FiveThirtyEight

Science Isn't Broken

It's just a hell of a lot harder than we give it credit for.

By Christie Aschwanden

Filed under Scientific Method
Published Aug 19, 2015

Graphics by Ritchie King

If you follow the headlines, your confidence in science may have taken a hit lately. Peer review? More like self-review. An investigation in November uncovered a scam in which researchers were rubber-stamping their own work, circumventing peer review at five high-profile publishers. Scientific journals? Not exactly a badge of legitimacy, given that the International Journal of Advanced Computer Technology recently accepted for publication a paper titled "Get Me Off Your Fucking Mailing List," whose text was nothing more than those seven words, repeated over and over for 10 pages. Two other journals allowed an engineer posing as Maggie Simpson and Edna Krabappel to publish a paper, "Fuzzy, Homogeneous Configurations." Revolutionary findings? Possibly fabricated. In May, a couple of University of California, Berkeley, grad students discovered irregularities in Michael LaCour's influential paper suggesting that an in-person conversation with a gay person could change how people felt about same-sex author could find no record of the data.

Taken together, headlines like these might suggest that science is a shady enterprise that spits out a bunch of dressed-up nonsense. But I’ve spent months investigating the problems hounding science, and I’ve learned that the headline-grabbing cases of misconduct and fraud are mere distractions. The state of our science is strong, but it’s plagued by a universal problem: Science is hard — really fucking hard.

If we’re going to rely on science as a means for reaching the truth — and it’s still the best tool we have — it’s important that we understand and respect just how difficult it is to get a rigorous result. I could pontificate about all the reasons why science is arduous, but instead I’m going to let you experience one of them for yourself. Welcome to the wild world of p-hacking.

FiveThirtyEight

Way To Scientific Glory

with a hunch: **The U.S. economy is affected by whether Republicans ffice.** Try to show that a connection exists, using real data going back to be publishable in an academic journal, you’ll need to prove that they ant” by achieving a low enough p-value.

Republicans	Rep.	Democrats	Dem.

-
t
-

eps.

-
-
-
-

If you tweaked the variables until you proved that Democrats are good for the economy, congrats; go vote for Hillary Clinton with a sense of purpose. But don’t go bragging

The data in our interactive tool can be narrowed and expanded (p-hacked) to make either hypothesis appear correct. That's because answering even a simple scientific question — which party is correlated with economic success — requires lots of choices that can shape the results. This doesn't mean that science is unreliable. It just means that it's more challenging than we sometimes give it credit for.

Which political party is best for the economy seems like a pretty straightforward question. But as you saw, it's much easier to get a *result* than it is to get an *answer*. The variables in the data sets you used to test your hypothesis had 1,800 possible combinations. Of these, 1,078 yielded a publishable p-value,[1] but that doesn't mean they showed that which party was in office had a strong effect on the economy. Most of them didn't.

The p-value reveals almost nothing about the strength of the evidence, yet a p-value of 0.05 has become the ticket to get into many journals. "The dominant method used [to evaluate evidence] is the p-value," said Michael Evans, a statistician at the University of Toronto, "and the p-value is well known not to work very well."

Scientists' overreliance on p-values has led at least one journal to decide it has had enough of them. In February, Basic and Applied Social Psychology announced that it will no longer publish p-values. "We believe that the p < .05 bar is too easy to pass and sometimes serves as an excuse for lower quality research,"the editors wrote in their announcement. Instead of p-values, the journal will require "strong descriptive statistics, including effect sizes."

After all, what scientists really want to know is whether their hypothesis is true, and if so, how strong the finding is. "A p-value does not give you that — it can never give you that," said Regina Nuzzo, a statistician and journalist in Washington, D.C., who wrote about the p-value problem in Nature last year. Instead, you can think of the p-value as an index of surprise. How surprising would these results be if you assumed your hypothesis was false?

As you manipulated all those variables in the p-hacking exercise above, you shaped your result by exploiting what psychologists Uri Simonsohn, Joseph Simmons and Leif Nelson call "researcher degrees of freedom," the decisions scientists make as they conduct a study. These choices include things like which observations to record, which ones to compare, which factors to control for, or, in your case, whether to measure the these calls as they go, and often there's no obviously correct way to proceed, which makes it tempting to try different things until you get the result you're looking for.

