CHAPTER 9

The Susceptibility of Science

SCIENCE IS HUMANITY'S GREATEST INVENTION. OUR SPECIES HAS evolved to act within a narrow band of time scales, ranging from a few milliseconds to a few decades. We have evolved to operate within a similarly narrow band of spatial scales, ranging from micrometers to kilometers. Yet temporal and spatial scales of the universe run from unimaginably larger to incomprehensibly smaller. Science allows us to transcend these limitations. It gives us the tools to understand what happened in the first picoseconds following the Big Bang, and how the universe has evolved over the 13.7 billion years since. Science allows us to model the internal geometry of a single atom and to grapple with cosmological distances so large that we record them in light-years.

For all that it does, however, it would be a mistake to conclude that science, as practiced today, provides an unerring conduit to the heart of ultimate reality. Rather, science is a haphazard collection of institutions, norms, customs, and traditions that have developed by trial and error over the past several centuries. Science is culturally contingent on its emergence from European natural philosophy as practiced a few centuries ago, and evolutionarily contingent on the cognitive capacities of the species that practices it. Science is jury-rigged atop the evolved psychology of one particular species of ape, designed to provide members of this species with adequate incentives to work in concert toward a better understanding of the world around them. To do this, science relies on our unique human attributes—from our reasoning abilities to our modes of communication to our psychological

motivations such as curiosity and status-seeking. If honeybees practiced their own form of science, it would likely be very different from our own. These observations do not demean science, but they do suggest that there may be value to lifting it down from its pedestal so that we can inspect it and evaluate its performance.

One of the reasons that science works so well is that it is self-correcting. Every claim is open to challenge and every fact or model can be overturned in the face of evidence. But while this organized skepticism makes science perhaps the best-known methodology for cutting through bullshit, it is not an absolute guarantor against such. There's plenty of bullshit in science, some accidental and some deliberate. Many scientific discoveries accurately reflect the workings of nature, but plenty of others—the notion of a geocentric universe, Giovanni Schiaparelli's Martian canals, cold fusion as an unlimited energy source—do not. In order to investigate faulty claims such as these, you'll need to be able to understand how they originate in the scientific literature.

First we want to look at what motivates scientists when they do research. Why do they put in all the hard hours and long nights working at their craft? Are all of them drawn by insatiable curiosity to probe the mysteries of nature in solitary concentration?*

* While Saint Jerome was a historian and theologian rather than a natural philosopher, Dürer's engraving captures the solitary quest for truth in our romanticized image of the scientist.

Philosophers who study science have often viewed science through this lens, focusing on how humans might ideally come to understand the natural world, while ignoring what inspires them to do so. Philosopher of science Philip Kitcher has a wonderful pair of terms for thinking about scientists and their motivations: *epistemically pure,* and *epistemically sullied.* An idealized scientist, interested only in advancing human understanding, is epistemically pure. Sir Francis Bacon, often credited as an early architect of the modern scientific method, exhorts his followers to epistemic purity in the preface of his *Instauratio Magna:*

> Lastly, I would address one general admonition to all; that they consider what are the true ends of knowledge, and that they seek it not either for pleasure of the mind, or for contention, or for superiority to others, or for profit, or fame, or power, or any of these inferior things; but for the benefit and use of life; and that they perfect and govern it in charity.

The rest of us—including every living scientist we know—are epistemically sullied. We act from the same human motivations as everyone else. This doesn't mean that scientists are irresponsible, untrustworthy, or unethical; it just means that they have other interests beyond the pure quest for understanding. To understand how science works and where things can go wrong, we need to understand what these interests are. So before addressing bullshit in science, let's look at what motivates its stewards.

While scientists tend to be intensely curious people who love solving puzzles, they are in most respects just like everyone else, striving to earn money and win status among their peers. We scientists want to understand how the world works, but we also want to impress our friends and colleagues, win the next promotion, and—if things go really, really well—make the big time with a guest appearance on the *Daily Show* or *Last Week Tonight*. Scientists seek both truth and recognition. In particular, scientists seek recognition for being the first to make a discovery. This is known in science as the *priority rule.*

Scientists build their reputations by publicizing their findings in scientific papers, typically articles of anywhere from two to fifty pages. Their results should be novel, relevant, complete, and correct.

The paper should describe experiments or observations that have not been previously reported. Its findings should tell us something we did not previously know about the world—even if they do so merely by strengthening the support for a previous argument. The article should relate to ongoing research questions that have already been deemed of interest to the community, or make a compelling case for the importance of a new research question. To be complete, a paper ought to describe the experiment or other work in enough detail that another expert in the field could reproduce the findings. Obviously a paper should not misreport findings, draw unfounded inferences, or make false claims. Finally, a paper should be of an appropriate scale. This is a matter of convention and varies across fields, but it relates to the scope of the work required to constitute a scientific publication. An afternoon's tinkering in the laboratory is insufficient, but years of efforts are typically broken into a series of publications.

Academic science relies on the process of peer review to uphold these standards. When an author wishes to publish a paper, she submits it to a scientific journal. The journal staff then send the unpublished article to a small number of referees: other scientists who have volunteered to read the paper, assess its quality, and suggest improvements. Journals occupy various positions in a prestige hierarchy. Publications in leading journals confer far more prestige than do other publications. The best journals are widely circulated, widely read, and set a high bar for the quality and importance of the articles that they publish. Other journals cater to narrower niche readership, and authors can often get papers accepted in these venues after they are rejected from the highest tier. There is even a tier of very low-quality journals that will publish more or less anything, often for a price.

Unlike industrial science, in which scientists' processes and discoveries are closely guarded, academic scientists compete to get their work into print, battle for attention on Twitter and Facebook, and vie to speak at conferences. Rewarding prestige instead of direct output is a wonderful trick for allowing a broad community of researchers to work together efficiently with minimal duplication of effort.

Findings are readily overturned when other researchers are unable to replicate them. In 1989, prominent electrochemists Martin Fleischmann and Stanley Pons announced that they had discovered cold fu-

sion. In experiments, they combined a heavy isotope of water, known as deuterium, with palladium metal, and applied an electric current. They observed that their apparatus seemed to periodically generate more energy, in the form of heat, than they had put into their system. They speculated that this was a consequence of room-temperature nuclear fusion occurring among deuterium molecules within the palladium metal. If they had been right, their discovery would have provided a world-changing, readily accessible source of clean power. But, sadly, this was not to be. Many leading research labs attempted to replicate their findings, but none could. Within the year, the research community determined that cold fusion was not a real phenomenon, and the issue was put to rest.

Even science's deepest foundations can be questioned, and at times replaced, when they prove incompatible with current discoveries. Geneticists and evolutionary biologists had long assumed that genes were the sole molecular vehicles of inheritance. Offspring resemble their parents because they share the same DNA sequences in their genomes. But when genetic sequencing became inexpensive and new molecular biology techniques gave us ways to measure how genes were being activated, strong evidence began accumulating that this was not the full picture. In addition to passing their genes to their offspring, parents sometimes pass a second layer of nongenetic information about what genes to activate, when. This became known as epigenetics. Because our scientific understanding of the world can change in light of new evidence, science has proven resilient to the occasional wrong turn and even to deliberate misdirection in the form of scientific fraud.

Around the turn of the twenty-first century, replication problems began to crop up in a number of fields at unexpectedly high rates. Occasionally these problems were the result of fraud or incompetence, but far more often there were no simple explanations. Solid work by respected investigators simply could not be replicated.

In 2012, a paper in the leading scientific journal *Nature* suggested that replication failures might be more common than most suspected. Authors C. Glenn Begley and Lee Ellis reported that scientists working in a commercial lab were able to reproduce only 6 of 53 important cancer biology studies published in the recent scientific literature. Shortly thereafter, the Open Science Collaboration—a large-scale

collective effort among dozens of researchers—reported that they were able to replicate only 39 out of 100 high-profile experiments in social psychology. In experimental economics, meanwhile, a similar effort was under way. One study revealed that only 11 of 18 experimental papers published in the very best economics journals could be replicated. Was science, one of our most trusted institutions, somehow generating inadvertent bullshit on a massive scale? And if so, why?

Scientific results could be irreproducible for a number of reasons. Perhaps the most obvious is outright fraud. If researchers have faked their data, we wouldn't expect to be able to replicate their experiments. Fraud generates enormous publicity, which can give a misleading impression of its frequency.* But outright fraud is rare. It might explain why one study in a thousand can't be replicated. It doesn't explain why half of the results in fields are irreproducible. How then do we explain the replication crisis? To answer this, it is helpful to take a detour and look at a statistic known as a *p*-value.

