their own worst instincts.

Regina Nuzzo

07 October 2015

Prir

Illustration by Dale Edwin Murray

In 2013, five years after he co-authored a paper showing that Democratic candidates in the United States could get more votes by moving slightly to the right on economic policy[1], Andrew Gelman, a statistician at Columbia University in New York City, was chagrined to learn of an error in the data analysis. In trying to replicate the work, an undergraduate student named Yang Yang Hu had discovered that Gelman had got the sign wrong on one of the variables.

Gelman immediately published a three-sentence correction, declaring that everything in the paper's crucial section should be considered wrong until proved otherwise.

Reflecting today on how it happened, Gelman traces his error back to the natural fallibility of the human brain: “The results seemed perfectly reasonable,” he says. “Lots of times with these kinds of coding errors you get results that are just ridiculous. So you know something's got to be wrong and you go back and search until you find the problem. If nothing seems wrong, it's easier to miss it.”

This is the big problem in science that no one is talking about: even an honest person is a master of self-deception. Our brains evolved long ago on the African savannah, where jumping to plausible conclusions about the location of ripe fruit or the presence of a predator was a matter of survival. But a smart strategy for evading lions does not necessarily translate well to a modern laboratory, where tenure may be riding on the analysis of terabytes of multidimensional data. In today's environment, our talent for jumping to conclusions makes it all too easy to find false patterns in randomness, to ignore alternative explanations for a result or to accept 'reasonable' outcomes without question — that is, to ceaselessly lead ourselves astray without realizing it.

Related stories

- First results from psychology’s largest reproducibility test
- Metascience could rescue the ‘replication crisis’
- Scientific method: Statistical errors

Failure to understand our own biases has helped to create a crisis of confidence about the reproducibility of published results, says statistician John Ioannidis, co-director of the Meta-Research Innovation Center at Stanford University in Palo Alto, California. The issue goes well beyond cases of fraud. Earlier this year, a large proje

replicate 100 psychology studies managed to reproduce only slightly more than one-third[2]. In 2

biotechnology firm Amgen in Thousand Oaks, California, reported that they could replicate only

studies in oncology and haematology[3]. And in 2009, Ioannidis and his colleagues described ho

microarray-based gene-expression studies[4].

Although it is impossible to document how often researchers fool themselves in data analysis,

explanation. The study of 100 psychology papers is a case in point: if one assumes that the va

diligent, then a large proportion of the problems can be explained only by unconscious biases.

realized that experimenters and subjects often unconsciously changed their behaviour to match expectations. From that insight, the double-blind standard was born.

“People forget that when we talk about the scientific method, we don't mean a finished product,” says Saul Perlmutter, an astrophysicist at the University of California, Berkeley. “Science is an ongoing race between our inventing ways to fool ourselves, and our inventing ways to avoid fooling ourselves.” So researchers are trying a variety of creative ways to debias data analysis — strategies that involve collaborating with academic rivals, getting papers accepted before the study has even been started and working with strategically faked data.

Read more: Replication studies: Bad copy

The problem

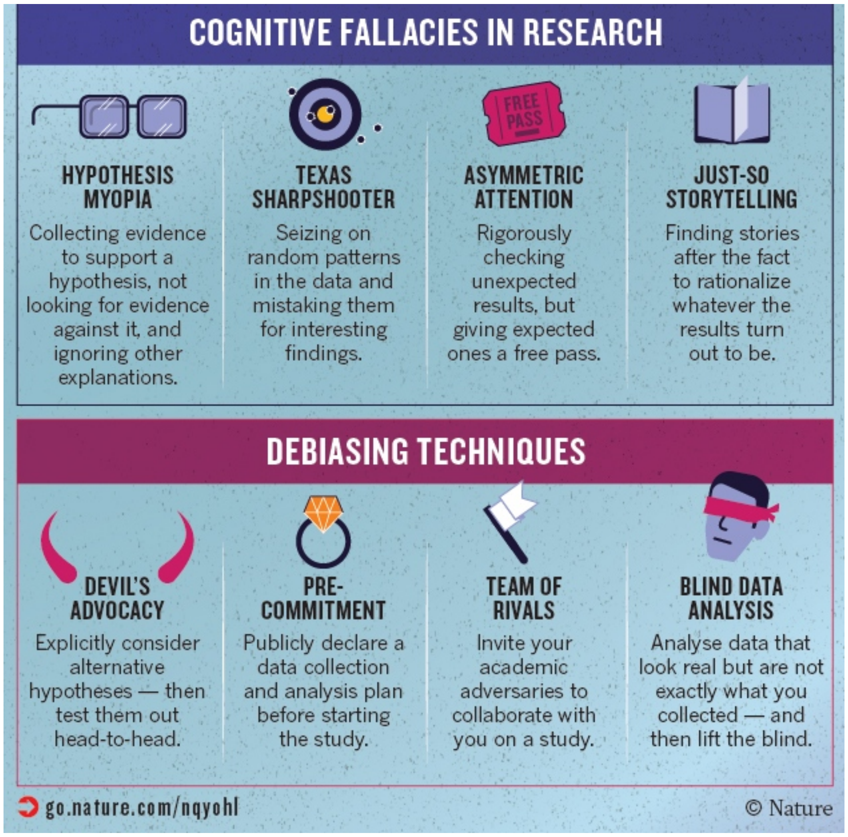
Although the human brain and its cognitive biases have been the same for as long as we have been doing science, some important things have changed, says psychologist Brian Nosek, executive director of the non-profit Center for Open Science in Charlottesville, Virginia, which works to increase the transparency and reproducibility of scientific research. Today's academic environment is more competitive than ever. There is an emphasis on piling up publications with statistically significant results — that is, with data relationships in which a commonly used measure of statistical certainty, the *p*-value, is 0.05 or less. “As a researcher, I'm not trying to produce misleading results,” says Nosek. “But I do have a stake in the outcome.” And that gives the mind excellent motivation to find what it is primed to find.