WHAT'S THE POINT: BAD INCENTIVES ARE BLOCKING GOOD SCIENCE

00:00 00:00

Subscribe to all the FiveThirtyEight podcasts.

Scientists who fiddle around like this — just about all of them do, Simonsohn told me — aren't usually committing fraud, nor are they intending to. They're just falling prey to natural human biases that lead them to tip the scales and set up studies to produce false-positive results.

Since publishing novel results can garner a scientist rewards such as tenure and jobs, there's ample incentive to p-hack. Indeed, when Simonsohn analyzed the distribution of p-values in published psychology papers, he found that they were suspiciously concentrated around 0.05. "Everybody has p-hacked at least a little bit," Simonsohn told me.

But that doesn't mean researchers are a bunch of hucksters, a la LaCour. What it means is that they're human. P-hacking and similar types of manipulations often arise from human biases. "You can do it in unconscious ways —*I've* done it in unconscious ways," Simonsohn said. "You really believe your hypothesis and you get the data and there's ambiguity about how to analyze it." When the first analysis you try doesn't spit out the result you want, you keep trying until you find one that does. (And if that doesn't work, you can always fall back on HARKing — hypothesizing after the results are known.)

Subtle (or not-so-subtle) manipulations like these plague so many studies that Stanford meta-science researcher John Ioannidis concluded, in a famous 2005 paper, that most published research findings are false. "It's really difficult to perform a good study," he told me, admitting that he has surely published incorrect findings too. "There are so many potential biases and errors and issues that can interfere with getting a reliable, he’s sworn to protect it.