THE PROSECUTOR'S FALLACY

As we've seen, most scientific studies look to patterns in data to make inferences about the world. But how can we distinguish a pattern from random noise? And how do we quantify how strong a particular pattern is? While there are a number of ways to draw these

* Just a few examples from the past decade: Social psychologist Diedrich Stapel became a superstar within his field by concocting fake data for dozens of experiments that he never actually conducted. His misdeeds were striking in their sheer scale. He now has fifty-eight retracted papers, with more likely to follow. Readers may be interested in *Faking Science*, Stapel's exploration of how his own psychology led him to this large-scale fraud. As an author, Stapel appears disarmingly honest in exploring his own flaws, so the book is hard to put down. But as the socially skilled perpetrator of such a massive-scale fraud, how can we not view Stapel as something like the ultimate unreliable narrator?

The highest echelons of Japanese biomedical research were deeply shaken when other researchers failed to replicate work by young research superstar Haruko Obokata. An investigation revealed that Obokata had inappropriately manipulated images in her scientific papers, raising false hopes for stem cell researchers everywhere. In a tragic turn, Obokata's supervisor, Yoshiki Sasai, subsequently committed suicide amid the fallout from the scandal.

UCLA political scientist Michael LaCour misled the scientific community, including his coauthor, with fabricated data suggesting that direct interaction with gay canvassers can have a drastic effect on attitudes toward gay marriage. The work was widely reported by national and international media outlets before the deception was uncovered.

distinctions, the most common is the use of *p-values*. Loosely speaking, a *p*-value tells us how likely it is that the pattern we've seen could have arisen by chance alone. If that is highly unlikely, we say that the result is *statistically significant*. But what does this really mean, and how should we interpret it? We will explore these questions by means of a short story.

Imagine that you are a prominent defense attorney, defending a mild-mannered biologist against the charge that he has committed the greatest art heist of the new century.

The crime was a spectacular one. A wealthy collector shipped her private collection of thirty European masterpieces, by guarded railcar, from her home in Santa Clara to an auction house in New York City. When the train reached its destination, the boxes were taken to the auction house to be unpacked. The journey had passed uneventfully and the boxes appeared untouched, but to everyone's shock, the four most valuable paintings had been cut from their frames and were missing! Neither the police nor the insurance investigators were able to find a single clue—save for one fingerprint on the frame of one of the missing paintings. The stolen paintings were never found.

Without any other leads, the police ran the fingerprint through the FBI's massive new fingerprint database until they found a match: your client. (He had provided his fingerprints to the Transportation Safety Administration in exchange for the convenience of leaving his shoes on when passing through airport security.) Upon questioning, your client was discovered to have no alibi: He claimed that he had been cut off from human contact for two weeks, tracking radio-collared grouse through the High Sierras as part of a research project.

Lack of alibi notwithstanding, you are convinced that your client could not be the culprit. He is more of a bumbling academic than a steely-nerved art thief. He already has two NSF grants to fund his work on grouse breeding and seems to have little desire for additional income. And as best as you can tell, he doesn't know a damn thing about art—he thinks Donatello is a Ninja Turtle, for goodness' sake.

Yet the case goes to trial, and you face a brilliant up-and-coming prosecutor. After presenting all of the other evidence against your client—weak and circumstantial as it may be—she moves on to her trump card, the fingerprint. She describes the computerized finger-

print matching process to the jury, and closes her statement by declaring "There is absolutely no chance of getting a fingerprint match this good simply by accident."

You counter. "You said there is absolutely no chance of a match this good. That can't be possible. Every test has at least some chance of being mistaken."

"Well, sure," she concedes, "the test *could* be mistaken in principle. But for all practical purposes, there is zero chance of it actually happening. The FBI's own studies have shown a one-in-ten-million chance that two different fingerprints would match as well as the ones we have here. One in ten million: That's certainly far beyond any reasonable doubt!"

This is just what you were looking for. You turn to the jury and sketch out a two-by-two table on a large pad of paper. It looks something like this:

	MATCH	NO MATCH
Guilty		
Innocent		

"I think we all agree," you continue, "that this crime was indeed committed by someone. And let us also assume that the true culprit is in the fingerprint database. He might not be," you say to the prosecutor, "but that would only weaken your case. So let's assume for now that he or she is included."

The prosecutor nods.

"Then the table looks something like this." You fill in the top row with a large red marker.

	MATCH	NO MATCH
Guilty	1 person	0 people
Innocent		

Turning to the prosecutor, you inquire, "Now how many people are in that FBI database of yours?"

She breaks in: "Objection, Your Honor! How is that possibly relevant?"

"Actually, it's the heart of the matter," you respond. "I think I'll make that very clear in the next few minutes."

"Objection overruled."

Your opponent concedes that with all US criminal prints, all civil background-check prints, and all of the prints from the TSA, there are approximately fifty million Americans represented in the database. And of course the majority of those people will have prints that do not match the fingerprint on the painting's frame.

"So now we can fill in more of the table," you respond. You add a 50,000,000 to the lower right corner.

	MATCH	NO MATCH
Guilty	1	0
Innocent		50,000,000

Now you point at the lower left corner—innocent people who nonetheless match—and you ask, "What do you think goes here?"

You look right at the prosecutor. "You said that there is a one in ten million chance that the algorithm finds a false match. That means that with fifty million people in the database, there should be about five people who match the fingerprint found at the scene of the crime. So we can complete our table as follows."

	MATCH	NO MATCH
Guilty	1	0
Innocent	5	50,000,000

"So look at this," you say, turning to the jury. "The prosecutor is trying to distract you by calling your attention to this comparison." You point along the bottom row (shaded above). "There is a one-in-ten-million chance of seeing a match by accident. But that's not relevant to what we're doing here in the courtroom. We don't care what the chance is of having a match, given that my client is innocent. We already *know* that we've got a match."

	MATCH	NO MATCH
Guilty	1	0
Innocent	5	50,000,000

"No, we want to know what chance there is that my client is innocent given that we've got a match." Now you point to the left-hand column. "That is a completely different matter, and it is illustrated by the left-hand column. We expect there to be about five innocent matches in the database, and one guilty person. So given a match, there is about a one-sixth chance that the accused actually committed the crime.

"Now, I can't prove beyond a shadow of a doubt that my client is innocent. All I've got is his word that he was following . . . dammit, what kind of bird was that again? Never mind. The point is that I don't have to *prove* that my client is innocent. Here in America, he is innocent until presumed guilty, and the standard of proof for guilt is 'beyond a reasonable doubt.' If there are five chances in six that my client is innocent, we're clearly nowhere near that standard. You must acquit."

In the story we've just told, your argument is not sophistry; it is correct. If your client had been found simply by scanning through the FBI's database until a match was made, there would be something like a five in six chance that he did not leave the fingerprint.*

You and the prosecutor have each stressed different *conditional probabilities*. A conditional probability is the chance that something is true, given other information. The prosecutor asks what the chance

* While our story here is fiction, our society will be confronting similar issues with increasing frequency as larger and larger databases of DNA evidence become available. Such information is particularly useful to investigators—but also particularly problematic from a false-accusation perspective—because the suspect himself or herself need not be in the database. His or her identity can be triangulated from DNA contributed by relatives. In one prominent case, the so-called Golden State Killer was identified in 2018 by screening a DNA sample against about two hundred thousand entries contributed voluntarily to a genealogy website. Before finding what appears to be the correct match, investigators initially fingered the wrong suspect who happened to share rare genes with the killer. Facial recognition systems suffer from similar problems. One system, tested by London's Metropolitan Police, was described as having a 0.1 percent error rate based on its rate of correctly identifying true negatives. But only eight of the twenty-two suspects apprehended by using the system were true positives, for a whopping error rate of 64 percent among the positive results returned.

is that there is a false match, given an innocent person chosen at random.* We can write this as P(match|innocent). You are asking the converse: What is the chance that your client is innocent, given that there is a match, which we write as P(innocent|match). People often assume that these two probabilities must be the same, but that is not true. In our example, P(match|innocent) = 1/10,000,000 whereas P(innocent|match) = 5/6.

This confusion is so common that it has its own name, the prosecutor's fallacy. Our story illustrates why. It can be a matter of life and death in the courtroom, but it's also a common source of confusion when interpreting the results of scientific studies.

