Normative scientific conflict is unavoidable and should be welcomed

Andrew Gelman[1]

26 Jul 2025

Abstract. The science reform movement involves several different sorts of normative conflicts involving scientific questions, research methods, and the political implications of scientific findings. After a brief review of some of the sources for scientific conflict, we consider internal and external incentives for bad scholarship and conclude with a recommendation of a fuller acceptance of normative conflicts regarding published science. We recommend a broader acceptance of normative conflict in science without expecting or even hoping that such conflict will go away.

The science reform movement involves several different sorts of normative conflicts. Most directly, there is a conflict between "reformers" who view the current system of publication and promotion as seriously flawed, versus "the establishment" who view the costs of reform as greater than its benefits. From the opposite direction, this can be framed as conflict between "doers" who want to just stick to the science and let the chips fall where they may and "critics" who want to get in the way of research. Everyone agrees some published and publicized work is in error, so the disputes arise at the level of particular papers, subfields, or reform proposals such as data sharing, preregistration, and post-publication review.

Another form of normative conflict involves politics, when weak evidence is used to promote views with potentially strong social implications. This can arise in different directions politically. Examples of unsupported claims on the left include purported large effects of early childhood intervention; see Gelman (2013) and Farran and Lipsey (2017). Examples on the right include claims of election fraud and gender-essentialist speculations about sex ratios; see Eggers, Garro, and Grimmer (2021) and Gelman and Weakliem (2009). This is not to say that implicit racism, preschool programs, election fraud, and evolutionary psychology should not be studied; rather, the point is that unavoidable scientific disputes regarding the evidence provided by p-values and the like will often get confounded with political disagreements.

Nor should we argue that political advocates should not do research. It makes sense that people who would advocate for a potential policy would want to study its effects. From a scientific perspective, what is important is to have open data and methods so that work can be independently evaluated—and for advocates to be open to the possibility that they have made

[1] Department of Statistics and Department of Political Science, Columbia University, New York, ag389@columbia.edu

mistakes. From a philosophy-of-knowledge perspective, we have to distinguish between evidence and truth. Researchers and advocates should feel free to believe that a certain pattern (for example, that a particular preschool program would have large effects were it to be applied to a new population, or that more attractive parents are more likely to have girls) while at the same time recognizing that available data are insufficient to address these beliefs empirically. Letting go of an inappropriate claim of strong evidence should not require letting go of a belief—although it should make you more welcoming of alternatives.

A scientific claim can also be controversial without having a clear political valence. For example, claims that elections are decided by irrelevant events such as college football games and shark attacks (see Fowler and Montagnes, 2015, and Fowler and Hall, 2018) can be taken to support the view that elections are not meaningful measures of public opinion, which can be a "left" or "right" view, depending on who is currently in power. Extravagant claims of the effectiveness of "nudging" or "libertarian paternalism" (Thaler and Sunstein, 2008) are vaguely on the left in the sense of providing tools for an activist government and are vaguely on the right to the extent that they would facilitate corporate control of employees. Refutations of such claims (e.g., Szászi et al., 2022) thus have politically ambiguous implications. Unsupported claims of the benefits of paleo diets or cold showers (see Gelman, 2023) have no direct political implications but have a vaguely right-leaning "manosphere" feel to them. In this era of partisan polarization, almost any scientific claim can become politically polarized—an example is the recycling of long-discredited claims of dangers of childhood vaccines, which until recently was not associated with the political left or right.

The past two decades have seen a replication crisis that has drawn attention to bad work in the social and biological sciences, including high-profile cases of failed replications, hopelessly noisy experiments, scientific implausibility, and flat-out fraud. It is not completely clear why these issues started receiving so much attention in this period (Gelman and Vazire, 2021). As with other cases of institutional failures, there is interest in tracking down what went wrong and how it can be fixed, tasks that are particularly challenging given that many of the people working on science reform, the author of the present article included, are part of the same academic structure that has rewarded, and continues to reward, fatally flawed work. We are in a position similar to the character in a film noir who is given the job to investigate (or possibly cover up; the mandate can be deliberately unclear) a crime perpetrated by a group within his organization (Farrow, 1948).