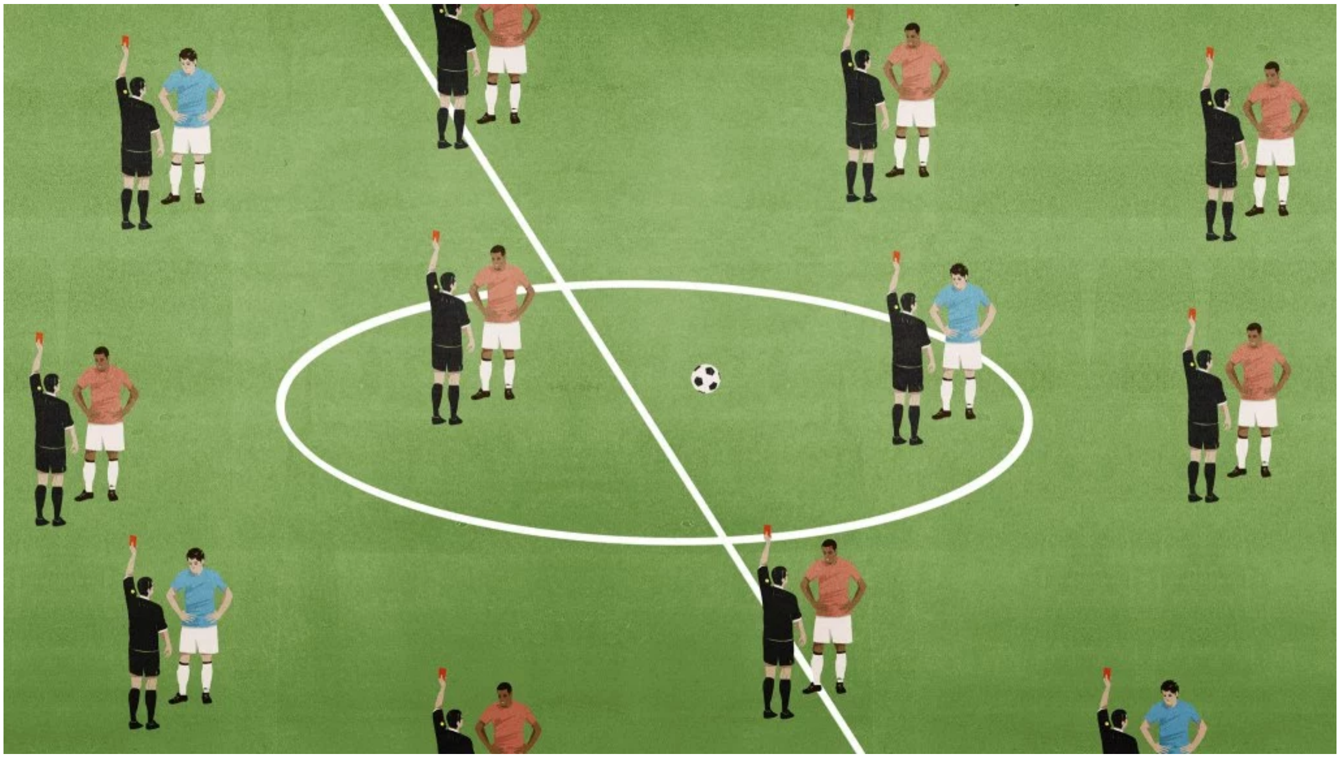


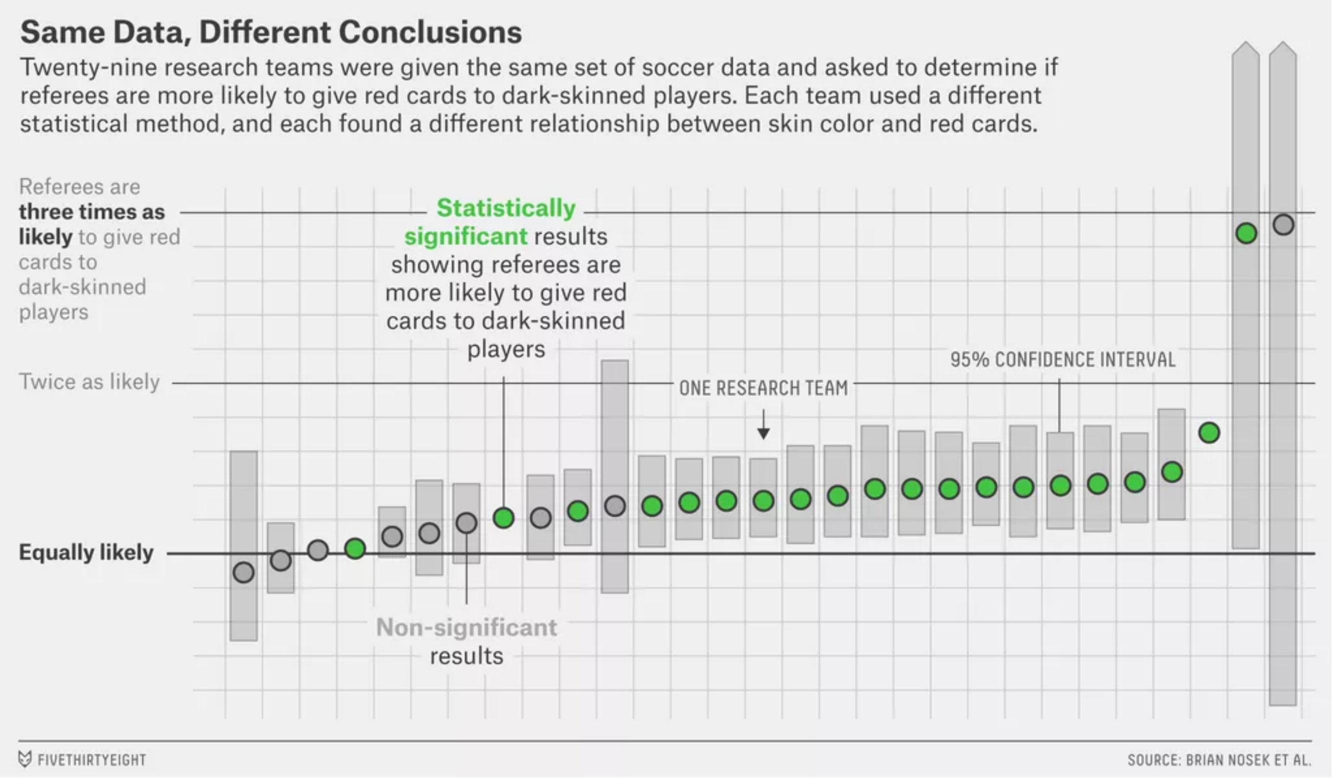
ILLUSTRATION BY SHOUT

P-hacking is generally thought of as cheating, but what if we made it compulsory instead? If the purpose of studies is to push the frontiers of knowledge, then perhaps playing around with different methods shouldn’t be thought of as a dirty trick, but encouraged as a way of exploring boundaries. A recent project spearheaded by Brian Nosek, a founder of the nonprofit Center for Open Science, offered a clever way to do this.

