

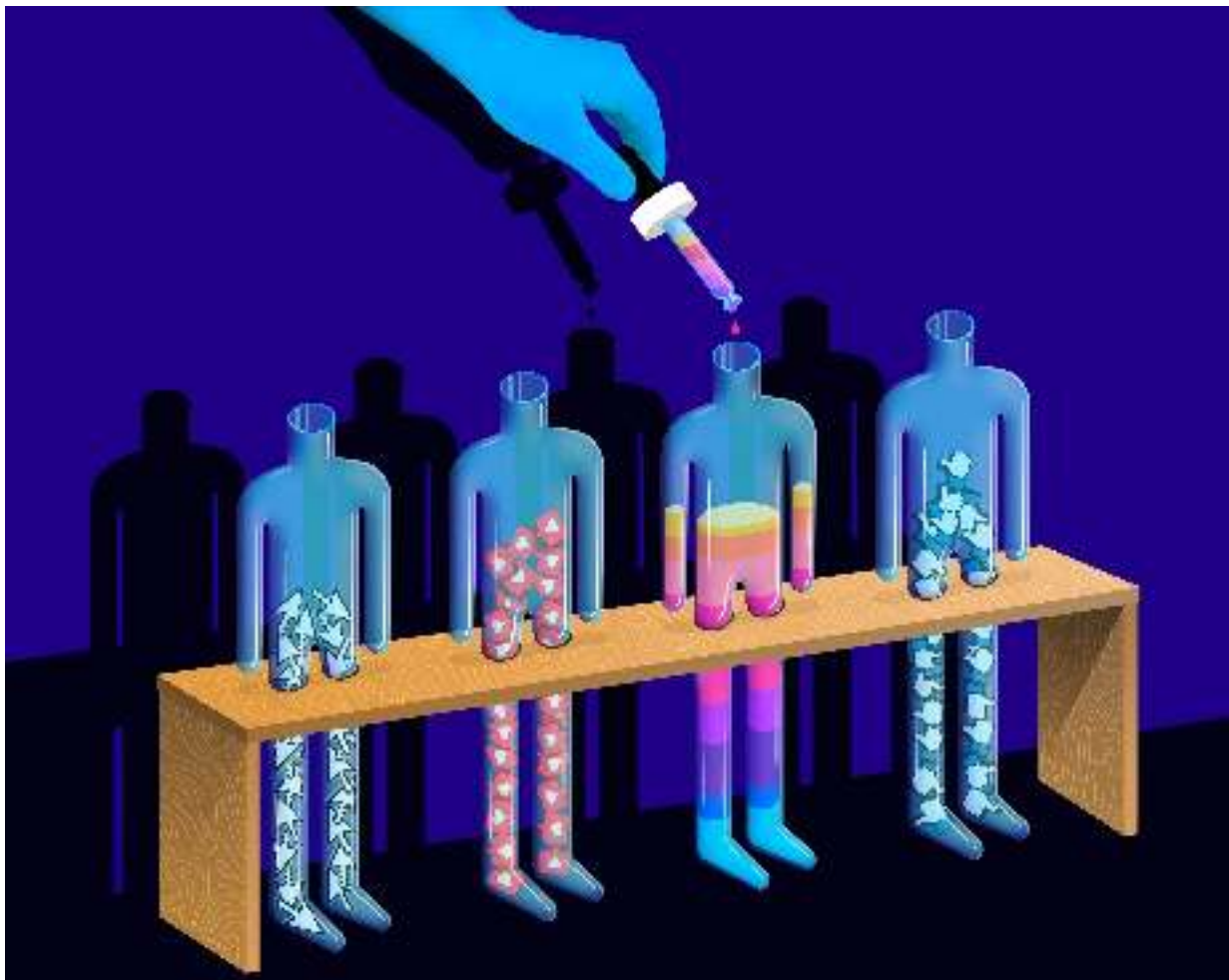
BOOKS MARCH 2, 2020 ISSUE

BIG TECH IS TESTING YOU

Large-scale social experiments are now ubiquitous, and conducted without public scrutiny. Has this new era of experimentation remembered the lessons of the old?