Why is there so much bad science? The simplest answer is that science is hard, and progress in a mature field will tend to be incremental. At the same time, there is a demand for a steady stream of breakthroughs. Top journals such as Science and Nature, along with news and promotional outlets such as NPR and TED, are always ready to report the latest success story. For example, in 2015 the magazine Pacific Standard was running a feature called Findings that was described as "a daily column by Pacific Standard staff writer Tom Jacobs, who scours the psychological-research journals to discover new insights into human behavior, ranging from the origins of our

political beliefs to the cultivation of creativity." Finding 365 new insights into human behavior each year—that's a lot of pressure! We came across that column when it was promoting a paper, also promoted by the British Psychology Society, claiming to find that "exposure to familiar, liked music leads to more compliance to a request implying harming a third person" (Gelman, 2015, Jacobs, 2015, Jarrett, 2015). A careful look revealed the paper to provide no strong evidence for anything—as explained in general terms by Simmons, Nelson, and Simonsohn (2011), uncontrolled researcher degrees of freedom can allow apparently strong statistical claims to be extracted from pure noise. Our point here is not to focus on this particular study but rather to demonstrate the clash between the reality of incremental, trial-and-error science and the demand for regular breakthroughs.

One might hope that in a world with millions of active scientists, enough of them will stumble across sufficiently exciting and replicable discoveries, but it doesn't have to work that way: a lot of popular-appeal studies in social and biological science are so noisy that they never have a chance to succeed. We illustrate with a study we have discussed before (Gelman and Weakliem, 2009; see also Denny, 2008). In the early 2000s, the Journal of Theoretical Biology published a series of papers with titles such as "Big and tall parents have more sons," "Engineers have more sons, nurses have more daughters," and "Violent men have more sons" (e.g., Kanazawa, 2007). Unfortunately—or, perhaps, fortunately—none of these studies had a chance of yielding any generalizable knowledge about sex ratios. The problem is mathematical. The probability of a girl birth is approximately 49%. These studies compare the sex ratios of different groups based on surveys with sample sizes ranging from 3000 to 25,000. Suppose you have a study at the high end of this range and you are comparing the proportion of girls among two groups (for example, children of tall versus short parents). The standard error of this difference is $\sqrt{(0.5^2/12500 + 0.5^2/12500)} = 0.006$: that is, six-tenths of one percentage point. For some purposes, this estimate would be very precise, but it is not so useful for studying sex ratios, given that observed population differences in sex ratios, comparing by factors such as birth order, ethnicity, and maternal age, typically very by less than one-half of one percentage point. A study where the standard error of estimation is larger than a realistic effect size is hopeless.

Indeed, this problem—experiments that through insufficient control of variation are doomed from the start—can be obscured under standard statistical frameworks. If a study has a statistical power of 10%, that says that, under certain assumptions, there is a 10% chance that it will yield a statistically significant result. Medical studies are typically required to have at least 80% power, and that makes some sense, as you would not want to put patients at risk without a good chance of success—but for a social survey or experiment that incurs zero risk and minimal cost to participants, or a reanalysis of existing data, one might think, why not conduct a study with 10% power: this represents a 1 in 10 chance of winning, and if 100 researchers at different labs are conducting such experiments, 10 will achieve discoveries, right? Actually, no, because any statistically significant finding in that setting will be unlikely to replicate in sign and magnitude (Button et al., 2013, Gelman and Carlin, 2014). Even in the absence of any questionable

research practices such as p-hacking, a scientific environment of noisy studies will lead to a regular production of statistically significant but unreplicable results.

