FIRST THINGS

SCIENTIFIC REGRESS

by

William A. Wilson

May 2016

The problem with science is that so much of it simply isn't. Last summer, the Open Science Collaboration announced that it had tried to replicate one hundred published psychology experiments sampled from three of the most prestigious journals in the field. Scientific claims rest on the idea that experiments repeated under nearly identical conditions ought to yield approximately the same results, but until very recently, very few had bothered to check in a systematic way whether this was actually the case. The OSC was the biggest attempt yet to check a field's results, and the most shocking. In many cases, they had used original experimental materials, and sometimes even performed the experiments under the guidance of the original researchers. Of the studies that had originally reported positive results, an astonishing 65 percent failed to show statistical significance on replication, and many of the remainder showed greatly reduced effect sizes.

Their findings made the news, and quickly became a club with which to bash the social sciences. But the problem isn't just with psychology. There's an unspoken rule in the pharmaceutical industry that half of all academic biomedical research will ultimately prove false, and in 2011 a group of researchers at Bayer decided to test it. Looking at sixty-seven recent drug discovery projects based on preclinical cancer biology research, they found that in more than 75 percent of cases the published data did not match up with their in-house attempts to replicate. These were not studies published in fly-by-night oncology journals, but blockbuster research featured in *Science*, *Nature*, *Cell*, and the like. The Bayer researchers were drowning in bad studies, and it was to this, in part, that they attributed the mysteriously declining yields of drug pipelines. Perhaps so many of these new drugs fail to have an effect because the basic

research on which their development was based isn't valid.

When a study fails to replicate, there are two possible interpretations. The first is that, unbeknownst to the investigators, there was a real difference in experimental setup between the original investigation and the failed replication. These are colloquially referred to as "wallpaper effects," the joke being that the experiment was affected by the color of the wallpaper in the room. This is the happiest possible explanation for failure to reproduce: It means that both experiments have revealed facts about the universe, and we now have the opportunity to learn what the difference was between them and to incorporate a new and subtler distinction into our theories.

The other interpretation is that the original finding was false. Unfortunately, an ingenious statistical argument shows that this second interpretation is far more likely. First articulated by John Ioannidis, a professor at Stanford University's School of Medicine, this argument proceeds by a simple application of Bayesian statistics. Suppose that there are a hundred and one stones in a certain field. One of them has a diamond inside it, and, luckily, you have a diamond-detecting device that advertises 99 percent accuracy. After an hour or so of moving the device around, examining each stone in turn, suddenly alarms flash and sirens wail while the device is pointed at a promising-looking stone. What is the probability that the stone contains a diamond?

Most would say that if the device advertises 99 percent accuracy, then there is a 99 percent chance that the device is correctly discerning a diamond, and a 1 percent chance that it has given a false positive reading. But consider: Of the one hundred and one stones in the field, only one is truly a diamond. Granted, our machine has a very high probability of correctly declaring it to be a diamond. But there are many more diamond-free stones, and while the machine only has a 1 percent chance of falsely declaring each of them to be a diamond, there *are* a hundred of them. So if we were to wave the detector over every stone in the field, it would, on average, sound twice—once for the real diamond, and once when a false reading was triggered by a stone. If we know only that the alarm has sounded, these two possibilities are roughly equally probable, giving us an approximately 50 percent chance that the stone really contains a diamond.

This is a simplified version of the argument that Ioannidis applies to the process of science itself. The stones in the field are the set of all possible testable hypotheses, the diamond is a hypothesized connection or effect that happens to be true, and the diamond-detecting device is the scientific method. A tremendous amount depends on the proportion of possible hypotheses which turn out to be true, and on the accuracy with which an experiment can discern truth from falsehood. Ioannidis shows that for a wide variety of scientific settings and fields, the values of these two parameters are not at all favorable.

For instance, consider a team of molecular biologists investigating whether a mutation in one of the countless thousands of human genes is linked to an increased risk of Alzheimer's. The probability of a randomly selected mutation in a randomly selected gene having precisely that effect is quite low, so just as with the stones in the field, a positive finding is more likely than not to be spurious—unless the experiment is unbelievably successful at sorting the wheat from the chaff. Indeed, Ioannidis finds that in many cases, approaching even 50 percent true positives requires unimaginable accuracy. Hence the eye-catching title of his paper: "Why Most Published Research Findings Are False."