By Hannah Fry

February 24, 2020



For Big Tech, users are ideal test subjects. But the research we unknowingly participate in is seldom revealed. Illustration by Alexander Glandien

Dr. John Haygarth knew that there was something suspicious about Perkins's Metallic Tractors. He'd heard all the theories about the newly patented medical device—about the way flesh reacted to metal, about noxious electrical fluids being expelled from the body. He'd heard that people plagued by rheumatism, pleurisy, and toothache swore the instrument offered them miraculous relief. Even George Washington was said to own a set. But Haygarth, a physician who had pioneered a method of preventing smallpox, sensed a sham. He set out to find the evidence.

The year was 1799, and the Perkins tractors were already an international phenomenon. The device consisted of a pair of metallic rods—rounded on one end and tapering, at the other, to a point. Its inventor, Elisha Perkins, insisted that gently stroking each tractor over the affected area in alternation would draw off the electricity and provide relief. Thousands of sets were sold, for twenty-five dollars each. People were even said to have auctioned off their horses just to get hold of a pair. And, in an era when your alternatives might be bloodletting, leeches, and purging, you could see the appeal.

Haygarth had a pair of dummy tractors created, carved carefully from wood and painted to resemble the originals. They were to be used on five unsuspecting patients at Bath General Hospital, in England, each suffering from chronic rheumatism. Using the lightest of touches, the fakes were drawn over the affected areas, with remarkable results. Four of the five patients declared that their pain was relieved. One reported a tingling sensation that lasted for two hours. Another regained the ability to walk.

The following day, Haygarth repeated his test using the true metallic tractors, with the same results. Other physicians soon followed his lead, using increasingly elaborate fakes of their own: nails, pencils, even old tobacco pipes in place of the tractors. Each brought the truth more clearly into focus: the tractors were no better than make-believe.

This humble experiment wasn't the only one of its kind. By the start of the nineteenth century, experimentation had already driven two centuries of significant changes in science. The Royal Society of London, the scientific academy of which Haygarth was an elected fellow, began insisting that all claims needed to be verified and reproduced before they could be accepted as scientific fact. A shakeup was under way. Astronomy had split off from astrology. Chemistry had become disentangled from alchemy. The motto of the society neatly encapsulated the new spirit of inquiry: *Nullius in Verba*. Translation: "Take nobody's word for it."

Physics, chemistry, and medicine have had their revolution. But now, driven by experimentation, a further transformation is in the air. That's the argument of "The Power of Experiments" (M.I.T.), by Michael Luca and Max H. Bazerman, both professors at the Harvard Business School. When it comes to driving our decisions in a world of data, they say, "the age of experiments is only beginning."

In fact, if you've recently used Facebook, browsed Netflix, or run a Google search, you have almost certainly participated in an experiment of some kind. Google alone ran fifteen thousand of them in 2018, involving countless unsuspecting Internet users. "We don't want high-level executives discussing whether a blue background or a yellow background will lead to more ad clicks," Hal Varian, Google's chief economist, tells the authors. "Why debate this point, since we can simply run an experiment to find out?"

Luca and Bazerman focus on a new breed of large-scale social experiments, the power of which has already been demonstrated in the public sector. As they note, governments have used experiments to find better ways to get their citizens to pay taxes on time, say, or to donate organs after death. N.G.O.s have successfully deployed experiments in developing countries to test the effects of everything from tampons to textbooks. The impact of a simple experiment can be dramatic, particularly in monetary terms.

A few years ago, if you searched for eBay on Google, the top two results would take you directly to the auction site's home page. The second one was produced organically by the Google algorithm; the first was an advertisement, paid for by eBay and meant to pop up whenever its name appeared as a keyword in someone's search.

Steve Tadelis, a professor of economics at the University of California, Berkeley, was spending a year at eBay at the time, and was suspicious about the value of placing such ads. Wouldn't people get to eBay anyway if they were searching for it, without the sponsored results? But, as Luca and Bazerman recount, eBay's marketing group defended the millions of dollars spent on the ads each year, noting that many people who clicked on them ended up buying things on eBay.