Over the years we have become aware of many internal incentives for bad scholarship. *Career advancement* can include direct advice from Ph.D. supervisors to avoid criticizing published work; see Luebbert and Pachter (2024). *Ideological conformity* can arise from intentional acts of political advocacy or be internalized in a subfield and enforced in hiring and through social media, or through closed research communities; see Gelman (2024). The *peer-review process* reifies existing methodological and theoretical frameworks within subfields, as notoriously demonstrated by the work in social psychology published by the journal *Psychological Science* from around 2010 through 2015. *Defensive or nonexistent responses* by authors and journals allow post-publication critiques to remain in the background. *University discipline processes* are weak even in clear cases of scholarly misconduct, and there are similar *failures of scientific societies* (for example, the American Statistical Association went to the trouble of renaming a lecture associated with the politically-controversial statistician Ronald Fisher but refused to retract an award they had given to a statistician who had plagiarized multiple papers). *Norms of lack of data sharing* make it harder to vet or replicate scientific studies. The *publication process* favors big claims and promotion of outlandish ideas including, notoriously, the "Bible Code" and extrasensory perception (see Bar Hillel et al., 1999, and Ritchie et al., 2012). A *tolerance of selection*, within papers as well as selection in what is published, produces papers that only show results that appear to confirm a preferred theory, thus creating a false impression of coherence of evidence. Authors, reviewers, and journals are often willing to publish *blatantly inaccurate research summaries*, for example, a psychology study described as "long term" even though it only lasted for three days, and a paper claiming that study participants "instantly become more powerful" even though it contained no measurements of power (Hasan et al., 2013, Carney et al., 2010). There is also a *research incumbency advantage* whereby published work is assumed to be correct, which we can attribute in part to an overvaluing of the peer-review process.

Other incentives for bad scholarship come from outside academia. As noted above, *news and social media* have an insatiable appetite for new discoveries, and scientists who make the most dramatic statements are the ones who will get attention. The quest for *outside funding* motivates conformity and claims of certainty. *Political interference* can affect the content of what is published, what areas are deemed worthy of research effort, and the active promotion of discredited work (see, e.g., BBC Verify Team, 2024).

With all these incentives for bad scholarship, what hope is there for progress? Proposed science reforms are controversial for the obvious reason that researchers in some subfields stand to lose from increased scrutiny. Science reformers do their sleuthing on their own time, with very rare exceptions (Retraction Watch, 2025) and, even when a paper is refuted in the published literature, the original article can continue to get cited at a higher rate than the critical update. There is some funding for open science and increasing support for open-science practices such as preregistered replication, so maybe the best we can hope for is a new equilibrium in which

research claims are open to post-publication review, a step that shifts the burden of trust from journal editors and referees to respected third parties, who can then themselves be called into question, and so forth.

That said, we do not think the situation is hopeless. Scientific consensus has been obtained on a wide range of topics, and areas of false consensus (such as the belief, until recently, of large and consistent priming effects in social psychology) have met skepticism outside their favored subfields. Unsubstantiated claims are not going away—there is the aforementioned belief of a link between vaccines and autism, and, for that matter, astrology is still going strong after thousands of years—but such areas are recognized to be speculative at best.

The most important step regarding normative scientific conflicts may well be to recognize that they exist. Unfortunately, much of the scientific publication process seems to be structured to downplay open conflict in the final product. There is lots of fighting during the referee process which sometimes gets ugly (see, e.g., Gelman, 2022), but once a paper is published, it is typically considered indecorous to publicly criticize it, a keep-the-disputes-within-the-family attitude. This pattern—conflict while reviewing and imputed certainty afterward—can be seen as a sort of mirroring of the long-contested significance testing framework (see Krantz, 1999) in which the statistically-significant replication of a null hypothesis is taken as strong positive evidence in favor of a posited theory. The difficulty of overcoming the publication barrier is considered to elevate a manuscript into the scientific canon. Given that statistically significant patterns can routinely arise by chance and p-hacking, we see an analogy to the subprime mortgage crisis, in which large numbers of shaky investments were "tranched" to obtain a misleading impression of stability (Gelman and Loken, 2014).

Recognizing legitimate scientific and political conflicts can be a first step toward more open resolutions of such disagreements. Resolution can come from reanalyses of data, replication studies, and, ideally, reassessments by researchers and others in their subfields. The point of recognizing incentives for bad scholarship is not to excuse such behavior or even to suggest immediate reforms—just for example, we have no idea how to disincentivize researchers from publishing blatantly inaccurate research summaries of their work—but more to move away from denial of post-publication conflict. We recommend a broader acceptance of normative conflict in science without expecting or even hoping that such conflict will go away.