> **“I'm not trying to produce misleading results — but I do have a stake in the outcome.”**

Another reason for concern about cognitive bias is the advent of staggeringly large multivariate data sets, often harbouring only a faint signal in a sea of random noise. Statistical methods have barely caught up with such data, and our brain's methods are even worse, says Keith Baggerly, a statistician at the University of Texas MD Anderson Cancer Center in Houston. As he told a conference on challenges in bioinformatics last September in Research Triangle Park, North Carolina, “Our intuition when we start looking at 50, or hundreds of, variables sucks.”

Andrew King, a management specialist at Dartmouth College in Hanover, New Hampshire, says that the widespread use of point-and-click data-analysis software has made it easy for researchers to sift through massive data sets without fully understanding the methods, and to find small *p*-values that may not actually mean anything. “I believe we are in the steroids era of social science,” he says. “I've been guilty of using some of these performance-enhancing practices myself. My sense is that most researchers have fallen at least once.”

Just as in competitive sport, says Hal Pashler, a psychologist at the University of California, San Diego, this can set up a vicious circle of chasing increasingly better results. When a few studies in behavioural neuroscience started reporting improbably strong correlations of 0.85, Pashler says, researchers who had more moderate (and plausible) results started to worry: “Gee, I just got a 0.4, so maybe I'm not really doing this very well.”



COGNITIVE FALLACIES IN RESEARCH

HYPOTHESIS MYOPIA
Collecting evidence to support a hypothesis, not looking for evidence against it, and ignoring other explanations.

TEXAS SHARPSHOOTER
Seizing on random patterns in the data and mistaking them for interesting findings.

ASYMMETRIC ATTENTION
Rigorously checking unexpected results, but giving expected ones a free pass.

JUST-SO STORYTELLING
Finding stories after the fact to rationalize whatever the results turn out to be.

DEBIASING TECHNIQUES

DEVIL'S ADVOCACY
Explicitly consider alternative hypotheses — then test them out head-to-head.

PRE-COMMITMENT
Publicly declare a data collection and analysis plan before starting the study.

TEAM OF RIVALS
Invite your academic adversaries to collaborate with you on a study.

BLIND DATA ANALYSIS
Analyse data that look real but are not exactly what you collected — and then lift the blind.

go.nature.com/nqyohl

© Nature

View How scientists fool themselves — and how they can stop

Hypothesis myopia

One trap that awaits during the early stages of research is what might be called hypothesis myopia: investigators fixate on collecting evidence to support just one hypothesis; neglect to look for evidence against it; and fail to consider other explanations. “People tend to ask questions that give 'yes' answers if their favoured hypothesis is true,” says Jonathan Baron, a psychologist at the University of Pennsylvania in Philadelphia.

For example, says Baron, studies have tried to show how disgust influences moral condemnation, “by putting the subject in a messy room, or a room with 'fart spray' in the air”. The participants are then asked to judge how to respond to moral transgressions; if those who have been exposed to clutter or smells favour harsher punishments, researchers declare their 'disgust hypothesis' to be supported[5]. But they have not considered competing explanations, he says, and so they ignore the possibility that participants are lashing out owing to anger at their foul treatment, not simply disgust. By focusing on one hypothesis, researchers might be missing the real story entirely.