What about accuracy? Here, too, the news is not good. First, it is a de facto standard in many fields to use one in twenty as an acceptable cutoff for the rate of false positives. To the naive ear, that may sound promising: Surely it means that just 5 percent of scientific studies report a false positive? But this is precisely the same mistake as thinking that a stone has a 99 percent chance of containing a diamond just because the detector has sounded. What it really means is that for each of the countless false hypotheses that are contemplated by researchers, we accept a 5 percent chance that it will be falsely counted as true—a decision with a considerably more deleterious effect on the proportion of correct studies.

Paradoxically, the situation is actually made worse by the fact that a promising connection is often studied by several independent teams. To see why, suppose that three groups of researchers are studying a phenomenon, and when all the data are analyzed, one group announces that it has discovered a connection, but the other two find nothing of note. Assuming that all the tests involved have a high

statistical power, the lone positive finding is almost certainly the spurious one. However, when it comes time to report these findings, what happens? The teams that found a negative result may not even bother to write up their non-discovery. After all, a report that a fanciful connection probably isn't true is not the stuff of which scientific prizes, grant money, and tenure decisions are made.

And even if they did write it up, it probably wouldn't be accepted for publication. Journals are in competition with one another for attention and "impact factor," and are always more eager to report a new, exciting finding than a killjoy failure to find an association. In fact, both of these effects can be quantified. Since the majority of all investigated hypotheses are false, if positive and negative evidence were written up and accepted for publication in equal proportions, then the majority of articles in scientific journals should report no findings. When tallies are actually made, though, the precise opposite turns out to be true: Nearly every published scientific article reports the presence of an association. There must be massive bias at work.

Ioannidis's argument would be potent even if all scientists were angels motivated by the best of intentions, but when the human element is considered, the picture becomes truly dismal. Scientists have long been aware of something euphemistically called the "experimenter effect": the curious fact that when a phenomenon is investigated by a researcher who happens to believe in the phenomenon, it is far more likely to be detected. Much of the effect can likely be explained by researchers unconsciously giving hints or suggestions to their human or animal subjects, perhaps in something as subtle as body language or tone of voice. Even those with the best of intentions have been caught fudging measurements, or making small errors in rounding or in statistical analysis that happen to give a more favorable result. Very often, this is just the result of an honest statistical error that leads to a desirable outcome, and therefore it isn't checked as deliberately as it might have been had it pointed in the opposite direction.

But, and there is no putting it nicely, deliberate fraud is far more widespread than the scientific establishment is generally willing to admit. One way we know that there's a great deal of fraud occurring is that if you phrase your question the right way, scientists will confess to it. In a survey of two thousand

research psychologists conducted in 2011, over half of those surveyed admitted outright to selectively reporting those experiments which gave the result they were after. Then the investigators asked respondents anonymously to estimate how many of their fellow scientists had engaged in fraudulent behavior, and promised them that the more accurate their guesses, the larger a contribution would be made to the charity of their choice. Through several rounds of anonymous guessing, refined using the number of scientists who would admit their own fraud and other indirect measurements, the investigators concluded that around 10 percent of research psychologists have engaged in outright falsification of data, and more than half have engaged in less brazen but still fraudulent behavior such as reporting that a result was statistically significant when it was not, or deciding between two different data analysis techniques after looking at the results of each and choosing the more favorable.

Many forms of statistical falsification are devilishly difficult to catch, or close enough to a genuine judgment call to provide plausible deniability. Data analysis is very much an art, and one that affords even its most scrupulous practitioners a wide degree of latitude. Which of these two statistical tests, both applicable to this situation, should be used? Should a subpopulation of the research sample with some common criterion be picked out and reanalyzed as if it were the totality? Which of the hundreds of coincident factors measured should be controlled for, and how? The same freedom that empowers a statistician to pick a true signal out of the noise also enables a dishonest scientist to manufacture nearly any result he or she wishes. Cajoling statistical significance where in reality there is none, a practice commonly known as "p-hacking," is particularly easy to accomplish and difficult to detect on a case-by-case basis. And since the vast majority of studies still do not report their raw data along with their findings, there is often nothing to re-analyze and check even if there were volunteers with the time and inclination to do so.