When Carl was a kid, he and his friends wanted to believe that their minds might have unrecognized powers, so they tried experimenting with mind reading and ESP. One time he opened a deck of playing cards, shuffled thoroughly, and had a friend turn over the cards one by one. Carl tried to call out the unseen suit of each card. He was hopeless—nowhere near 100 percent—so they quickly gave up. But suppose we wanted to go back today and analyze the results from that little experiment. It wouldn't take anything like 100 percent success to suggest something interesting was happening. Because there are four suits of cards, you would expect to get only one in four of the guesses right. If instead you got one in three correct, that might suggest that something curious was going on. But it's not at all obvious how much better than one in four Carl would have had to do to suggest that there was something other than random guessing at play. Suppose he correctly guessed the suit of 19 of the 52 cards. That's more than expected; on average one would guess only 13 correctly. But is 19 different enough from 13 to be meaningful?

This is where *p*-values come in. We can think of our task as trying to distinguish between two different hypotheses. The first, called the *null hypothesis* and written H0, is that Carl's guesses were no better than random. The second, called the *alternative hypothesis* and written H1, is that Carl was able to identify the cards at a rate higher than chance. The *p*-value associated with the experiment tells us how likely

* Note straight off that your client isn't a randomly chosen person; he was chosen specifically because the computer found a match.

it is that by guessing at random Carl would get 19 or more cards correct. We can use statistical theory to calculate the appropriate *p*-value. In this particular case, it turns out that there is only a 4.3 percent chance that one would do this well at random,* so we say that the *p*-value is 0.043.

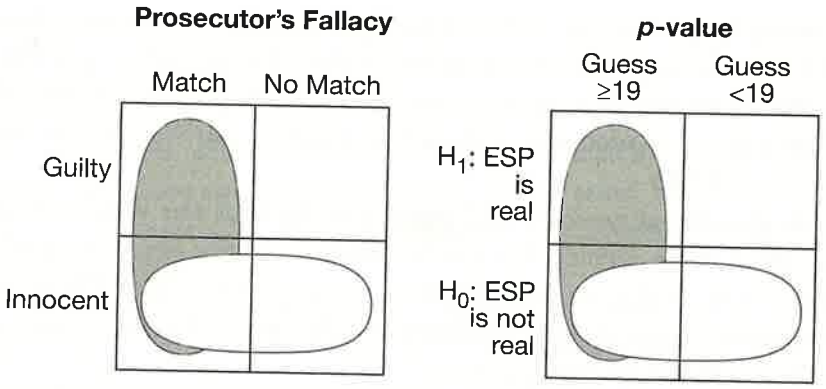
So guessing at random, 95.7 percent of the time one would fail to get 19 or more correct. But the key thing to notice here is that this does not mean that we are 95.7 percent sure that H0 is false. These are two very different claims—and they closely parallel the two claims we looked at in the art heist example.

In the courtroom drama, the prosecutor calls the jury's attention to the probability of a match by chance, i.e., given that the client is innocent: one in ten million. This is analogous to the probability of getting 19 or more cards right simply by chance. But in the courtroom, we already knew our client matched, and in the ESP example, we already know that Carl got 19 cards right. So these probabilities are not the ones we want to calculate. Rather, we want to figure out what to believe after the experiment is done. In the courtroom we wanted to determine the probability that your client was guilty given a match. To assess the results of the mind-reading test we want to know the probability that something other than random chance was responsible for Carl's score.

So here's the dirty secret about *p*-values in science. When scientists report *p*-values, they're doing something a bit like the prosecutor did in reporting the chance of an innocent person matching the fingerprint from the crime scene. They would like to know the probability that their null hypothesis is wrong, in light of the data they have observed. But that's not what a *p*-value is. A *p*-value describes the probability of getting data at least as extreme as those observed, if the null hypothesis were true. Unlike the prosecutor, scientists aren't trying to trick anybody when they report this. Scientists are stuck using *p*-values because they don't have a good way to calculate the probability of the alternative hypothesis.

The image following illustrates the parallel. What the jury wants to know is the probability that the defendant is innocent, given the

* This assumes that the deck is shuffled between each guess, so that Carl can't gain a slight edge by counting cards.



match. This is a matter of comparing the probabilities in the shaded vertical oval. Instead, the prosecutor tells them the probability of a match, given that the defendant is innocent. This is the comparison in the unshaded horizontal oval. Scientific *p*-values are the same. We want to know what the probability is that ESP is not real, given that Carl guessed at least 19 cards right. The shaded vertical oval illustrates this comparison. Instead, we are told the probability that Carl would guess at least 19 cards right, assuming that ESP is not real. The unshaded horizontal oval represents this comparison.

Why is it difficult to calculate the probability of the alternative hypothesis? That probability depends on how likely we thought it was before we did the experiment—and people seldom agree on that issue. Suppose the alternative explanation to random chance is:

> H1a: *There really is such a thing as ESP, but despite years of looking, occultists and spiritualists and even scientists never managed to observe it until two kids came along and did an experiment with playing cards in a suburban living room in Ann Arbor, Michigan, in the late 1970s.*

How likely would you have thought this H1a was before you knew that Carl correctly guessed 19 of 52 cards? A one-in-a-million chance? One in a billion? One in a trillion? Something extremely improbable, in any case. Given that, even after you know that he guessed 19 of 52 cards correctly, you have little reason to suspect that H1 is true. The odds that Carl and his friend were the first to observe human telepathy

are vanishingly small, whereas at nearly 5 percent the probability of getting 19 right by chance is far higher.

But suppose instead that we consider a different alternative hypothesis for how Carl did so well.

> H1b: *Carl's friend wanted to make him believe in ESP, so he sometimes reported that Carl had guessed correctly, even when he hadn't.*

If you'd known Carl's friend (who cheated at everything from basketball to solitaire), you would almost expect him to do something like this. So in this case, after you found out that Carl guessed 19 cards right, you'd say "That figures—Arnie probably lied about his score." The point here is that the probability of the alternative hypothesis after we see the data, $P(H_1 \mid \text{data})$, depends on the probability of the alternative hypothesis before we see the data, and that's something that isn't readily measured and incorporated into a scientific paper. So as scientists we do what we can do instead of what we want to do. We report P(data at least as extreme as we observed $\mid H_0$), and this is what we call a *p*-value.

So what does all of this have to do with bullshit? Well, even scientists sometimes get confused about what *p*-values mean. Moreover, as scientific results pass from the scientific literature into press releases, newspapers, magazines, television programs, and so forth, *p*-values are often described inaccurately. For example, in 2012 scientists using the Large Hadron Collider in Geneva reported exciting results that supported the existence of the Higgs boson, an elementary particle that had long been predicted but never observed directly. Reporting on the story, *National Geographic* wrote that scientists "are more than 99 percent certain they've discovered the Higgs boson, aka the God particle—or at least a brand-new particle exactly where they expected the Higgs to be." What *National Geographic* should have reported is that the *p*-value for the experiment was 0.01. The results obtained with the Large Hadron Collider would have had a 1 percent probability of appearing by chance even if there had been such a thing as a Higgs boson. This does not mean that scientists were 99 percent sure that the Higgs boson was real. In the case of the Higgs boson, there were already good reasons to expect that the Higgs boson would

exist, and its existence was subsequently confirmed. But this is not always the case.* The important thing to remember is that a very unlikely hypothesis remains unlikely even after someone obtains experimental results with a very low *p*-value.

P-HACKING AND PUBLICATION BIAS

Purely as a matter of convention, we often use a *p*-value of 0.05 as a cutoff for saying that a result is *statistically significant*.† In other words, a result is statistically significant when $p < 0.05$, i.e., when it would have less than 5 percent probability of arising due to chance alone.

Researchers are more interested in reading about statistically significant "positive" results than nonsignificant "negative" results, so both authors and journals have strong incentives to present significant results. Why are researchers and journals uninterested in negative results? It's not entirely clear to us, but there are a number of plausible contributing factors. Some of it may lie in our own psychology; to most of us negative results feel sort of boring. "These two groups don't differ." "This treatment doesn't change the outcome." "Knowing x doesn't help us predict y." Reading sentences like these, we feel like we're right back where we started, instead of learning something interesting about the world.

Negative results may also be associated with the inability to carry out a technical experiment. When Carl was working in a microbiology lab, he was often unable to grow his study organism *E. coli* on an agar plate. This wasn't an interesting scientific result; it just demonstrated his utter incompetence in a laboratory environment.