Nature special collection: reproducibility

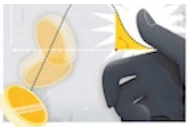
Courtrooms face a similar problem. In 1999, a woman in Britain called Sally Clark was found guilty of murdering two of her sons, who had died suddenly as babies. A factor in her conviction was the presentation of statistical evidence that the chances of two children in the same family dying of sudden infant death syndrome (SIDS) were only 1 in 73 million — a figure widely interpreted as fairly damning. Yet considering just one hypothesis leaves out an important part of the story. “The jury needs to weigh up two competing explanations for the babies' deaths: SIDS or murder,” wrote statistician Peter Green on behalf of the Royal Statistical Society in 2002 (see go.nature.com/ochsja). “The fact that two deaths by SIDS is quite unlikely is, taken alone, of little value. Two deaths by murder may well be even more unlikely. What matters is the relative likelihood of the deaths under each explanation, not just how unlikely they are under one explanation.” Mathematician Ray Hill of the University of Salford, UK, later estimated[6] that a double SIDS death would occur in roughly 1 out of 297,000 families, whereas two children would be murdered by a parent in roughly 1 out of 2.7 million families — a likelihood ratio of 9 to 1 against murder. In 2003, Clark's conviction was overturned on the basis of new evidence. The Attorney General for England and Wales went on to release two other women who had been convicted of murdering their children on similar statist

The Texas sharpshooter

A cognitive trap that awaits during data analysis is illustrated by the fable of the Texas sharpsh
bullets at the side of a barn, draws a target around the biggest clump of bullet holes, and point

His bullseye is obviously laughable — but the fallacy is not so obvious to gamblers who believe
people who see supernatural significance when a lottery draw comes up as all odd numbers.

hacking': "Exploiting — perhaps unconsciously — researcher degrees of freedom until $p < 0.05$." In 2012, a study of more than 2,000 US psychologists[7] suggested how common *p*-hacking is. Half had selectively reported only studies that 'worked', 58% had peeked at the results and then decided whether to collect more data, 43% had decided to throw out data only after checking its impact on the *p*-value and 35% had reported unexpected findings as having been predicted from the start, a practice that psychologist Norbert Kerr of Michigan State University in East Lansing has called HARKing, or hypothesizing after results are known. Not only did the researchers admit to these *p*-hacking practices, but they defended them.



The *P* value - not all it's cracked up to be

This May, a journalist described how he had teamed up with a German documentary filmmaker and demonstrated that creative *p*-hacking, carried out over one "beer-fueled" weekend, could be used to 'prove' that eating chocolate leads to weight loss, reduced cholesterol levels and improved well-being (see go.nature.com/blkpke). They gathered 18 different measurements — including weight, blood protein levels and sleep quality — on 15 people, a handful of whom had eaten some extra chocolate for a few weeks. With that many comparisons, the odds were better than 50–50 that at least one of them would look statistically significant just by chance. As it turns out, three of them did — and the team cherry-picked only those to report.

Asymmetric attention

The data-checking phase holds another trap: asymmetric attention to detail. Sometimes known as disconfirmation bias, this happens when we give expected results a relatively free pass, but we rigorously check non-intuitive results. "When the data don't seem to match previous estimates, you think, 'Oh, boy! Did I make a mistake?'" MacCoun says. "We don't realize that probably we would have needed corrections in the other situation as well."

> **"When the data don't seem to match previous estimates, you think, 'oh, boy! Did I make a mistake?'"**

The evidence suggests that scientists are more prone to this than one would think. A 2004 study[8] observed the discussions of researchers from 3 leading molecular-biology laboratories as they worked through 165 different lab experiments. In 88% of cases in which results did not align with expectations, the scientists blamed the inconsistencies on how the experiments were conducted, rather than on their own theories. Consistent results, by contrast, were given little to no scrutiny.

In 2011, an analysis of over 250 psychology papers found[9] that more than 1 in 10 of the *p*-values was incorrect — and that when the errors were big enough to change the statistical significance of the result, more than 90% of the mistakes were in favour of the researchers' expectations, making a non-significant finding significant.