If we accept the desirability of open normative scientific conflict, what steps can be taken to promote it? Some steps have already been taken, for example many journals now require data and code to be shared. Journals could also post referee reports, host post-publication criticism online so it is directly accessible along with the published paper, schedule formal post-publication reviews of frequently cited papers (Gelman and King, 2025), and publish serious criticism of their papers—without holding the criticism to a higher standard than the original paper—along with responses from the original authors if they want to do so.

It should also be the norm to publish unsuccessful replication attempts, but this is tricky. In 2011, a claim of extra-sensory perception were published in a top psychology journal and featured on the front page of the *New York Times*. The failed replication attempts appeared in *PLoS One*, an archival journal that publishes just about everything. But this makes some sense: a true discovery of ESP really would have been a scientific breakthrough, whereas a null finding is entirely unsurprising. We would not want our leading scholarly journals to fill themselves with papers confirming that gravity is real, perpetual motion machines don't actually work, World War II really happened, etc. There is an uncomfortable asymmetry that false claims are often more exciting than the truth. But there is no need to demand that journals publish every replication; any given journal should only have the burden of publishing (online) all the serious attempted replications of papers it has published. This seems fair: along with the potential glory and publicity would come the responsibility to handle the aftermath.

So, we recommend that, instead of only publishing criticisms and replication attempts in exceptional circumstances, journals should publish them default unless they have obvious problems. (If there is a concern that critics and replicators could then get cheap publication credit in this way, these submissions could be labeled as such, for example as *Journal of X: Criticisms and Replications*.) And we also recommend automatic invitations of authors to respond to criticisms, with all of this published online and accessible from the original published paper.

But what is the endpoint of all this? With nobody getting the last word, and with well-known human patterns of stubbornness—not to mention the disincentives for admitting error—it is not realistic to expect this to end in any sort of satisfactory "closure." The fight has no winner and can continue indefinitely. But if we think of normative scientific conflict as a good in itself, this is fine. However much we might wish for closure in any given example, we should swallow our frustrations in service of the larger goal of facilitating conflict.

Lest this position seem to anodyne and obvious—as Karl Popper and others have emphasized, it is the nature of research for its future directions to be unexpected—let us remember that the current journal publication system is nearly the opposite of what is proposed here: as things stand now, once a paper is published, there is a push toward enforced consensus.

Here's an example. In 2024, the magazine *Daedalus* published an issue on Understanding Implicit Bias: Insights & Innovations. One of the articles was called "The Science of Implicit Race Bias: Evidence from the Implicit Association Test," another was "The Implicit Association Test," and there were others on the effects of implicit bias—but no skeptical views were presented. A couple of the articles briefly mentioned and cited dissenting takes but only very briefly and in a context where the IAT was treated as legitimate. Given the level of controversy on this topic (see, for example, Goldhill, 2017), I was surprised that all 18 of the articles in this journal issue took the same position. Shouldn't they have had a few saying, "Yeah, sure, implicit bias is a thing, it's just not a thing that you should try to measure using the Implicit Association

Test," or something like that? Or shouldn't the introduction have found space to say something like, "The Implicit Association Test is controversial and has been subjected to serious criticisms (refs.); nonetheless we include several papers that use it because . . ."?

A quick answer to that question is, No, it's their journal, they can do what they want, and it's not like this publication will dispel the controversy. The normative scientific dispute remains—in this case, with methodological, substantive, and political dimensions—, so really the only question is whether the dispute also occurs within *Daedalus* or whether that journal presents a unified front within a larger frame of disagreement. I would prefer the former, but I admit I cannot offer any evidence in favor of this view.

The unfortunate flip side of perpetually unresolved normative scientific disputes is that various incorrect and even dangerous ideas will not go away. A current example is the U.S. Secretary of Health and Human Services promoting anti-vaccine ideology based on discredited, falsified, and even completely made-up studies (Jacobs, 2025). Less immediately consequential examples include continuing beliefs in ghosts, alien visitations, and various extreme versions of social priming. The problem here is not that people are holding implausible beliefs—after all, some implausibilities turn out to be true, and one of the benefits of a large scientific community is that it has room for the pursuit of minority opinions, longshots in the communal scientific portfolio. Rather, our problem is when people misrepresent the evidence in favor of such beliefs. Science may be self-correcting, but there will always be political or personal motivations for people to misrepresent evidence, even within an economically and politically open society.