* In 2019, CNN reported on a physics paper that claimed to have evidence for a fifth fundamental force of nature. Referring to a *p*-value of one in one trillion, the article alleges that "there was only a one-in-a-trillion chance that the results were caused by anything other than the X17 particle, and this new fifth force." This is completely wrong. As in the Higgs boson example, the statement confuses the *p*-value for the probability of the hypothesis. But here it does so with a stronger *p*-value in support of a hypothesis that scientists find considerably less likely. Worse still, it implies that the existence of a fifth fundamental force is the *only* possible explanation for the results other than experimental error. That is not a valid inference. Something out of the ordinary seems to have happened in the experiment, but rejecting the null hypothesis does not guarantee that a researcher's favorite alternative hypothesis is true.

† Since the choice of 0.05 as the threshold for statistical significance is an arbitrary one, there is no reason it could not be set to some other value. A debate has recently emerged in the scientific literature about whether the cutoff value of 0.05 is a reasonable choice, or whether we should be using a more stringent cutoff such as 0.005.

A third possibility is that negative propositions are a dime a dozen. It's easy to state hypotheses that aren't true. Assemble words into sentences at random, and they'll usually be false. "Tulips bite." "Snowflakes melt iron." "Elephants are birds." Finding true statements in this sea of false ones is like looking for a needle of meaning in a haystack of nonsense. Think of it like the old board game Battleship. Most of the spaces on the grid are open water, so you don't learn much when you miss. But when you hit, you learn a great deal—and you can work from there to learn even more.

For all of these reasons, negative results simply do not get a huge amount of attention. We've never seen someone secure a job or win a prize by giving a lecture solely about the things she couldn't get to work in the lab.

Very few scientists would commit scientific fraud to get the *p*-values they want, but there are numerous gray areas that still undermine the integrity of the scientific process. Researchers sometimes try different statistical assumptions or tests until they find a way to nudge their *p*-values across that critical $p = 0.05$ threshold of statistical significance. This is known as *p*-hacking, and it's a serious problem. Or they will alter the outcomes they are testing. A clinical trial may set out to measure the effect of a new drug on survival after five years, but after finding no change, the researchers might mine the data and pull out an apparent improvement after three years in quality of life.

When you analyze data you have collected, you often get to make a large set of choices about what exactly to include in your study. For example, suppose I want to study how election results influence painkiller consumption in the US. I might tabulate election results, collect survey reports about painkiller use, and acquire data on painkiller sales over time. There are many degrees of freedom here. What elections do I look at? US president, US senator, US representative, state governor, state senator, state representative, mayor, city council member, etc.? Do I look at consumption by men, women, or both? Consumption by young adults, middle-aged adults, those over sixty-five, teens, all of the above? Am I looking at the effect of Democratic versus Republican candidates taking office, or am I looking at the effects of having a preferred versus nonpreferred candidate in office? In other words, do I control for the political stance of the user? And what counts as a painkiller? Aspirin, Advil, Tyle-

nol, hydrocodone, OxyContin? Do I want to compare painkiller use in one location before and after the election, or do I want to look only after the election and compare different locations? There are a huge number of decisions I need to make before I can analyze my data. Given the many combinations, the likelihood is high that at least one of those combinations will show a statistically significant result—even if there is no causal relationship between election results and painkiller use.

To avoid this pitfall, researchers ought to specify all of these choices before they look at the data, and then test the one hypothesis they've committed to in advance.* For example, I might decide to test whether adult men and women of voting age take more painkillers after their preferred gubernatorial candidate has lost an election. Or I might test to see whether Children's Tylenol sales drop in districts where a Republican is elected to replace a Democrat in the US House of Representatives. Whatever it is that I choose to look at, the important thing is that I specify it clearly before I analyze the data. Otherwise by looking at enough different hypotheses, I'll always find some significant results, even if there are no real patterns.

But look at it from the researcher's perspective. Imagine that you have just spent months collecting a massive data set. You test your main hypothesis and end up with results that are promising—but nonsignificant. You know you won't be able to publish your work in a good journal this way, and you may not be able to publish at all. But surely the hypothesis is true, you think—maybe you just don't have quite enough data. So you keep collecting data until your p-value drops below 0.05, and then stop right away lest it drift back above that threshold.

Or maybe you try a few other statistical tests. Since the data are close to significant, the right choice of measures and of tests might allow you to clear the critical $p = 0.05$ bar. Sure enough, with a bit of tinkering, you find an approach that gives you a significant result.

Or maybe your hypothesis appears to be correct only for men, and the meaningful pattern is getting swamped out by including women

* If researchers want to test multiple hypotheses, there are statistical methods such as the Bonferroni correction that allow this. Each individual test then requires stronger evidence to be considered significant, so that there is roughly a one-in-twenty chance that *any* of the tested hypotheses would appear significant if the null hypothesis were true.

in your sample. You look—and lo and behold, once you look at men only, you've got a statistically significant result. What are you going to do? Scrap the whole project, abandon thousands of dollars invested, ask your graduate student to delay her graduation by another six months . . . or just write up the result for men and submit to a top journal? Under those circumstances it may feel easy to rationalize the latter approach. "I'm sure the trend is really there," you might tell yourself. "I was thinking about omitting women from the study right from the start."

Congratulations. You've just *p*-hacked your study.*

IMAGINE A THOUSAND RESEARCHERS of unimpeachable integrity, all of whom refuse to *p*-hack under any circumstances. These virtuous scholars test a thousand hypotheses about relationships between political victories and analgesic use, all of which are false. Simply by

* To illustrate how powerful *p*-hacking techniques can be, Joseph Simmons and colleagues Leif Nelson and Uri Simonsohn tested a pair of hypotheses they were pretty sure were untrue. One was an unlikely hypothesis; the other was impossible.

The unlikely hypothesis was that listening to children's music makes people feel older than they really are. Volunteers listened to either a children's song or a control song, and later were asked how old they felt. With a bit of *p*-hacking, the researchers concluded that listening to a children's song makes people feel older, with statistical significance at the $p < 0.05$ level.

While suggestive, the initial study was not the most persuasive demonstration of how *p*-hacking can mislead. Maybe listening to a children's song really does make you feel old. So the authors raised the bar and tested a hypothesis that couldn't possibly be true. They hypothesized that listening to the classic Beatles song "When I'm Sixty-Four" doesn't just make people feel younger, it literally makes them younger. Obviously this is ridiculous, but they conducted a scientific experiment testing it anyway. They ran a randomized controlled trial in which they had each subject listen either to the Beatles song or to a control song. Remarkably, they found that while people who listened to each song should have been the same age, people who heard "When I'm Sixty-Four" were, on average, a year and a half younger than people who heard the control. Moreover, this difference was significant at the $p < 0.05$ level! Because the study was a randomized controlled trial, the usual inference would be that the treatment—listening to the song—had a causal effect on age. Thus the researchers could claim (albeit tongue in cheek) to have evidence that listening to "When I'm Sixty-Four" actually makes people younger.

To reach these impossible conclusions, the researchers deliberately *p*-hacked their study in multiple ways. They collected information about a number of characteristics of their study subjects, and then controlled for the one that happened to give them the result they were looking at. (It was the age of the subject's father, for what that's worth.) They also continued the experiment until they got a significant result, rather than predetermining the sample size. But such decisions would be hidden in a scientific report if the authors chose to do so. They could simply list the final sample size without acknowledging that it was not set in advance, and they could report controlling for the father's age without acknowledging that they had also collected several additional pieces of personal information, which they ended up discarding because they did not give the desired result.

The paper makes a compelling case. If *p*-hacking can reverse the flow of time, what can it not do?

chance, roughly fifty of these hypotheses will be statistically supported at the $p = 0.05$ level. The fifty lucky researchers will write up their results and send them to journals where they are accepted and published. Of the other 950 researchers, only a handful will bother to write up their negative results, and of these only a few will be able to get their negative results published.

When a reader comes along and looks at the literature, she'll see fifty studies showing links between political outcomes and painkiller consumption, and perhaps a handful that find no association. It would be natural for her to conclude that politics heavily influences the use of painkillers, and the studies that didn't work must have simply measured the wrong quantities or looked for the wrong patterns. But the reality is the other way around. There was no relationship. The appearance of a connection is purely an artifact of what kinds of results are considered worth publishing.