Nosek’s team invited researchers to take part in a crowdsourcing data analysis project. The setup was simple. Participants were all given the same data set and prompt: Do soccer referees give more red cards to dark-skinned players than light-skinned ones? They were then asked to submit their analytical approach for feedback from other teams before diving into the analysis.

Twenty-nine teams with a total of 61 analysts took part. The researchers used a wide variety of methods, ranging — for those of you interested in the methodological gore — from simple linear regression techniques to complex multilevel regressions and Bayesian in their analyses.

Despite analyzing the same data, the researchers got a variety of results. Twenty teams concluded that soccer referees gave more red cards to dark-skinned players, and nine teams found no significant relationship between skin color and red cards.



Same Data, Different Conclusions

Twenty-nine research teams were given the same set of soccer data and asked to determine if referees are more likely to give red cards to dark-skinned players. Each team used a different statistical method, and each found a different relationship between skin color and red cards.

SOURCE: BRIAN NOSEK ET AL.

The variability in results wasn't due to fraud or sloppy work. These were highly competent analysts who were motivated to find the truth, said Eric Luis Uhlmann, a psychologist at the Insead business school in Singapore and one of the project leaders. Even the most skilled researchers must make subjective choices that have a huge impact on the result they find.

But these disparate results don't mean that studies can't inch us toward truth. "On the one hand, our study shows that results are heavily reliant on analytic choices," Uhlmann told me. "On the other hand, it also suggests there's a *there* there. It's hard to look at that data and say there's no bias against dark-skinned players." Similarly, most of the permutations you could test in the study of politics and the economy produced, at best, only weak effects, which suggests that if there's a relationship between the number of Democrats or Republicans in office and the economy, it's not a strong one.

answer. Every result is a temporary truth, one that's subject to change when someone else comes along to build, test and analyze anew.