When asked if he had strong political views, the historian A. J. P. Taylor replied, "No. Extreme views weakly held." This is perhaps how we should approach science, at least as individuals. A difficulty here, for which I see no solution, is scientists play three roles, sometimes at the same time. The first role involves conjecture and a sort of scientific, but not necessarily political, advocacy: you believe an idea, you want others to come to your view, and you seek out evidence and arguments in favor of it. The second role is the self-critic, following the dictum of Feynman (1974) that "The first principle is that you must not fool yourself—and you are the easiest person to fool." When playing these two roles, you embody scientific conflict within yourself—a trait which perhaps is not as common as it should be among practicing scientists. However, there is also a third role, which is the dispassionate judge of evidence. We do need to come to tentative conclusions, and at that point some maturity is called for. Accepting and even welcoming normative scientific conflict goes along with the willingness to come to reasonable conclusions given the data at hand.

References

Maya Bar-Hillel, Dror Bar-Natan, Gil Kalai, and Brendan McKay (1999). Solving the Bible Code puzzle. *Statistical Science* 14, 150-173.

BBC Verify Team (2024). Fact-checking RFK Jr.'s views on health policy. *BBC*, 15 Nov. https://www.bbc.com/news/articles/c0mzk2y41zvo/

Katherine S. Button, John P. A. Ioannidis, Claire Mokrysz, Brian A. Nosek, Jonathan, Emma S. J. Robinson, and Marcus R. Munafò (2013). Power failure: Why small sample size undermines the reliability of neuroscience. *Nature Reviews Neuroscience* 14, 365-376.

Dana R. Carney, Amy J. C. Cuddy, and Andy J. Yap (2010). Power posing: Brief nonverbal displays affect neuroendocrine levels and risk tolerance. *Psychological Science* 21, 1363-1368.

Kevin Denny (2008). Big and tall parents do not have more sons. *Journal of Theoretical Biology* 250, 752-753.

Andrew C. Eggers, Haritz Garro, and Justin Grimmer (2021). No evidence for systematic voter fraud: A guide to statistical claims about the 2020 election. *Proceedings of the National Academy of Sciences* 118, e2103619118.

Dale C. Farran and Mark W. Lipsey (2017). Misrepresented evidence doesn't serve pre-K programs well. Brookings Institution, 24 Feb. https://www.brookings.edu/articles/misrepresented-evidence-doesnt-serve-pre-k-programs-well/

John Farrow (1948). *The Big Clock*. Paramount.

Richard P. Feynman (1974). Cargo cult science. Commencement address, California Institute of Technology. https://calteches.library.caltech.edu/51/2/CargoCult.htm

Anthony Fowler and Andrew B. Hall (2018). Do shark attacks influence presidential elections? Reassessing a prominent finding on voter competence. *Journal of Politics* 80, 1423-1437.

Anthony Fowler and B. Pablo Montagnes (2015). College football, elections, and false-positive results in observational research. *Proceedings of the National Academy of Sciences* 112, 13800-13804.

Andrew Gelman (2013). Childhood intervention and earnings. *Symposium Magazine*, 3 Nov. https://web.archive.org/web/20131118074128/http://www.symposium-magazine.com/childhood-intervention-and-earnings/

Andrew Gelman (2015). The aching desire for regular scientific breakthroughs. *Statistical Modeling, Causal Inference, and Social Science*, 16 Sep. https://statmodeling.stat.columbia.edu/2015/09/16/harsh/

Andrew Gelman (2022). How do things work at top econ journals, exactly? *Statistical Modeling, Causal Inference, and Social Science*, 25 Jan. https://statmodeling.stat.columbia.edu/2022/01/25/how-do-things-work-at-top-econ-journals-exactly-this-is-one-weird-ass-story/

Andrew Gelman (2023). Before reading this post, take a cold shower: A Stanford professor says it's "great training for the mind"! *Statistical Modeling, Causal Inference, and Social Science*, 8 Jul. https://statmodeling.stat.columbia.edu/2023/07/08/before-reading-this-post-take-a-cold-shower-a-stanford-professor-its-great-training-for-the-mind/

Andrew Gelman (2024). Implicitly denying the controversy associated with the Implicit Association Test. *Statistical Modeling, Causal Inference, and Social Science*, 20 Aug. https://statmodeling.stat.columbia.edu/2024/08/20/when-is-it-appropriate-to-give-a-one-sided-perspective-implicit-association-test-example/

Andrew Gelman and John B. Carlin (2014). Beyond power calculations: Assessing Type S (sign) and Type M (magnitude) errors. *Perspectives on Psychological Science* 9, 641-651.