The fundamental problem here is that the chance that a paper gets published is not independent of the *p*-value that it reports. As a consequence, we slam head-on into a selection bias problem. The set of published papers represents a *biased sample* of the set of all experiments conducted. Significant results are strongly overrepresented in the literature, and nonsignificant results are underrepresented. The data from experiments that generated nonsignificant results end up in scientists' file cabinets (or file systems, these days). This is what is sometimes called the *file drawer effect*.

Remember Goodhart's law? "When a measure becomes a target, it ceases to be a good measure." In a sense this is what has happened with *p*-values. Because a *p*-value lower than 0.05 has become essential for publication, *p*-values no longer serve as a good measure of statistical support. If scientific papers were published irrespective of *p*-values, these values would remain useful measures of the degree of statistical support for rejecting a null hypothesis. But since journals have a strong preference for papers with *p*-values below 0.05, *p*-values no longer serve their original purpose.*

In 2005, the epidemiologist John Ioannidis summarized the consequences of the file drawer effect in an article with the provocative title

* This is not a new insight. The statistician Theodore Sterling observed in 1959 that when we read a paper that has been selected for publication, "the risk [e.g., the *p*-value] stated by the author cannot be accepted at face value once the author's conclusions appear in print."

"Why Most Published Research Findings Are False." To explain Ioannidis's argument, we need to make a slight digression and explore a statistical trap known as the *base rate fallacy*.

Imagine that you are the physician treating a young man who is concerned that he may have contracted Lyme disease while on a fishing trip in Maine. He has felt poorly ever since but has not had the characteristic circular rash associated with Lyme. Mostly to relieve his concern, you agree to test his blood for antibodies against the bacteria that cause the disease.

To his dismay and yours, the test comes back positive. The test itself is reasonably accurate, but not exceptionally so. It shows a false positive about 5 percent of the time. What are the chances that your patient has Lyme disease?

Many people, including many physicians, would expect the answer to be somewhere around 95 percent. This is incorrect. Ninety-five percent is the chance that someone who doesn't have Lyme disease would test negative. You want to know the chance that someone who tests positive has Lyme disease. It turns out that this is a low probability, because Lyme disease is quite rare. In areas where it is endemic, only about one person out of one thousand is infected. So imagine testing 10,000 people. You'd expect to have about 10 true positives, and about $0.05 \times 10{,}000 = 500$ false positives. Fewer than 1 in 50 of those who test positive are actually infected. Thus you expect your patient would have less than a 2 percent chance of having the disease, *even after testing positive*.

This confusion—thinking that there is an about 95 percent chance that the patient is infected when actually the chances are less than 2 percent—should be a familiar mistake. This is our old friend, the prosecutor's fallacy, dressed up in new clothes. We sometimes call it the *base rate fallacy* because it involves ignoring the base rate of the disease in a population when interpreting the results of a test.

Recall our table from the prosecutor's fallacy:

	MATCH	NO MATCH
Guilty	1	0
Innocent	5	50,000,000

The analogous table for the base rate fallacy is as follows:

	POSITIVE TEST	NEGATIVE TEST
Infected	10	0
Uninfected	500	10,000

In each case, the mistake is to compare probabilities along the bottom row instead of down the left column.

The base rate fallacy is less of an issue if you are testing for a condition that is very common. Suppose you are treating a young Caucasian woman from the upper Midwest for stomach problems, and you decide to test for the presence of *Helicobacter pylori,* a stomach pathogen associated with peptic ulcers. As with the antibody test for Lyme disease, about 5 percent of uninfected people test positive when they use the urea breath test. If your patient tests positive, what are the chances that she is carrying *Helicobacter*? Is it also 1 in 100? No, it's far greater, because *Helicobacter* is a common pathogen. In the United States, about 20 percent of Caucasians carry *Helicobacter*. So imagine testing 10,000 people for this pathogen. You would see about 2,000 true positives, and about 5 percent of the remaining 8,000 people, i.e., about 400 people, would get false positive results. Thus among US Caucasians, roughly 5 in 6 of those who test positive for *Helicobacter* are actually carrying it.

With that out of the way, let's come back to Ioannidis. In his paper "Why Most Published Research Findings Are False," Ioannidis draws the analogy between scientific studies and the interpretation of medical tests. He assumes that because of publication bias, most negative findings go unpublished and the literature comprises mostly positive results. If scientists are testing improbable hypotheses, the majority of positive results will be false positives, just as the majority of tests for Lyme disease, absent other risk factors, will be false positives.

That's it. That's the whole argument. We can't really argue with Ioannidis's mathematics. His conclusions are correct, given his model. He also has some empirical support from the papers we discussed earlier, those demonstrating that many experiments published in good journals cannot be reproduced. If many of the positive findings for

those experiments were false positives, this is exactly what we would expect.

What we can argue about are Ioannidis's assumptions. For most published findings to be false, scientific experiments would have to be like rare diseases: highly unlikely to generate a true positive result. Science doesn't work like that, because scientists get to *choose* what hypotheses they want to test. We've seen that scientists are finely attuned to the reward structure of their professional world, that rewards accrue mostly for publishing interesting work, and that it is difficult to publish negative results. So we would expect scientists to test hypotheses that, while undecided, seem reasonably likely to be true. This moves us toward the domain of the *Helicobacter pylori* example, where the majority of positive results are true positives. Ioannidis is overly pessimistic because he makes unrealistic assumptions about the kinds of hypotheses that researchers decide to test.

Of course, this is all theoretical speculation. If we want to actually measure how big of a problem publication bias is, we need to know (1) what fraction of tested hypotheses are actually correct, and (2) what fraction of negative results get published. If both fractions are high, we've got little to worry about. If both are very low, we've got problems.

We've argued that scientists will tend to test hypotheses with a decent chance of being correct. The chance may be 10 percent or 50 percent or 75 percent, but it is unlikely to be 1 percent or 0.1 percent. What about publishing negative results? How often does that happen? Across the sciences at large, about 15 percent of the results published are negative. Within biomedicine, 10 percent. Social psychology, a mere 5 percent. The problem is that we don't know from these data whether psychologists are less likely to publish negative results, or whether they are choosing experiments that are more likely to generate positive results. The fraction of published results that are negative is not what we really want to know. We want to know the fraction of negative results that are published.

But how can we get at that? We would have to somehow know about all of the unpublished results of experiments—and these tend to be buried in file drawers. Erick Turner at the US Food and Drug Administration (FDA) found an ingenious way to get around this problem. In the United States, whenever a team of investigators wants

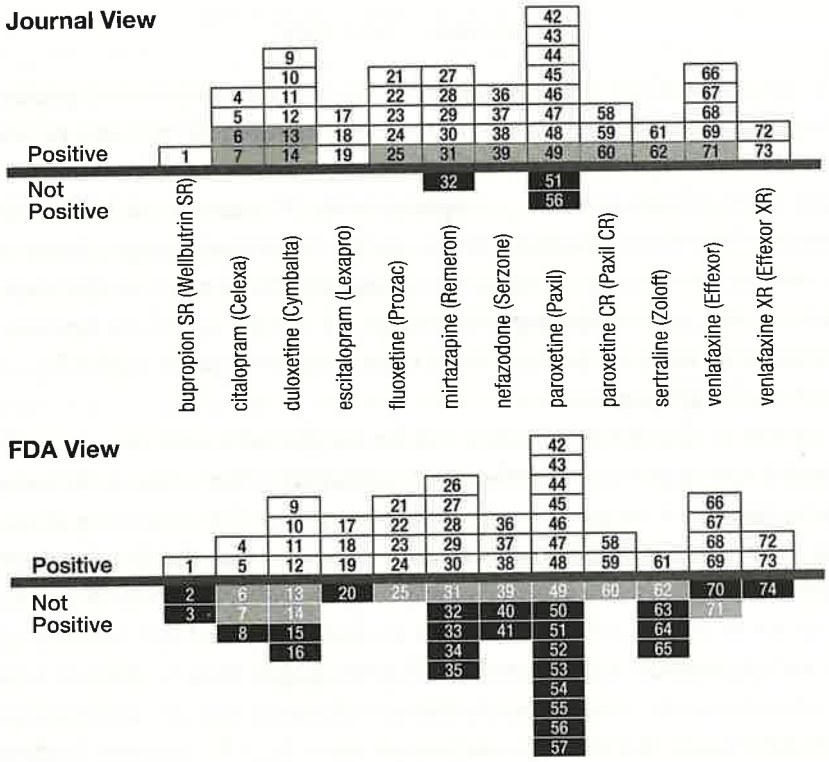
to run a clinical trial—an experiment using human subjects to test outcomes of medical treatments—they are required by law to register this trial with the FDA. This involves filing paperwork that explains what the trial is designed to test, how the trial will be conducted, and how the outcomes will be measured. Once the trial is completed, the team is also required to report the results to the FDA. However, they are not required to publish the results in a scientific journal.