One creative attempt to estimate how widespread such dishonesty really is involves comparisons between fields of varying "hardness." The author, Daniele Fanelli, theorized that the farther from physics one gets, the more freedom creeps into one's experimental methodology, and the fewer constraints there are on a

scientist's conscious and unconscious biases. If all scientists were constantly attempting to influence the results of their analyses, but had more opportunities to do so the "softer" the science, then we might expect that the social sciences have more papers that confirm a sought-after hypothesis than do the physical sciences, with medicine and biology somewhere in the middle. This is exactly what the study discovered: A paper in psychology or psychiatry is about five times as likely to report a positive result as one in astrophysics. This is not necessarily evidence that psychologists are all consciously or unconsciously manipulating their data—it could also be evidence of massive publication bias—but either way, the result is disturbing.

Speaking of physics, how do things go with this hardest of all hard sciences? Better than elsewhere, it would appear, and it's unsurprising that those who claim all is well in the world of science reach so reliably and so insistently for examples from physics, preferably of the most theoretical sort. Folk histories of physics combine borrowed mathematical luster and Whiggish triumphalism in a way that journalists seem powerless to resist. The outcomes of physics experiments and astronomical observations seem so matter-of-fact, so concretely and immediately connected to underlying reality, that they might let us gingerly sidestep all of these issues concerning motivated or sloppy analysis and interpretation. "*E pur si muove*," Galileo is said to have remarked, and one can almost hear in his sigh the hopes of a hundred science journalists for whom it would be all too convenient if Nature were always willing to tell us whose theory is more correct.

And yet the flight to physics rather gives the game away, since measured any way you like—volume of papers, number of working researchers, total amount of funding—deductive, theory-building physics in the mold of Newton and Lagrange, Maxwell and Einstein, is a tiny fraction of modern science as a whole. In fact, it also makes up a tiny fraction of modern physics. Far more common is the delicate and subtle art of scouring inconceivably vast volumes of noise with advanced software and mathematical tools in search of the faintest signal of some hypothesized but never before observed phenomenon, whether an astrophysical event or the decay of a subatomic particle. This sort of work is difficult and beautiful in its

own way, but it is not at all self-evident in the manner of a falling apple or an elliptical planetary orbit, and it is very sensitive to the same sorts of accidental contamination, deliberate fraud, and unconscious bias as the medical and social-scientific studies we have discussed. Two of the most vaunted physics results of the past few years—the announced discovery of both cosmic inflation and gravitational waves at the BICEP2 experiment in Antarctica, and the supposed discovery of superluminal neutrinos at the Swiss-Italian border—have now been retracted, with far less fanfare than when they were first published.

Many defenders of the scientific establishment will admit to this problem, then offer hymns to the self-correcting nature of the scientific method. Yes, the path is rocky, they say, but peer review, competition between researchers, and the comforting fact that there *is* an objective reality out there whose test every theory must withstand or fail, all conspire to mean that sloppiness, bad luck, and even fraud are exposed and swept away by the advances of the field.

So the dogma goes. But these claims are rarely treated like hypotheses to be tested. Partisans of the new scientism are fond of recounting the "Sokal hoax"—physicist Alan Sokal submitted a paper heavy on jargon but full of false and meaningless statements to the postmodern cultural studies journal *Social Text*, which accepted and published it without quibble—but are unlikely to mention a similar experiment conducted on reviewers of the prestigious *British Medical Journal*. The experimenters deliberately modified a paper to include eight different major errors in study design, methodology, data analysis, and interpretation of results, and not a single one of the 221 reviewers who participated caught all of the errors. On average, they caught fewer than two—and, unbelievably, these results held up even in the subset of reviewers who had been specifically warned that they were participating in a study and that there might be something a little odd in the paper that they were reviewing. In all, only 30 percent of reviewers recommended that the intentionally flawed paper be rejected.