Just-so storytelling

As data-analysis results are being compiled and interpreted, researchers often fall prey to just-so storytelling — a fallacy named after the Rudyard Kipling tales that give whimsical explanations for things such as how the leopard got its spots. The problem is that post-hoc stories can be concocted to justify anything and everything — and so end up truly explaining nothing. Baggerly says that he has seen such stories in genetics studies, when an analysis implicates a huge number of genes in a particular trait or outcome. "It's akin to a Rorschach test," he said at the bioinformatics conference. Researchers will find a story, he says, "whether it's there or not. The problem is that occasionally it ain't real."

Another temptation is to rationalize why results should have come up a certain way but did not — what might be called JARKing, or justifying after results are known. Matthew Hankins, a statistician at King's College London, has collected more than 500 creative phrases that researchers use to convince readers that their non-significant results are worthy of attention (see go.nature.com/pwctoq). These include "flirting with conventional levels of significance ($p > 0.1$)", "on the very fringes of significance ($p = 0.099$)" and "not absolutely significant but very probably so ($p > 0.05$)".

Read more: *P* values are just the tip of the shoddy stats iceberg

The solutions

In every one of these traps, cognitive biases are hitting the accelerator of science: the process of spotting potentially important scientific relationships. Countering those biases comes down to strengthening the 'brake': the ability to slow down, be sceptical of findings and eliminate false positives and dead ends.

> **"Science is an ongoing race between our inventing ways to fool ourselves, and our inventing ways to avoid fooling ourselves."**

One solution that is piquing interest revives an old tradition: explicitly considering competing hypotheses, and if possible working to develop experiments that can distinguish between them. This approach, called strong inference[10], attacks hypothesis myopia head on. Furthermore, when scientists make themselves explicitly list alternative explanations for their observations, they can re

In 2013, researchers reported[11] using strong-inference tec (*Engystomops pustulosus*) during mating calls. The existir theories — one in which females have a preset neural template for mating calls, and another ir signals such as the appearance of the males' vocal sacs. So the researchers developed an ex predictions. The results showed that females can use multisensory cues to judge attractivenes

Transparency

An even more radical extension of this idea is the introduction of registered reports: publications in which scientists present their research plans for peer review before they even do the experiment. If the plan is approved, the researchers get an 'in-principle' guarantee of publication, no matter how strong or weak the results turn out to be. This should reduce the unconscious temptation to warp the data analysis, says Pashler. At the same time, he adds, it should keep peer reviewers from discounting a study's results or complaining after results are known. “People are evaluating methods without knowing whether they're going to find the results congenial or not,” he says. “It should create a much higher level of honesty among referees.” More than 20 journals are offering or plan to offer some format of registered reports.

Team of rivals

When it comes to replications and controversial topics, a good debiasing approach is to bypass the typical academic back-and-forth and instead invite your academic rivals to work with you. An adversarial collaboration has many advantages over a conventional one, says Daniel Kahneman, a psychologist at Princeton University in New Jersey. “You need to assume you're not going to change anyone's mind completely,” he says. “But you can turn that into an interesting argument and intelligent conversation that people can listen to and evaluate.” With competing hypotheses and theories in play, he says, the rivals will quickly spot flaws such as hypothesis myopia, asymmetric attention or just-so storytelling, and cancel them out with similar slants favouring the other side.

Psychologist Eric-Jan Wagenmakers of the University of Amsterdam has engaged in this sort of proponent–sceptic collaboration, when he teamed up with another group in an attempt[12] to replicate its research suggesting that horizontal eye movements help people to retrieve events from their memory. It is often difficult to get researchers whose original work is under scrutiny to agree to this kind of adversarial collaboration, he says. The invitation is “about as attractive as putting one's head on a guillotine — there is everything to lose and not much to gain”. But the group that he worked with was eager to get to the truth, he says. In the end, the results were not replicated. The sceptics remained sceptical, and the proponents were not convinced by a single failure to replicate. Yet this was no stalemate. “Although our adversarial collaboration has not resolved the debate,” the researchers wrote, “it has generated new testable ideas and has brought the two parties slightly closer.” Wagenmakers suggests several ways in which this type of collaboration could be encouraged, including a prize for best adversarial collaboration, or special sections for such collaborations in top journals.

Blind data analysis

One debiasing procedure has a solid history in physics but is little known in other fields: blind data analysis (see page 187). The idea is that researchers who do not know how close they are to desired results will be less likely to find what they are unconsciously looking for[13].

One way to do this is to write a program that creates alternative data sets by, for example, adding random noise or a hidden offset, moving participants to different experimental groups or hiding demographic categories. Researchers handle the fake data set as usual — cleaning the data, handling outliers, running analyses — while the computer faithfully applies all of their actions to the real data. They might even write up the results. But at no point do the researchers know whether their results are scientific treasures or detritus. Only at the end do they lift the blind and see their true results — after which, any further fiddling with the analysis would be obvious cheating.