An experiment was in order. By turning Google ads on and off, Tadelis and his research team tracked the traffic coming to their site and discovered that—as Tadelis had suspected—much of the money

eBay had been shelling out was wasted. The marketing team had an exaggerated notion of how valuable those ads were: without the sponsored result, searchers would simply click on the free organic links instead. The company could (and did) save itself millions.

There's an important point in all of this: instead of going by our possibly unreliable intuition, we can, in a range of cases, know for sure whether an intervention has an effect by running a trial and collecting the evidence. It's a step that Esther Duflo, who shared a Nobel Prize in Economics for her work using experiments to study how global poverty can be alleviated, makes a particularly strong case for. Without gathering and analyzing the evidence, she has said, "we are not any better than the medieval doctors and their leeches."

The most reliable way to test an intervention is by using a type of experiment known as a "randomized controlled trial" (R.C.T.). You randomly assign subjects to groups and then treat each group differently. One group will receive the intervention, while another, the "control" group, will not. Control here is key. The aim is to make the groups as similar as possible, to constrain as many variables as you can manage, because if the only thing allowed to change freely is the intervention itself you can study its true effect. In the tech world, the "intervention" might simply be a different Web-page layout or a new pricing plan. Here, the usual term is "A/B testing," but the objective is the same: to create some basis for comparison.

Such studies tell you whether something works, though not why. Haygarth's experiment wasn't a randomized trial by modern standards, but he nonetheless proved the power of experimenting: by directly comparing the experiences of patients on the day they got treated with the tractors with their experiences on the day they were treated with the fakes, he could show that the tractors were duds. The second set of observations served as a kind of control group.

Without a properly randomized control group, there is no real way to measure whether something is working. Take the case of the Scared Straight program, developed in the United States to discourage at-risk kids from choosing a life of crime. The theory seemed sound. By taking young offenders on organized visits to prison and allowing them to meet murderers and armed robbers, they'd see the terrifying consequences of breaking the law, and be less likely to do so themselves in the future.

At first, the program appeared to be a roaring success. Few kids who had been through the program were later involved in crime (as many as ninety-four per cent steered clear, according to one study). The intervention gained public approval and was copied in a number of countries around the world. There was just one tiny problem: no one had stopped to wonder what would have happened to a similar set of kids who hadn't gone through the experience.

When a series of proper R.C.T.s was run, and a direct comparison was made between kids who went through the Scared Straight program and similar groups of kids who didn't, it became clear that the program was not working as intended. The intervention was, in fact, *increasing* the chances that kids would become criminals. Fewer juveniles over all would have ended up committing further crimes if they'd never been taken to visit the jails.

You can see why these experiments have become ubiquitous. They're the ultimate fact finder, overriding the need to rely on intuition. But, for some, their popularity is also a cause for concern, because they aren't always used to nudge our behavior toward the greater good. They're also helping organizations to manipulate us in ways that might not be in our interests.

As "The Power of Experiments" makes clear, there are times when this happens in irritating but relatively harmless ways—a company making a small tweak to a Web site that elevates profits over customer experience, for instance. Consider an experiment that StubHub, the ticket-resale company, ran to determine where best to notify users about its ticketing costs. Should it be up front about them from the moment you land on the page? Or surprise you at checkout? StubHub discovered, after experimenting, that hiding the fees until the last minute led to thirteen per cent more sales, plus tickets that were 5.73 per cent more expensive on average. As Luca and Bazerman explain, "People were buying better, higher-priced tickets when the fees were hidden." The technique did make people less likely to return to the Web site in the following months, but that falloff was not enough to counter the increase in ticket sales and prices.

There are also times when manipulation leaves people feeling cheated. For instance, in 2018 the *Wall Street Journal* reported that Amazon had been inserting sponsored products in its consumers' baby registries. "The ads look identical to the rest of the listed products in the registry, except for a small gray 'Sponsored' tag," the *Journal* revealed. "Unsuspecting friends and family clicked on the ads and

purchased the items,” assuming they’d been chosen by the expectant parents. Amazon’s explanation when confronted? “We’re constantly experimenting,” a spokesperson said. (The company has since ended the practice.)