If peer review is good at anything, it appears to be keeping unpopular ideas from being published. Consider the finding of another (yes, another) of these replicability studies, this time from a group of cancer researchers. In addition to reaching the now unsurprising conclusion that only a dismal 11 percent

of the preclinical cancer research they examined could be validated after the fact, the authors identified another horrifying pattern: The "bad" papers that failed to replicate were, on average, cited far more often than the papers that did! As the authors put it, "some non-reproducible preclinical papers had spawned an entire field, with hundreds of secondary publications that expanded on elements of the original observation, but did not actually seek to confirm or falsify its fundamental basis."

What they do not mention is that once an entire field has been created—with careers, funding, appointments, and prestige all premised upon an experimental result which was utterly false due either to fraud or to plain bad luck—pointing this fact out is not likely to be very popular. Peer review switches from merely useless to actively harmful. It may be ineffective at keeping papers with analytic or methodological flaws from being published, but it can be deadly effective at suppressing criticism of a dominant research paradigm. Even if a critic is able to get his work published, pointing out that the house you've built together is situated over a chasm will not endear him to his colleagues or, more importantly, to his mentors and patrons.

Older scientists contribute to the propagation of scientific fields in ways that go beyond educating and mentoring a new generation. In many fields, it's common for an established and respected researcher to serve as "senior author" on a bright young star's first few publications, lending his prestige and credibility to the result, and signaling to reviewers that he stands behind it. In the natural sciences and medicine, senior scientists are frequently the controllers of laboratory resources—which these days include not just scientific instruments, but dedicated staffs of grant proposal writers and regulatory compliance experts—without which a young scientist has no hope of accomplishing significant research. Older scientists control access to scientific prestige by serving on the editorial boards of major journals and on university tenure-review committees. Finally, the government bodies that award the vast majority of scientific funding are either staffed or advised by distinguished practitioners in the field.

All of which makes it rather more bothersome that older scientists are the most likely to be invested in the regnant research paradigm, whatever it is, even if it's based on an old experiment that has never successfully been replicated. The quantum physicist Max Planck famously quipped: "A new scientific truth does not triumph by convincing its opponents and making them see the light, but rather because its opponents eventually die, and a new generation grows up that is familiar with it." Planck may have been too optimistic. A recent paper from the National Bureau of Economic Research studied what happens to scientific subfields when star researchers die suddenly and at the peak of their abilities, and finds that while there is considerable evidence that young researchers are reluctant to challenge scientific superstars, a sudden and unexpected death does not significantly improve the situation, particularly when "key collaborators of the star are in a position to channel resources (such as editorial goodwill or funding) to insiders."

In the idealized Popperian view of scientific progress, new theories are proposed to explain new evidence that contradicts the predictions of old theories. The heretical philosopher of science Paul Feyerabend, on the other hand, claimed that new theories frequently *contradict* the best available evidence—at least at first. Often, the old observations were inaccurate or irrelevant, and it was the invention of a new theory that stimulated experimentalists to go hunting for new observational techniques to test it. But the success of this "unofficial" process depends on a blithe disregard for evidence while the vulnerable young theory weathers an initial storm of skepticism. Yet if Feyerabend is correct, and an unpopular new theory can ignore or reject experimental data long enough to get its footing, how much longer can an old and creaky theory, buttressed by the reputations and influence and political power of hundreds of established practitioners, continue to hang in the air even when the results upon which it is premised are exposed as false?

The hagiographies of science are full of paeans to the self-correcting, self-healing nature of the enterprise. But if raw results are so often false, the filtering mechanisms so ineffective, and the self-correcting mechanisms so compromised and slow, then science's approach to truth may not even be monotonic.

That is, past theories, now "refuted" by evidence and replaced with new approaches, may be closer to the truth than what we think now. Such regress has happened before: In the nineteenth century, the (correct) vitamin C deficiency theory of scurvy was replaced by the false belief that scurvy was caused by proximity to spoiled foods. Many ancient astronomers believed the heliocentric model of the solar system before it was supplanted by the geocentric theory of Ptolemy. The Whiggish view of scientific history is so dominant today that this possibility is spoken of only in hushed whispers, but ours is a world in which things once known can be lost and buried.