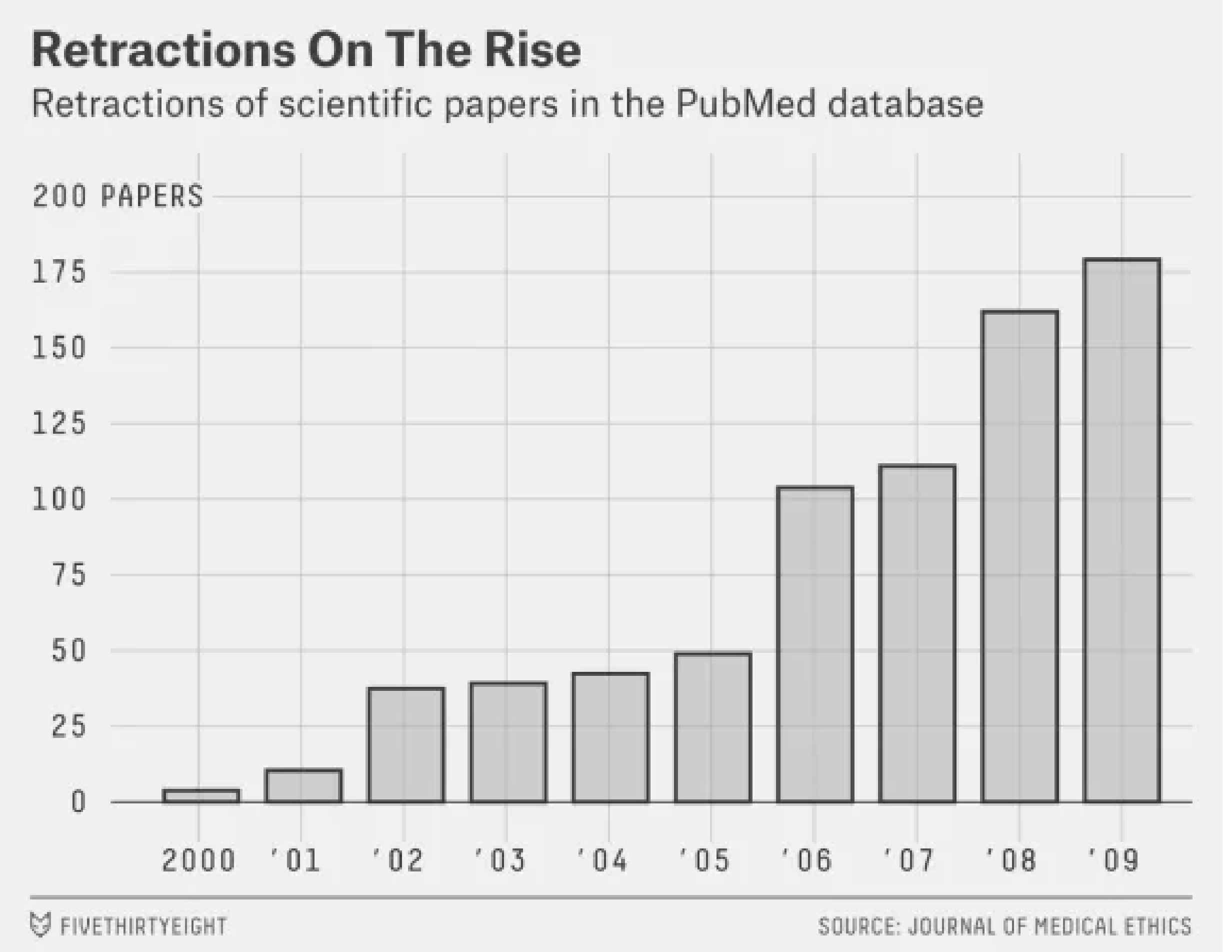
What makes science so powerful is that it's self-correcting — sure, false findings get published, but eventually new studies come along to overturn them, and the truth is revealed. At least, that's how it's supposed to work. But scientific publishing doesn't have a great track record when it comes to self-correction. In 2010, Ivan Oransky, a physician and editorial director at MedPage Today, launched a blog called Retraction Watch with Adam Marcus, managing editor of Gastroenterology & Endoscopy News and Anesthesiology News. The two had been professional acquaintances and became friendly while covering the case against Scott Reuben, an anesthesiologist who in 2009 was caught faking data in at least 21 studies.

The first Retraction Watch post was titled "Why write a blog about retractions?" Five years later, the answer seems self-evident: Because without a concerted effort to pay attention, nobody will notice what was wrong in the first place. "I thought we might do one post a month," Marcus told me. "I don't think either of us thought it would become two or three a day." But after an interview on public radio and media attention highlighting the blog's coverage of Marc Hauser, a Harvard psychologist caught fabricating data, the tips started rolling in. "What became clear is that there was a very large number of people in science who were frustrated with the way that misconduct was being handled, and these people found us very quickly," Oransky said. The site now draws 125,000 unique views each month.

While the site still focuses on retractions and corrections, it also covers broader misconduct and errors. Most importantly, "it's a platform where people can discuss and uncover instances of data fabrication," said Daniele Fanelli, a senior research scientist at Stanford's Meta-Research Innovation Center. Reader tips have helped create a surge in content, and the site now employs several staff members and is building a comprehensive, freely available database of retractions with help from a $400,000 MacArthur Foundation grant.

Marcus and Oransky contend that retractions shouldn't automatically be viewed as a stain on the scientific enterprise; instead, they signal that science is fixing its mistakes.

(rigging images from microscopes or gels, for instance, to show the desired results) are the two most common ones, Marcus told me. While outright fabrications are relatively rare, most errors aren’t just honest mistakes. A 2012 study by University of Washington microbiologist Ferric Fang and his colleagues concluded that two-thirds of retractions were due to misconduct.