This system provided Turner and his colleagues with a way of tallying both the published and unpublished trials in one particular area of research. Turner compiled a list of 74 clinical trials aimed at assessing the effectiveness of 12 different antidepressant medications. Results from 51 of these trials were published, 48 with positive results (the drug is effective) and 3 with negative results. Looking at the published literature, a researcher would conclude that these antidepressant drugs tend to work. But with access to the experiments as initially registered, the FDA sees a very different picture. They see 74 trials of which 38 yield positive results, 12 yield questionable results, and 24 yield negative results. From those numbers one would reach a more pessimistic conclusion—that some antidepressants seem to work somewhat under some circumstances.

What happened? How did clinical trials with a 51 percent success rate end up being reported as successful in 94 percent of the published papers? For one thing, almost all of the positive results were published, whereas fewer than half of the questionable or negative results were published. Yet more problematically, of the 14 questionable or negative results that were published, 11 were recast as positive findings.*

Turner illustrates these results with a graph that looks something like the upcoming graph. The dark rectangles represent negative results, the open ones positive results. Lightly shaded rectangles represent studies that originally returned questionable or negative results, but were published as positive findings.

* We caution against overly extrapolating from Turner's study. Clinical trials differ considerably from many other types of experiments, and they may be more—or less—subject to publication bias. On one hand, commercial interests may be involved to an unusual degree and may have some role in deterring the publication of negative results or encouraging that these results are reframed as positive. On the other hand, clinical research tends to be expensive, time-consuming, and generally involves a large number of investigators—so there may be strong incentives to publish results, whatever they may be.



Just as a sailor sees only the portion of the iceberg above the water's surface, a researcher reads only the positive results in the scientific literature. This makes it difficult to tell how many negative results are lying beneath the waterline. They are rarely published and, if they do appear, are often spun to appear in the guise of positive results. If there isn't much below the waterline, we have strong support for whatever is being tested. But if there is considerable mass lurking below the surface, the impression one gathers from viewing only the surface can be highly misleading.

Fortunately, there are ways to estimate the size of the submerged piece of the iceberg. One of the most powerful approaches involves meta-analysis: looking at multiple studies simultaneously. Doing so, we may be able to see when the published literature is likely to be representative of the set of all experiments conducted, and when, instead, the published literature reflects problematic practices such as *p*-hacking or publication bias. Figuring out how best to do this has become a hot area in statistics research.

CLICKBAIT SCIENCE

Members of the general public are sometimes skeptical of professional scientists and their opinions. For example, 88 percent of the members of the AAAS—a professional association for scientists—feel that GMO foods are safe, compared with 37 percent of US adults overall. Of AAAS members, 87 percent believe that climate change is mostly due to human activity, compared with 50 percent of US adults overall. An overwhelming 98 percent of AAAS members are convinced that humans have evolved over time, compared with 65 percent of the US population.

Some of this public distrust can be attributed to deliberate, well-funded campaigns to manufacture uncertainty. The tobacco industry spent decades trying to sow doubts about the evidence linking smoking with cancer. The oil industry continues to attack the link between carbon emissions and climate change. Religiously motivated groups attempt to create public mistrust of evolutionary biology, in favor of so-called creation science and intelligent design theory. This is how cover-ups work at scale. People like to believe in massive conspiracies that keep some story—alien corpses on ice in Area 51, a moon landing staged in a Hollywood studio, CIA involvement in 9/11—undercover. But no one can keep a secret that big. Rather, the really big cover-ups take place out in the open, as climate denial does. The smoking gun is there for everyone to see; the cover-up is providing people with alternative reasons to believe it might be smoking.

Still, a fair share of the blame lies squarely on the shoulders of scientists and science reporters. For one thing, science reporting amplifies the publication bias problem that we see in the scientific literature. Newspapers and other sources of science news eagerly report potential breakthroughs, but many of these studies fail to pan out. This isn't necessarily a problem. Science works like that—only a few of the exciting new leads that we uncover withstand subsequent experiments. The problem arises because news sources often fail to clearly indicate the preliminary nature of the findings that they report, and worse yet, they seldom report when the studies covered previously fail to pan out.

Estelle Dumas-Mallet and her colleagues attempted to estimate the magnitude of this bias in news reporting. They looked at over five

thousand papers about disease risk, of which 156 were reported in the popular press. All of the papers that received press attention reported positive results, in the sense that they suggested links between disease and genetic or behavioral risk factors. Thirty-five of these papers (about which 234 news articles had been written) were disconfirmed by subsequent research. Yet when that happened, only four news articles were published pointing out that the original stories were incorrect. Negative results simply aren't exciting.

Think about the endless parade of newspaper stories about the health effects of red wine. One week, a single daily glass of wine increases heart disease risk, and the next week that same glass of wine decreases this risk. So long as a research study on this topic finds a relationship, it seems newsworthy. But when a follow-up study finds that the relationship was spurious, the popular press isn't interested. No wonder members of the public feel jerked around by the scientists who can't decide whether wine is good or bad, and no wonder they quickly become cynical about the entire enterprise.

Popular science writing often allows, if not actively encourages, a fundamental misunderstanding about what the results of a single study mean for science. In the news media and even in textbooks, scientific activity is often portrayed as a *collecting process,* and a scientific paper as the report of what has been collected. By this view, scientists search out the facts that nature has hidden away; each uncovered fact is published in a scientific paper like a stamp laid in a collector's album; textbooks are essentially collections of such facts.

But science doesn't work like this. The results of an experiment are not a definitive fact about nature. The results of experiments involve chance, as well as a web of assumptions about how such results are properly assessed. The interpretation of experimental results involves a yet more complex network of models and assumptions about how the world works. Rather than representing a definitive fact about nature, each experiment or collection of observations merely represents an argument in favor of some particular hypothesis. We judge the truth of hypotheses by weighing the evidence offered from multiple papers, each of which comes at the issue from a different perspective.

This distinction has important implications for how we ought to interpret media reports about scientific studies. Suppose a new study reports cardiac benefits associated with moderate consumption of red

wine. This is not a new fact to be engraved in our canon of knowledge. Rather, it represents yet one additional contribution to an oft-studied question, marginally shifting our beliefs in the direction of the hypothesis "red wine has cardiac benefits." A single study doesn't tell you much about what the world is like. It has little value at all unless you know the rest of the literature and have a sense about how to integrate these findings with previous ones.

Practicing scientists understand this. Researchers don't make up their minds about a complex question based on an individual research report, but rather, weigh the evidence across multiple studies and strive to understand why multiple studies often produce seemingly inconsistent results. But the popular press almost never reports the story in this way. It's boring.

Even when popular science articles discuss multiple studies, they do not always do so in a representative fashion. The epithet "cafeteria Catholic" suggests a worshipper who picks and chooses from the tenets of the faith, ignoring those that are uncomfortable or unpleasant. Science writers sometimes engage in an analogous practice that we call cafeteria science. They pick and choose from a broad menu of studies, to extract a subset that tells a consistent and compelling story.

Scientists are not altogether innocent of this practice either. In 1980, two researchers published a brief hundred-word note in *The New England Journal of Medicine,* reporting a low rate of addiction to narcotic painkillers based on medical records of hospitalized patients. Following the release of the opioid painkiller OxyContin, this paper was widely cited in the medical literature as evidence that opioids rarely cause addiction—a massive overstatement of its findings. Some scholars go so far as to ascribe a portion of the ongoing opioid crisis to the uncritical use of the paper to minimize concerns about addiction. In 2017, the editors of the *New England Journal* took the highly unusual step of issuing a warning that now appears atop the article. While the warning does not question the paper's findings, it cautions: "For reasons of public health, readers should be aware that this letter has been 'heavily and uncritically cited' as evidence that addiction is rare with opioid therapy."

Worse still, there is a powerful selection bias in what scientific studies we end up hearing about in the popular press and on social media. The research studies that are reported in the popular press are not a

random sample of those conducted, or even of those published. The most surprising studies are the ones that make the most exciting articles. If we fail to take this into account, and ignore all of the less surprising findings from related studies, we can end up with an unrealistic picture of how scientific knowledge is developing.