And even if self-correction does occur and theories move strictly along a lifecycle from less to more accurate, what if the unremitting flood of new, mostly false, results pours in faster? Too fast for the sclerotic, compromised truth-discerning mechanisms of science to operate? The result could be a growing body of true theories completely overwhelmed by an ever-larger thicket of baseless theories, such that the proportion of true scientific beliefs shrinks even while the absolute number of them continues to rise. Borges's Library of Babel contained every true book that could ever be written, but it was useless because it also contained every false book, and both true and false were lost within an ocean of nonsense.

Which brings us to the odd moment in which we live. At the same time as an ever more bloated scientific bureaucracy churns out masses of research results, the majority of which are likely outright false, scientists themselves are lauded as heroes and science is upheld as the only legitimate basis for policy-making. There's reason to believe that these phenomena are linked. When a formerly ascetic discipline suddenly attains a measure of influence, it is bound to be flooded by opportunists and charlatans, whether it's the National Academy of Science or the monastery of Cluny.

This comparison is not as outrageous as it seems: Like monasticism, science is an enterprise with a superhuman aim whose achievement is forever beyond the capacities of the flawed humans who aspire toward it. The best scientists know that they must practice a sort of mortification of the ego and cultivate a dispassion that allows them to report their findings, even when those findings might mean the dashing of hopes, the drying up of financial resources, and the loss of professional prestige. It should be no

surprise that even after outgrowing the monasteries, the practice of science has attracted souls driven to seek the truth regardless of personal cost and despite, for most of its history, a distinct lack of financial or status reward. Now, however, science and especially science bureaucracy is a *career*, and one amenable to social climbing. Careers attract careerists, in Feyerabend's words: "devoid of ideas, full of fear, intent on producing some paltry result so that they can add to the flood of inane papers that now constitutes 'scientific progress' in many areas."

If science was unprepared for the influx of careerists, it was even less prepared for the blossoming of the Cult of Science. The Cult is related to the phenomenon described as "scientism"; both have a tendency to treat the body of scientific knowledge as a holy book or an a-religious revelation that offers simple and decisive resolutions to deep questions. But it adds to this a pinch of glib frivolity and a dash of unembarrassed ignorance. Its rhetorical tics include a forced enthusiasm (a search on Twitter for the hashtag "#sciencedancing" speaks volumes) and a penchant for profanity. Here in Silicon Valley, one can scarcely go a day without seeing a t-shirt reading "Science: It works, b—es!" The hero of the recent popular movie *The Martian* boasts that he will "science the sh— out of" a situation. One of the largest groups on Facebook is titled "I f—ing love Science!" (a name which, combined with the group's penchant for posting scarcely any actual scientific material but a lot of pictures of natural phenomena, has prompted more than one actual scientist of my acquaintance to mutter under her breath, "What you truly love is *pictures*"). Some of the Cult's leaders like to play dress-up as scientists—Bill Nye and Neil deGrasse Tyson are two particularly prominent examples— but hardly any of them have contributed any research results of note. Rather, Cult leadership trends heavily in the direction of educators, popularizers, and journalists.

At its best, science is a human enterprise with a superhuman aim: the discovery of regularities in the order of nature, and the discerning of the consequences of those regularities. We've seen example after example of how the human element of this enterprise harms and damages its progress, through incompetence, fraud, selfishness, prejudice, or the simple combination of an honest oversight or slip with plain bad luck. These failings need not hobble the scientific enterprise broadly conceived, but only if

scientists are hyper-aware of and endlessly vigilant about the errors of their colleagues . . . and of themselves. When cultural trends attempt to render science a sort of religion-less clericalism, scientists are apt to forget that they are made of the same crooked timber as the rest of humanity and will necessarily imperil the work that they do. The greatest friends of the Cult of Science are the worst enemies of science's actual practice.

William A. Wilson is a software engineer in the San Francisco Bay Area.