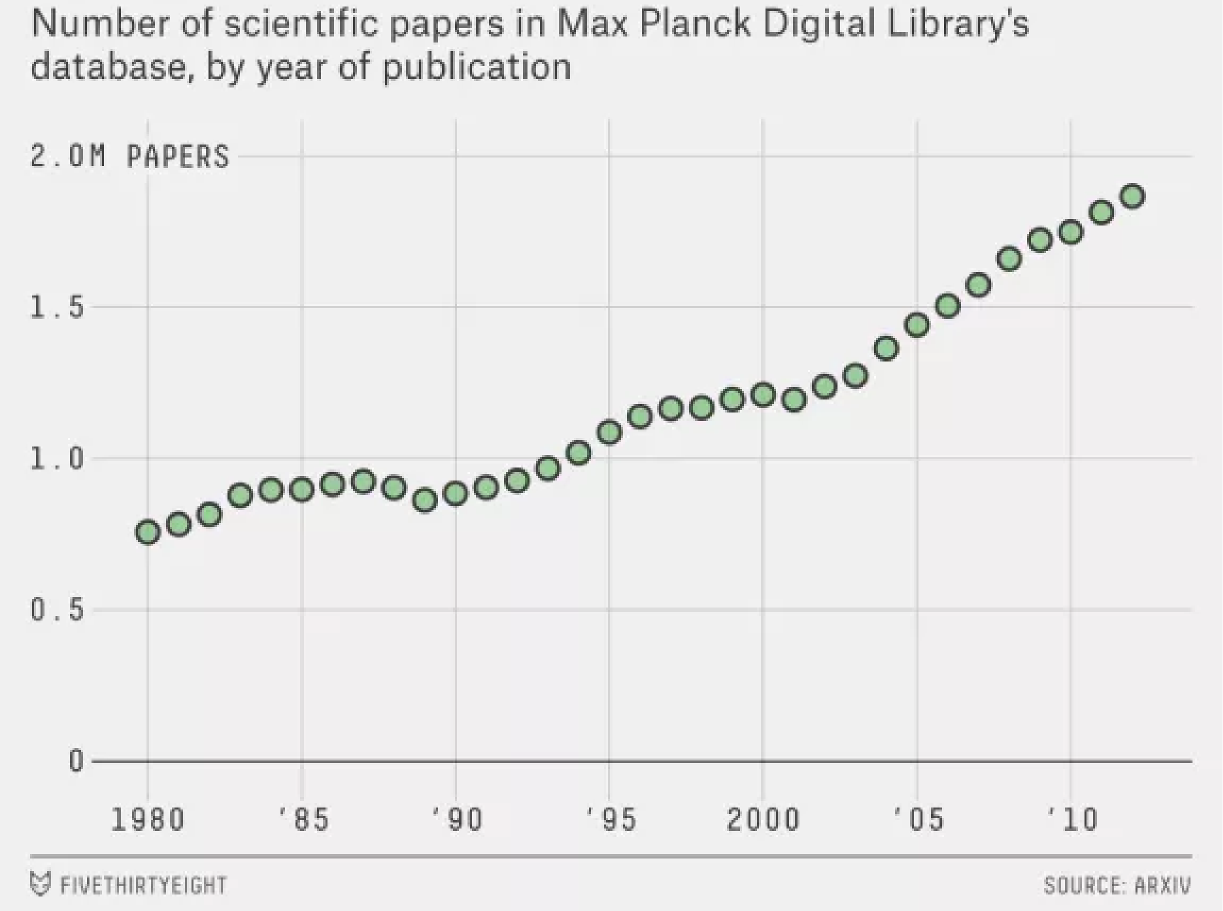
From 2001 to 2009, the number of retractions issued in the scientific literature rose tenfold. It remains a matter of debate whether that’s because misconduct is increasing or is just easier to root out. Fang suspects, based on his experiences as a journal editor, that misconduct has become more common. Others aren’t so sure. “It’s easy to show — I’ve done it — that all this growth in retractions is accounted for by the number of new journals that are retracting,” Fanelli said. Still, even with the rise in retractions, fewer than 0.02 percent of publications are retracted annually.

Peer review is supposed to protect against shoddy science, but in November, Oransky, Marcus and Cat Ferguson, then a staff writer at Retraction Watch, uncovered a ring of fraudulent peer reviewing in which some authors exploited flaws in publishers’ computer systems so they could review their own papers (and those of close colleagues).

statistical editor at the journal European Urology and a biostatistician at Memorial Sloan Kettering Cancer Center. A few years back, he decided to write up guidelines for contributors describing common statistical errors and how to avoid them. In preparation for writing the list, he and some colleagues looked back at papers their journal had already published. “We had to go back about 17 papers before we found one without an error,” he told me. His journal isn’t alone — similar problems have turned up, he said, in anesthesia, pain, pediatrics and numerous other types of journals.