When science writers cast scientific studies in popular language, they sometimes omit important caveats, offer stronger advice than is justified, present correlations as causal relationships, or extrapolate from findings in lab animals to suggest immediate application in humans. Very often these errors are not introduced by the reporters; they are already present in the press releases issued by universities and scientific journals. Charged with producing more and more material in less and less time, science writers often rely heavily on these releases as source material. Less reputable news outlets and blogs often go one step further, republishing press releases in the guise of original reporting.

In the spring of 2015, astronaut Scott Kelly boarded a rocket for a trip to the International Space Station. Three months later he returned to Earth a changed man. A NASA press release reported that

> Researchers now know that 93% of Scott's genes returned to normal after landing. However, the remaining 7% point to possible longer term changes in genes related to his immune system, DNA repair, bone formation networks, hypoxia, and hypercapnia.

Major news networks interpreted this as meaning that Kelly's genome changed by 7 percent. "Astronaut's DNA No Longer Matches That of His Identical Twin, NASA Finds," announced CNN. "NASA Twins Study Confirms Astronaut's DNA Actually Changed in Space," reported *Newsweek*. "After Year in Space, Astronaut Scott Kelly No Longer Has Same DNA as Identical Twin," declared the *Today* show. On Twitter, Scott Kelly joked, "What? My DNA changed by 7%! Who knew? I just learned about it in this article. This could be good news! I no longer have to call @ShuttleCDRKelly my identical twin brother anymore."

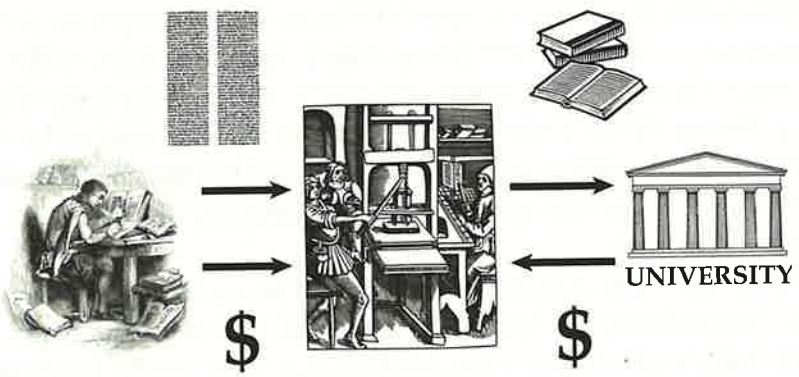
But wait—this doesn't even pass a very basic plausibility check. A chimpanzee's genome is only 2 percent different from a human's. In fact, Kelly's genes themselves didn't change. The way his genes were

expressed did. In other words, there were persistent changes in the rates at which about 7 percent of his (unchanged) genes were being translated into proteins. This is unsurprising. Changes in environmental conditions commonly drive changes in gene expression. Geneticists and other biologists responded immediately, trying to correct the misinformation on social media and in news interviews. NASA issued a correction to its press release. Some news outlets wrote entire stories about the debacle. But most people had moved on and the corrections received only a fraction of the attention given to the original story. Jonathan Swift had it right: "Falsehood flies, and truth comes limping after it."

THE MARKET FOR BULLSHIT SCIENCE

In a typical economic market for consumer goods, money flows in one direction; goods flow in the other. A steel company provides raw materials to an auto manufacturer, which assembles cars and sells them to consumers. The consumers pay the manufacturer, which in turn pays the steel company. The market for scholarly journals is different. A scholar pours her heart and soul into writing a scholarly paper. She provides her work to an academic publisher, which collects a number of such articles and bundles them together as a journal issue. The publisher charges libraries to subscribe, but doesn't pass any of the revenue on to the authors.

Some journals even require that the author pay to have her work published. Why would anyone pay? Recall what motivates academic scholars. As we discussed earlier in the chapter, scholars are rewarded



for the reputations they accrue, and publishing is how reputations are made.

To be clear, there's nothing inherently wrong with the idea of charging authors to publish their work. Many scientists enthusiastically support *open access* publishing, in which journal publishers take money from the authors and dispense with subscription fees entirely. This has a number of advantages. The papers published in this way are free for anyone in the world to read at any time. With open access, scientists at institutions in the developing world do not have to struggle to access the papers they need; medical patients and their families can view the latest papers about the relevant diseases; the public can read about the research results that they have already funded through government-sponsored research grants. And increased readership is a good thing for the authors.

That's the full half of the half-full glass. There's an empty half as well. Open access publishing creates a niche for low-quality journals that do not impose adequate standards. When a journal charges for subscriptions, its editors have strong incentives to publish high-quality articles that have been thoroughly vetted through peer review. Journals filled with low-quality articles sell few subscriptions. When a journal charges academics who are eager to publish for the service of publishing their work, multiple business models come into play. Most open access journals seek to build their reputations by publishing only high-quality work. But for some, money is money—they'll publish anything provided that the check clears.

Here is Goodhart's law's again: When a measure becomes a target, it ceases to be a good measure. This has happened to a substantive swath of the scientific literature. When scientists started judging one another by the number of papers published, a market sprung up for journals willing to publish low-quality work. At the bottom of that particular barrel are journals produced by so-called predatory publishers. These parasites on the scientific publishing system provide little to nothing by way of rigorous peer review. Today they are sucking tens of millions of dollars from the academic system and polluting the literature with millions of unreliable articles.

It could be useful to have a central authority validating journals as "scientific," but none exists. In the age of print, the cost of publishing a physical journal and the need to sell subscriptions to libraries created

barriers that prevented scammers from entering the publishing game. Now that journals are viewed online, articles are delivered electronically, and publishers can bill authors rather than libraries, those barriers are gone. Today, a rudimentary understanding of Web design and a willingness to defraud people is all it takes to become a predatory publisher.

While the primary customers for predatory journals may be marginal academics looking to boost their résumés with additional publications, other parties take advantage of these journals' services as well. Climate skeptics, antivaxxers, creationists, and HIV denialists publish their work in predatory journals and then point to these papers as "peer-reviewed" science supporting their fringe beliefs. Con artists publish fabricated data purporting to show the power of their snake-oil diet supplements or disease cures. Unscrupulous political operatives publish unfounded claims about their political opponents in an effort to rally support. Conspiracy theorists document the collaborations between the CIA, the Illuminati, and the aliens that built the great pyramids. Crackpots can publish grand theories of quantum gravity and the universal consciousness permeating all of space-time, or incantations written in lost languages that will summon a demonic archon to oversee the destruction of the world.

To make money, predatory journals need to attract submissions. Here they turn to sending spam email. If you publish an academic paper or two, your email in-box will start to fill up with messages from predatory journals, each saying something along the lines of: "With great interest we read your paper [paper title] in the journal [journal title]. We congratulate you on this excellent work. If you have additional research in this area, we would urge you to consider our journal for your next publication. . . ." Of course, no one from the journal has actually read your paper; its title and your email address have simply been scraped from the Internet—but if you are new to academic publishing, you wouldn't necessarily know this. Sometimes these communications list the publication fees, but many predatory journal publishers hide these fees until after a paper has been accepted.

In a typical week we each get a few dozen of these emails. Some of our colleagues have found amusing ways to fight back. Scientific editor John McCool was invited to publish in the *Urology & Nephrology*

Open Access Journal. McCool has no medical background, but he is a devoted *Seinfeld* fan, and he recalled an episode that might be of interest to the journal in question. In season three's episode "The Parking Garage," Jerry Seinfeld forgets where he parked his car, and after hours of searching without a bathroom break he is arrested for public urination. At the police station, he tries to talk his way out of his predicament. "Why would I [urinate in public] unless I was in mortal danger? I know it's against the law," Jerry tells the officer who is booking him into jail. He then answers his own question: "Because I could get uromycitisis poisoning and die! That's why." Of course there is no such thing as uromycitisis. Seinfeld just made up a scary-sounding word off the top of his head.

McCool's paper, "Uromycitisis Poisoning Results in Lower Urinary Tract Infection and Acute Renal Failure: Case Report," lays out a complete summary of the episode's plot in the guise of a case report. It begins: "A 37-year-old white male was in a large suburban mall parking garage and was unable to locate his car. After more than an hour of walking up and down flights of stairs and through row after row of cars, searching fruitlessly for his own car, he felt a powerful urge to urinate. With no restroom available in the garage, and knowing that he suffers from uromycitisis, he feared that if he did not urinate immediately he would develop uromycitisis poisoning." The peer reviewers, if such existed at all, must not be *Seinfeld* fans (or critical thinkers). The ridiculous case report was accepted for publication in a matter of days, in hope that McCool would pay the $799 author fee. He never did.

