 0 as $n \to \infty$. If $\nu > 1$, on the other hand, any manuscript on theme τ will be accepted with certainty once the posterior belief for the theme being true exceeds a certain threshold level which is strictly below 1. In addition, the scientists will submit on the popular theme regardless their own signal if the posterior for that theme exceeds a (different) threshold. Therefore, if the posterior belief for $\tau = A$ gets sufficiently close to 1 or 0, both the author's theme selection and the reviewer's decision are uniquely determined by the prevailing posterior independently of the scientist's own signal. Once this stage is reached, then the continuation history is uniquely determined (irrespective of whether $\tau = A$ or B) unlike (i) above and, consequently, publication outcomes reveal no further information and the posterior remains at the same level forever. Therefore, the expected misperception never converges to 0 and remains fixed at some positive level within finite time with probability 1.

For M2 and M3, by the same token the expected misperception never converges to 0 and gets stuck at some positive level once the posterior belief reaches a level such that the author's theme selection is dictated by herding independently of their own signal.

27. Bayes, T. & Price, R. An essay towards solving a problem in the doctrine of chances. *Philosophical Transactions (1683–1775)* **53,** 370–418 (1763).
28. Bikhchandani, S., Hirshleifer, D. & Welch, I. A theory of fads, fashion, custom, and cultural change in informational cascades. *J. Polit. Econ.* **100,** 992–1026 (1992).
29. Billingsley, P. *Probability and Measure* 3rd edn (John Wiley & Sons, 1995).