But there are times when the experiments go further still, leaving some to question whether they should be allowed at all. There was a notorious experiment run by Facebook in 2012, in which the number of positive and negative posts in six hundred and eighty-nine thousand users’ news feeds was tweaked. The aim was to see how the unwitting participants would react. As it turned out, those who saw less negative content in their feeds went on to post more positive stuff themselves, while those who had positive posts hidden from their feeds used more negative words.

A public backlash followed; people were upset to discover that their emotions had been manipulated. Luca and Bazerman argue that this response was largely misguided. They point out that the effect was small. A person exposed to the negative news feed “ended up writing about four additional negative words out of every 10,000,” they note. Besides, they say, “advertisers and other groups manipulate consumers’ emotions all the time to suit their purposes. If you’ve ever read a Hallmark card, attended a football game or seen a commercial for the ASPCA, you’ve been exposed to the myriad ways in which products and services influence consumers’ emotions.”

In my view, the authors are too sanguine on this issue. It isn’t helpful to excuse an experiment based on the average effect it had on hundreds of thousands of people. This wasn’t four words each, shared out equally among everyone. Nor is there a way of knowing the initial conditions of those involved in the study. It may be that some people were emotionally stable enough for a tiny effect to make no difference but that others were already close to a tipping point.

It’s true that we inhabit a world in which all manner of companies are continually seeking to influence our behavior. But the sequence of messages you see during halftime at the Super Bowl isn’t algorithmically curated for your eyes only, based on vast swathes of your highly personalized data. We don’t want to live in “The Truman Show.” Whatever you think about the particulars of the Facebook study, though, it raised an important question about the care with which experiments on human subjects are handled.

Medicine has already been through this. In the early twentieth century, without a set of ground rules on how people should be studied, medical experimentation was like the Wild West. Alongside a great deal of good work, a number of deeply unethical studies took place—including the horrifying experiments conducted by the Nazis and the appalling Tuskegee syphilis trial, in which hundreds of African-American men were denied treatment by scientists who wanted to see how the lethal disease developed. As a result, there are now clear rules about seeking informed consent whenever medical experiments use human subjects, and institutional procedures for reviewing the design of such experiments in advance. We've learned that researchers aren't always best placed to assess the potential harm of their work.

So what happens now that data scientists have arrived on the scene? They, too, are conducting experiments on people, but they're doing it without that history of public outrage, without a long list of successive scandals, and, crucially, without the scrutiny and ethical framework you'll find in the biomedical sciences. Luca and Bazerman advocate greater voluntary transparency; I'm not convinced that this approach is enough to protect the public from being exploited.

That's not to say that we need a review board every time an online publication runs an A/B test to see which of two headlines draws more readers. Still, if we don't hold everyone to some reasonable ethical standards, a social-media company, in theory, would have free rein to manipulate the amount of politically extreme or anti-climate-change stories displayed in your feed. For that matter, it's possible to imagine a ride-hailing app experimenting to see whether it could wring more money from you when your phone battery was running low. Or a driverless car being deployed on the roads before it was fully functional, leaving bystanders as unwitting participants in a live experiment—although, come to think of it, that scenario doesn't exactly require imagining.

Of course, there's a lot that we don't understand about human behavior which experiments are perfectly placed to explore. They can seem to be self-interpreting, telling us that some treatment or intervention works even when we don't know why. In what was perhaps the first-ever controlled trial, a naval doctor in the seventeen-forties demonstrated that sailors who ate citrus fruits were less afflicted by scurvy, without knowing that the disease was caused by a Vitamin C deficiency. We're still not sure how general anesthesia works, despite having heavily relied on it since the eighteen-fifties; the same is true of many psychiatric drugs.

But, as Luca and Bazerman note, this particular power of randomized experiments—their ability to yield outcomes in advance of explanations—can also be their greatest weakness. When the results arrive without a sound theory of what caused them, we can easily overgeneralize or make erroneous inferences. The conclusions researchers reach can be fantastically sensitive to the way the trial is designed.