So how do you tell whether a scientific article is legitimate? The first thing to recognize is that *any scientific paper can be wrong*. That's the nature of science; nothing is above questioning. No matter where a paper is published, no matter who wrote it, no matter how well supported its arguments may be, any paper can be wrong. Every hypothesis, every set of data, every claim, and every conclusion is subject to reexamination in light of future evidence. Linus Pauling was a brilliant scientist who remains the only person to have received two unshared Nobel Prizes, the prize in chemistry and the peace prize—but he also published papers and books that turned out to be completely wrong, from his proposed triple-helix structure of DNA to his views about the benefits of high doses of vitamin C. *Nature* and *Science* are

the two most prestigious journals in the basic sciences, but they too have published some howlers. In 1969, *Science* published a paper about a nonexistent polymer of water known as polywater, which contributed to fears among defense researchers of a "polywater gap" between the US and Russia. *Nature* published an erroneous paper in 1988 purporting to show that homeopathy can be effective.

The second thing to understand is the role of peer review. While it's an important part of the scientific process, *peer review does not guarantee that published papers are correct*. Peer reviewers carefully read a paper to make sure its methods are reasonable and its reasoning logical. They make sure a paper accurately represents what it adds to the literature, and that its conclusions follow from its results. They suggest ways to improve a paper, and sometimes recommend additional experiments. But peer reviewers can make mistakes and, more important, peer reviewers cannot possibly check every aspect of the work. Peer reviewers don't redo the laboratory experiments, repeat the field observations, rewrite the computer code, rederive all of the mathematics, or even delve too deeply into the data in most cases. Though helpful, peer review cannot catch every innocent mistake, let alone uncover well-concealed acts of scientific misconduct.

As a result, there is no surefire way for you, as the reader, to know, beyond a shadow of a doubt, that any particular scientific paper is correct. Usually the best you can hope to do is to determine that a paper is *legitimate*. By legitimate, we mean that a paper is (1) written in good faith, (2) carried out using appropriate methodologies, and (3) taken seriously by the relevant scientific community.

A quick way to evaluate the legitimacy of a published paper is to find out about the journal in which it is published. A number of websites purport to rank journal quality or prestige, typically based on citations.* Highly cited journals are thought to be better than their

* Clarivate's journal impact factor is the most commonly used of these metrics. Journal impact factor measures the ratio of citations received over a two-year window to the number of "citable" articles published in that same window. Unfortunately, impact factor scores are available only in bulk via a subscription service known as the Clarivate *Journal Citation Reports* (JCR). While what constitutes an impressive impact factor varies from one field to another, it is a reasonable rule of thumb to consider that any journal listed in the JCR is at least somewhat reputable, any journal with an impact factor of at least 1 is decent, and any journal with an impact factor of at least 10 is outstanding. Several free alternatives to journal impact factor are also available. We the authors provide a set of journal indicators at http://www.eigenfactor.org. These metrics cover the same set of journals included in the JCR. The large commercial publisher Elsevier provides an alternative set of

seldom-cited competitors. Knowing the publisher is helpful as well. Major publishers and reputable scientific societies usually ensure that the articles in their journals meet basic standards of quality.

Another issue to be on the lookout for is whether the claims in a paper are commensurate with the venue in which it is published. As we mentioned, journals occupy different positions in a hierarchy of prestige. All else equal, papers published in top journals will represent the largest advances and have the highest credibility. Less interesting or less credible findings will be relegated to less prestigious outlets. Be wary of extraordinary claims appearing in lower-tier venues. You can think of this as the scientist's version of "if you're so smart, why aren't you rich?"

Thus if a paper called "Some Weights of Riparian Frogs" lists the weights of some frogs in the little-known *Tasmanian Journal of Australian Herpetology,* there is relatively little cause for concern. A table of frogs' weights, while possibly useful to some specialists in the area, is far from a scientific breakthrough and is well suited for the journal in question. If, however, a paper called "Evidence That Neanderthals Went Extinct during the Hundred Years' War" appears in the equally low-profile *Journal of Westphalian Historical Geography,* this is serious cause for concern. That finding, if true, would revolutionize our understanding of hominin history, and shake our notion of what it means to be human. Such a finding, if true, would appear in a high-profile journal.

While the examples above are hypothetical, real examples abound. For example, in 2012 television personality Dr. Mehmet Oz used his show to promote a research paper purporting to demonstrate that green coffee extract has near miraculous properties as a weight-loss supplement. Despite this remarkable claim, the paper did not appear in a top medical journal such as *JAMA, The Lancet,* or *NEJM.* Rather, it appeared in a little-known journal titled *Diabetes, Metabolic Syndrome and Obesity: Targets and Therapy* from a marginal scientific publisher called Dove Press. The journal is not even listed in some of the major scientific indices. This should set off alarm bells for any reader—and

---

metrics based on their Scopus database. Scopus covers a larger set of journals than the JCR, but we have concerns about the possible conflicts of interest that arise when a major journal publisher begins ranking its own journals against those of its competitors. Google Scholar provides journal rankings of its own as well.

indeed a look at the paper reveals that the results are based on a clinical trial with an absurdly small sample size of sixteen, too small a sample to justify the strong claims that the paper makes. The small sample size and inglorious venue turned out to be only part of the story. The paper was subsequently retracted because its data could not be validated.

While retractions are uncommon, it can be a good idea to check for a retraction or correction before staking too much on the results of a scientific paper. The easiest way to do this is simply to check the paper on the publisher's website or—if it is a paper in biology or medicine—in the PubMed database.*

WHY SCIENCE WORKS

We've run through a fairly long litany of problems that plague contemporary science. It would be easy to feel overwhelmed, and start to wonder whether science works at all. Fortunately, there are a number of reasons why the institutions of science are effective despite everything we've discussed in this chapter.

We've already discussed how most of the hypotheses that get tested are *a priori* reasonably likely to be correct. Scientists tend to test hypotheses that have a decent chance of being true rather than wasting their time on long shots. When such is the case, most positive findings will be true positives, not false positives.

Science is a cumulative process. Even though experiments are not often directly replicated, science proceeds when researchers build upon previous results. If a result is false, people cannot build on it effectively. Their attempts will fail, they'll go back and reassess the original findings, and in this way the truth will come out. Suppose that while studying fruit flies, I discover a biochemical pathway that makes it easy to edit an organism's genes. Other researchers may not try to directly replicate my experiment, but people working with mice or nematodes or other model organisms will want to see if the same mechanism works in their systems. Or perhaps people interested in technological applications may try to work out better methods of using this pathway in a directed fashion. If I were wrong in the first

* PubMed is freely available online at https://www.ncbi.nlm.nih.gov/pubmed/.

place, none of this will work and my mistakes will be corrected. If I am right, the successes of these subsequent researchers will confirm my discovery.

Experiments often test not only whether there is an effect but the direction of the effect. Suppose several research teams test whether a new antidepressant also helps with memory loss among seniors. If this drug has no real effect on memory, there are two different kinds of false positive results we could find: The drug could improve or decrease performance on memory tests. Because in either case the treatment group taking the drug differs from the control group taking a placebo, either result could be published. If a treatment has mixed effects in the published literature, some beneficial and some harmful, we have reason to suspect that we are looking at statistical noise instead of a true signal.

As science moves toward accepting some claim as a fact, experiments that contradict that claim become noteworthy, are treated almost as positive results, and become far easier to publish. For example, physicists have typically assumed that the speed of light in a vacuum is a fundamental constant of physics and is the same everywhere and at all times in the universe. If a few decades ago someone tried to measure differences in the speed of light across time and space and found no difference, this would have been a negative result and difficult to publish in a high-profile journal. But recently several lines of evidence have accumulated suggesting that the speed of light in a vacuum may vary. Now that these claims are being discussed in the physics community, careful experiments showing no difference would be of substantial interest.

Finally, irrespective of the problems we have discussed in this chapter, science just plain works. As we stated at the start of this chapter, science allows us to understand the nature of the physical world at scales far beyond what our senses evolved to detect and our minds evolved to comprehend. Equipped with this understanding, we have been able to create technologies that would seem magical to those only a few generations prior. Empirically, science is successful. Individual papers may be wrong and individual studies misreported in the popular press, but the institution as a whole is strong. We should keep this in perspective when we compare science to much of the other human knowledge—and human bullshit—that is out there.