Many reviewers just don’t check the methods and statistics sections of a paper, and Arthur Caplan, a medical ethicist at New York University, told me that’s partly because they’re not paid or rewarded for time-consuming peer review work.

Some studies get published with no peer review at all, as so-called “predatory publishers” flood the scientific literature with journals that are essentially fake, publishing any author who pays. Jeffrey Beall, a librarian at the University of Colorado at Denver, has compiled a list of more than 100 so-called “predatory” journal publishers. These journals often have legit-sounding names like the International Journal of Advanced Chemical Research and create opportunities for crackpots to give their unscientific views a veneer of legitimacy. (The fake “get me off your fucking mailing list” and “Simpsons” papers were published in such journals.)



Predatory journals flourish, in part, because of the sway that publication records have when it comes to landing jobs and grants, creating incentives for researchers to pad their CVs with extra papers.

But the Internet is changing the way scientists distribute and discuss their ideas and data, which may make it harder to pass off shoddy papers as good science. Today when researchers publish a study, their peers are standing by online to discuss and critique it. Sometimes comments are posted on the journal's own website in the form of "rapid responses," and new projects such as PubMed Commons and PubPeer provide forums for rapid, post-publication peer review. Discussions about new publications also commonly take place on science blogs and social media, which can help spread information about disputed or corrected results.

"One of the things we've been campaigning for is for scientists, journals and universities to stop acting as if fraud is something that never happens," Oransky told me. There are bad players in science just as there are in business and politics. "The difference is that science actually has a mechanism for self-correction. It's just that it doesn't always work." Retraction Watch's role as a watchdog has forced more accountability. The

Watch's criticisms that it hired a publications ethics manager to help its scientific record become more self-correcting. Retraction Watch has put journals on notice — if they try to retract papers without comment, they can expect to be called out. The discussion of science's shortcomings has gone public.

After the deluge of retractions, the stories of fraudsters, the false positives, and the high-profile failures to replicate landmark studies, some people have begun to ask: "Is science broken?"I've spent many months asking dozens of scientists this question, and the answer I've found is a resounding no. Science isn't broken, nor is it untrustworthy. It's just more difficult than most of us realize. We can apply more scrutiny to study designs and require more careful statistics and analytic methods, but that's only a partial solution. To make science more reliable, we need to adjust our expectations of it.

Science is not a magic wand that turns everything it touches to truth. Instead, "science operates as a procedure of uncertainty reduction," said Nosek, of the Center for Open Science. "The goal is to get less wrong over time." This concept is fundamental — whatever we know now is only our best approximation of the truth. We can never presume to have everything right.

"Science is great, but it's low-yield. Most experiments fail. That doesn't mean the challenge isn't worth it, but we can't expect every dollar to turn a positive result. Most of the things you try don't work out — that's just the nature of the process."

"By default, we're biased to try and find extreme results," Ioannidis, the Stanford meta-science researcher, told me. People want to prove something, and a negative result doesn't satisfy that craving. Ioannidis's seminal study is just one that has identified ways that scientists consciously or unconsciously tip the scales in favor of the result they're seeking, but the methodological flaws that he and other researchers have identified explain only *how* researchers arrive at false results. To get to the bottom of the problem, we have to understand *why* we're so fundamental: the biased ways that the human mind forms beliefs.

Some of these biases are helpful, at least to a point. Take, for instance, naive realism — the idea that whatever belief you hold, you believe it because it’s true. This mindset is almost essential for doing science, quantum mechanics researcher Seth Lloyd of MIT told me. “You have to believe that whatever you’re working on right now is *the* solution to give you the energy and passion you need to work.” But hypotheses are usually incorrect, and when results overturn a beloved idea, a researcher must learn from the experience and keep, as Lloyd described it, “the hopeful notion that, ‘OK, maybe that idea wasn’t right, but this next one will be.’”