Read more: Blind analysis: Hide results to seek the truth

Perlmutter used this method for his team's work on the Supernova Cosmology Project in the mid-2000s. He knew that the potential for the researchers to fool themselves was huge. They were using new techniques to replicate estimates of two crucial quantities in cosmology — the relative abundances of matter and of dark energy — which together reveal whether the Universe will expand forever or eventually collapse into a Big Crunch. So their data were shifted by an amount known only to the computer, leaving them with no idea what their findings implied until everyone agreed on the analyses and the blind could be safely lifted. After the big reveal, not only were the researchers pleased to confirm earlier findings of an expanding Universe[14], Perlmutter says, but they could be more confident in their conclusions. “It's a lot more work in some sense, but I think it leaves you feeling much safer as you do your analysis,” he says. He calls blind data analysis “intellectual hygiene, like washing your hands”.

Data blinding particularly appeals to young researchers, Perlmutter says — not least because of the sense of suspense it gives. He tells the story of a recent graduate student who had spent two years under a data blind as she analysed pairs of supernova explosions. After a long group meeting, Perlmutter says, the student presented all her analyses and said that she was ready to unblind if everyone agreed.

“It was 6 o'clock in the evening and time for dinner,” says Perlmutter. And everyone in the audi
a very disappointing evening, and she's going to have to think really hard about what she's goi
morning.”

“And we all looked at each other, and we said, 'Nah! Let's unblind now!' So we unblinded, and
applauded.”

Nature **526,** 182–185 (08 October 2015) doi:10.1038/526182a

References

1. Gelman, A. & Cai, C. J. *Ann. Appl. Stat.* **2**, 536–549 (2008).
   Show context — Article
2. Open *Science* Collaboration. *Science* http://dx.doi.org/10.1126/science.aac4716 (2015).
   Show context — PubMed
3. Begley, C. G. & Ellis, L. M. *Nature* **483**, 531–533 (2012).
   Show context — Article PubMed ChemPort
4. Ioannidis, J. P. A. *et al. Nature Genet.* **41**, 149–155 (2009).
   Show context — Article PubMed ChemPort
5. Landy, J. F. & Goodwin, G. P. *Perspect. Psychol. Sci.* **10**, 518–536 (2015).
   Show context — Article PubMed
6. Hill, R. *Paediatr. Perinatal Epidemiol.* **18**, 320–326 (2004).
   Show context — Article
7. John, L. K., Loewenstein, G. & Prelec, D. *Psychol. Sci.* **23**, 524–532 (2012).
   Show context — Article PubMed
8. Fugelsang, J. A., Stein, C. B., Green, A. E. & Dunbar, K. N. *Can. J. Exp. Psychol.* **58**, 86–95 (2004).
   Show context — Article PubMed
9. Bakker, M. & Wicherts, J. M. *Behav. Res. Meth.* **43**, 666–68 (2011).
   Show context — Article
10. Platt, J. R. *Science* **146**, 347–353 (1964).
   Show context — Article PubMed ChemPort
11. Taylor, R. C. & Ryan, M. J. *Science* **341**, 273–274 (2013).
   Show context — Article PubMed ChemPort
12. Matzke, D. *et al. J. Exp. Psychol. Gen.* **144**, e1–e15 (2015).
   Show context — Article PubMed
13. MacCoun, R. & Perlmutter, S. in *Psychological Science Under Scrutiny: Recent Challenges and Proposed Solutions* (eds Lilienfeld, S. O. & Waldman, I.) (Wiley, in the press); Preprint available at http://ssrn.com/abstract=2563337
   Show context
14. Conley, A. *et al. Astrophys. J.* **644**, 1–20 (2006).
   Show context — Article

Related stories and links

From nature.com

- **First results from psychology's largest reproducibility test**
  30 April 2015
- **Metascience could rescue the 'replication crisis'**
  04 November 2014
- **Scientific method: Statistical errors**
  12 February 2014
- **Reproducibility: Six red flags for suspect work**
  22 May 2013
- **Replication studies: Bad copy**
  16 May 2012

From elsewhere

- **Center for Open Science**
- **Robert MacCoun**
- **Retraction Watch**

For the best commenting experience, please login or register as a user and agree to our Community Guidelines. You will be re-directed back to this page where you will see comments updating in real-time and have the ability to recommend comments to other users.

Commenting is currently unavailable.

Nature ISSN 0028-0836 EISSN 1476-4687