Andrew Gelman and Andrew King (2025). Social science is broken. Here's how to fix it. *Chronicle of Higher Education*, 25 Feb.

Andrew Gelman and Eric Loken (2014). The AAA tranche of subprime science. *Chance* 27 (1), 51-56.

Andrew Gelman and Simine Vazire (2021). Why did it take so many decades for the behavioral sciences to develop a sense of crisis around methodology and replication? *Journal of Methods and Measurement in the Social Sciences* 12, 37-41.

Andrew Gelman and David Weakliem (2009). Of beauty, sex, and power: Statistical challenges in estimating small effects. *American Scientist* 97, 310-316.

Olivia Goldhill (2017). The world is relying on a flawed psychological test to fight racism. *Quartz*, 3 Dec. https://qz.com/1144504/the-world-is-relying-on-a-flawed-psychological-test-to-fight-racism/

Youssef Hasan, Laurent Bègue, Michael Scharkow, and Brad Bushman (2013). The more you play, the more aggressive you become: A long-term experimental study of cumulative violent video game effects on hostile expectations and aggressive behavior. *Journal of Experimental Social Psychology* 49, 224-227.

Phie Jacobs (2025). Trump officials downplay fake citations in high-profile report on children's health. *Science*, 30 May. https://www.science.org/content/article/trump-officials-downplay-fake-citations-high-profile-report-children-s-health/

Tom Jacobs (2015). The dark side of music. *Pacific Standard*, 31 Aug. https://psmag.com/social-justice/the-dark-side-of-the-power-of-music/

Christian Jarrett (2015). Background positive music increases people's willingness to do others harm. *British Psychological Society Research Digest*, 15 Sep. https://web.archive.org/web/20150917175734/http://digest.bps.org.uk/2015/09/background-positive-music-increases.html

Satoshi Kanazawa (2007). Beautiful parents have more daughters: A further implication of the generalized Trivers-Willard hypothesis. *Journal of Theoretical Biology* 244, 133-140.

David H. Krantz (1999). The null hypothesis testing controversy in psychology. *Journal of the American Statistical Association* 44, 1372-1381.

Laura Luebbert and Lior Pachter (2014). The journal of scientific integrity. *Bits of DNA*, 2 Jul. https://liorpachter.wordpress.com/2024/07/02/the-journal-of-scientific-integrity/

Retraction Watch (2025). Meet the first two Retraction Watch Sleuths in Residence. *Retraction Watch*, 27 May. https://retractionwatch.com/2025/05/27/meet-the-first-two-retraction-watch-sleuths-in-residence/

Stuart J. Ritchie, Richard Wiseman, and Christopher C. French (2012). Failing the future: Three unsuccessful attempts to replicate Bem's "retroactive facilitation of recall" effect. PLoS One 7, e33423.

Joseph P. Simmons, Leif D. Nelson, and Uri Simonsohn (2011). False-positive psychology: Undisclosed flexibility in data collection and analysis allows presenting anything as significant. *Psychological Science* 22, 1359-1366.

Barnabás Szászi, Anthony C. Higney, Aaron B. Charlton, Andrew Gelman, Ignazio Ziano, Balacs Aczel, Daniel G. Goldstein, David S. Yeager, and Elizabeth Tipton (2022). No reason to expect large and consistent effects of nudge interventions. *Proceedings of the National Academy of Sciences* 119, e2200732119.

A. J. P. Taylor (1977). Accident prone, or what happened next. *Journal of Modern History* 49, 1-18.

Richard H. Thaler and Cass R. Sunstein (2008). *Nudge: Improving Decisions About Health, Wealth, and Happiness*. Yale University Press.