At the gym chain 24 Hour Fitness, a team of behavioral scientists wanted to see how they might nudge people into exercising more. Over twenty-eight days, fifty-two interventions were tested (among them text reminders and supportive video material). All were successful at helping people increase their attendance. Had the scientists left the study there, it might have appeared as though they'd found a multitude of ways to get us all into shape. But the scientists knew better. They followed up with the participants of the study beyond the initial period, and discovered that, in fact, none of the interventions produced any lasting change. Tweak the question just slightly, adjust the time frame of investigation, and the answer may come out different, too.

The extreme sensitivity of experiments certainly extends to the selection of participants. Even a slight imbalance can wildly throw off the conclusions. Luca and Bazerman offer the cautionary tale of hormone-replacement therapy (H.R.T.), which, for two decades, was widely prescribed to postmenopausal women as a preventative drug for various conditions associated with aging, including heart disease. A series of studies had demonstrated its impact. Unfortunately, the women in the studies who'd been given H.R.T. had, on average, a higher socioeconomic status than those who hadn't. They were more likely to have good access to medical care, more likely to have health issues diagnosed and treated early. This meant that no one could be sure it was the hormone therapy itself that was yielding the benefit. Indeed, a later large-scale experiment, better randomized across the groups, suggested that H.R.T. was actually increasing the odds of heart disease for some women. (Yet more studies have, in turn, questioned this conclusion; the therapy remains in medical limbo.)

Context matters. The fact that an intervention has been proved to work in one setting doesn't guarantee that it will work in another. This is a particular problem in the behavioral sciences. Angus Deaton, a Nobel laureate in economics, has argued that any such experiment that has been constrained enough to be scientifically rigorous might be too narrow to provide useful guidance for large-scale interventions.

There's a cautionary tale to be found in an experiment that started in Tamil Nadu, a state in southern India that had a serious problem with infant mortality. Babies were being born malnourished after their mothers, worried about the dangers of giving birth to a large baby, were deliberately undereating as they approached labor. Aid agencies designed and tested a program that offered mothers reassurance about advances in maternity care, food supplements, advice, and education. It was a great success. Malnutrition went way down.

Satisfied that a remedy had been found, the project team took the program to Bangladesh, which was suffering from a similar problem: pregnant women worried about childbirth restricting their diet before labor, resulting in malnourished babies fighting to survive. And yet in Bangladesh the intervention made no meaningful difference to the women or their children. Why? The two settings were demographically similar, but there were subtle cultural differences. Part of the answer, as Michael Blastland explains in his new book, "The Hidden Half," was that "in Bangladesh it was not the mother who controlled the family food. It was the mother-*in-law*."

It's tempting to look for laws of people the way we look for the laws of gravity. But science is hard, people are complex, and generalizing can be problematic. Although experiments might be the ultimate truth-tellers, they can also lead us astray in surprising ways.

That may be the most important point to remember in this new age of experiments. Science knows that it makes mistakes. It's why the Royal Society, since its inception, has placed so much emphasis on verifying claims and reproducing results. It's why Robert K. Merton, the American sociologist who laid out an "ethos of science" in the nineteen-forties, included skepticism as a fundamental tenet. And it's why he insisted that science should be based on a pillar of communality, recognizing that every idea in science is built on the ideas of others, and should therefore belong to the scientific community. Knowledge shouldn't be owned by its immediate discoverer.

There's untold good that can be done by experimentation in the digital age. It can help us to understand the impact of screen time or the Like button on our well-being; to find and fix discriminatory practices; to identify ways of promoting healthier life styles. But where these experiments are being done away from public scrutiny, the ethos of science is compromised. The Big Tech companies can tell us their findings. I'm just not sure it's enough to take their word for it. ♦

Published in the print edition of the March 2, 2020, issue, with the headline “Experiments on Trial.”

Hannah Fry is a professor at University College London’s Centre for Advanced Spatial Analysis. Her latest book is “Hello World.”

More: [Experiments](#) [Science](#) [Sociology](#)