“Science is great, but it’s low-yield,” Fang told me. “Most experiments fail. That doesn’t mean the challenge isn’t worth it, but we can’t expect every dollar to turn a positive result. Most of the things you try don’t work out — that’s just the nature of the process.” Rather than merely avoiding failure, we need to court truth.

Yet even in the face of overwhelming evidence, it’s hard to let go of a cherished idea, especially one a scientist has built a career on developing. And so, as anyone who’s ever tried to correct a falsehood on the Internet knows, the truth doesn’t always win, at least not initially, because we process new evidence through the lens of what we already believe. Confirmation bias can blind us to the facts; we are quick to make up our minds and slow to change them in the face of new evidence.

A few years ago, Ioannidis and some colleagues searched the scientific literature for references to two well-known epidemiological studies suggesting that vitamin E supplements might protect against cardiovascular disease. These studies were followed by several large randomized clinical trials that showed no benefit from vitamin E and one meta-analysis finding that at high doses, vitamin E actually increased the risk of death.

Human fallibilities send the scientific process hurtling in fits, starts and misdirections instead of in a straight line from question to truth.

Despite the contradictory evidence from more rigorous trials, the first studies continued to be cited and defended in the literature. Shaky claims about beta carotene’s ability to reduce cancer risk and estrogen’s role in staving off dementia also persisted, even after they’d been overturned by more definitive studies.

remove from the conventional wisdom.

Sometimes scientific ideas persist beyond the evidence because the stories we tell about them *feel* true and confirm what we already believe. It’s natural to think about possible explanations for scientific results — this is how we put them in context and ascertain how plausible they are. The problem comes when we fall so in love with these explanations that we reject the evidence refuting them.

The media is often accused of hyping studies, but scientists are prone to overstating their results too.

Take, for instance, the breakfast study. Published in 2013, it examined whether breakfast eaters weigh less than those who skip the morning meal and if breakfast could protect against obesity. Obesity researcher Andrew Brown and his colleagues found that despite more than 90 mentions of this hypothesis in published media and journals, the evidence for breakfast’s effect on body weight was tenuous and circumstantial. Yet researchers in the field seemed blind to these shortcomings, overstating the evidence and using causative language to describe associations between breakfast and obesity. The human brain is primed to find causality even where it doesn’t exist, and scientists are not immune.

As a society, our stories about how science works are also prone to error. The standard way of thinking about the scientific method is: ask a question, do a study, get an answer. But this notion is vastly oversimplified. A more common path to truth looks like this: ask a question, do a study, get a partial or ambiguous answer, then do another study, and then do another to keep testing potential hypotheses and homing in on a more complete answer. Human fallibilities send the scientific process hurtling in fits, starts and misdirections instead of in a straight line from question to truth.

Media accounts of science tend to gloss over the nuance, and it’s easy to understand why. For one thing, reporters and editors who cover science don’t always have training on how to interpret studies. And headlines that read “weak, unreplicated study finds tenuous link between certain vegetables and cancer risk” don’t fly off the newsstands or bring in the clicks as fast as ones that scream “foods that fight cancer!”

People often joke about the herky-jerky nature of science and health headlines in the media — coffee is good for you one day, bad the next — but that back and forth embodies exactly what the scientific process is all about. It’s hard to measure the impact of diet on

Isolating how coffee affects health requires lots of studies and lots of evidence, and only over time and in the course of many, many studies does the evidence start to narrow to a conclusion that's defensible. "The variation in findings should not be seen as a threat," Nosek said. "It means that scientists are working on a hard problem."

The scientific method is the most rigorous path to knowledge, but it's also messy and tough. Science deserves respect exactly because it is difficult — not because it gets everything correct on the first try. The uncertainty inherent in science doesn't mean that we can't use it to make important policies or decisions. It just means that we should remain cautious and adopt a mindset that's open to changing course if new data arises. We should make the best decisions we can with the current evidence and take care not to lose sight of its strength and degree of certainty. It's no accident that every good paper includes the phrase "more study is needed" — there is always more to learn.

CORRECTION (Aug. 19, 12:10 p.m.): An earlier version of the p-hacking interactive in this article mislabeled one of its economic variables. It was GDP, not productivity.

Footnotes

[1] A p-value less than or equal to 0.05 is considered statistically significant, at least in psychology and the biosciences. Physics and some other fields use even more stringent thresholds.