CHAPTER 1

The Sin of Bias

The human understanding when it has once adopted an opinion . . . draws all things else to support and agree with it.

—Francis Bacon, 1620

History may look back on 2011 as the year that changed psychology forever. It all began when the *Journal of Personality and Social Psychology* published an article called "Feeling the Future: Experimental Evidence for Anomalous Retroactive Influences on Cognition and Affect."[1] The paper, written by Daryl Bem of Cornell University, reported a series of experiments on *psi* or "precognition," a supernatural phenomenon that supposedly enables people to see events in the future. Bem, himself a reputable psychologist, took an innovative approach to studying *psi*. Instead of using discredited parapsychological methods such as card tasks or dice tests, he selected a series of gold-standard psychological techniques and modified them in clever ways.

One such method was a reversed priming task. In a typical priming task, people decide whether a picture shown on a computer screen is linked to a positive or negative emotion. So, for example, the participant might decide whether a picture of kittens is pleasant or unpleasant. If a word that "primes" the same emotion is presented immediately before the picture (such as the word "joy" followed by the picture of kittens), then people find it easier to judge the emotion of the picture, and they respond faster. But if the prime and target trigger opposite emotions then the task becomes more difficult because the emotions conflict (e.g., the word "murder" followed by kittens). To test for the existence of precognition, Bem reversed the order of this experiment and found that primes delivered *after* people had responded seemed to influence their reaction times. He also reported similar "retroactive" effects on memory. In one of his experiments, people were overall better at recalling specific words from a list that were also included in a practice task, with the catch that the so-called practice was undertaken *after* the recall task rather than before. On this basis, Bem argued that the participants were able to benefit in the past from practice they had completed in the future.

As you might expect, Bem's results generated a flood of confusion and controversy. How could an event in the future possibly influence someone's reaction time or memory in the past? If precognition truly did exist, in even a tiny minority of the population, how is it that casinos or stock markets turn profits? And how could such a bizarre conclusion find a home in a reputable scientific journal?

Scrutiny at first turned to Bem's experimental procedures. Perhaps there was some flaw in the methods that could explain his results, such as failing to randomize the order of events, or some other subtle experimental error. But these aspects of the experiment seemed to pass muster, leaving the research community facing a dilemma. If true, precognition would be the most sensational discovery in modern science. We would have to accept the existence of time travel and reshape our entire understanding of cause and effect. But if false, Bem's results would instead point to deep flaws in standard research practices—after all, if accepted practices could generate such nonsensical findings, how can any published findings in psychology be trusted? And so psychologists faced an unenviable choice between, on the one hand, accepting an impossible scientific conclusion and, on the other hand, swallowing an unpalatable professional reality.

The scientific community was instinctively skeptical of Bem's conclusions. Responding to a preprint of the article that appeared in late 2010, the psychologist Joachim Krueger said: "My personal view is that this is ridiculous and can't be true."[2] After all, extraordinary claims require extraordinary evidence, and despite being published in a prestigious journal, the statistical strength of Bem's evidence was considered far from extraordinary.

Bem himself realized that his results defied explanation and stressed the need for independent researchers to replicate his findings. Yet doing so proved more challenging than you might imagine. One replication attempt by Chris French and Stuart Ritchie showed no evidence whatsoever of precognition but was rejected by the same journal that published Bem's paper. In this case the journal didn't even bother to peer review French and Ritchie's paper before rejecting it, explaining that it "does not publish replication studies, whether successful or unsuccessful."[3] This decision may sound bizarre, but, as we will see, contempt for replication is common in psychology compared with more established sciences. The most prominent psychology journals selectively publish findings that they consider to be original, novel, neat, and above all positive. This *publication bias*, also known as the "file-drawer effect," means that studies that fail to show statistically significant effects, or that reproduce the work of others, have such low priority that they are effectively censored from the scientific record. They either end up in the file drawer or are never conducted in the first place.

Publication bias is one form of what is arguably the most powerful fallacy in human reasoning: *confirmation bias*. When we fall prey to confirmation bias, we seek out and favor evidence that agrees with our existing beliefs, while at the same time ignoring or devaluing evidence that doesn't. Confirmation bias corrupts psychological science in several ways. In its simplest form, it favors the publication of positive results—that is, hypothesis tests that reveal statistically significant differences or associations between conditions (e.g., A is greater than B; A is related to B, vs. A is the same as B; A is unrelated to B). More insidiously, it contrives a measure of scientific reproducibility in which it is possible to replicate but never falsify previous findings, and it encourages altering the hypotheses of experiments after the fact to "predict" unexpected outcomes. One of the most troubling aspects of psychology is that the academic community has refused to unanimously condemn such behavior. On the contrary, many psychologists acquiesce to these practices and even embrace them as survival skills in a culture where researchers must *publish or perish*.

Within months of appearing in a top academic journal, Bem's claims about precognition were having a powerful, albeit unintended, effect on the psychological community. Established methods and accepted publishing practices fell under renewed scrutiny for producing results that appear convincing but are almost certainly false. As psychologist Eric-Jan Wagenmakers and colleagues noted in a statistical demolition of Bem's paper: "Our assessment suggests that something is deeply wrong with the way experimental psychologists design their studies and report their statistical results."[4] With these words, the storm had broken.

A Brief History of the "Yes Man"

To understand the different ways that bias influences psychological science, we need to take a step back and consider the historical origins and basic research on confirmation bias. Philosophers and scholars have long recognized the "yes man" of human reasoning. As early as the fifth century BC, the historian Thucydides noted words to the effect that "[w]hen a man finds a conclusion agreeable, he accepts it without argument, but when he finds it disagreeable, he will bring against it all the forces of logic and reason." Similar sentiments were echoed by Dante, Bacon, and Tolstoy. By the mid-

twentieth century, the question had evolved from one of philosophy to one of science, as psychologists devised ways to measure confirmation bias in controlled laboratory experiments.

Since the mid-1950s, a convergence of studies has suggested that when people are faced with a set of observations (data) and a possible explanation (hypothesis), they favor tests of the hypothesis that seek to confirm it rather than falsify it. Formally, what this means is that people are biased toward estimating the probability of data if a particular hypothesis is *true*, *p*(data|hypothesis) rather than the opposite probability of it being false, *p*(data|~hypothesis). In other words, people prefer to ask questions to which the answer is "yes," ignoring the maxim of philosopher Georg Henrik von Wright that "no confirming instance of a law is a verifying instance, but . . . any disconfirming instance is a falsifying instance."[5]

Psychologist Peter Wason was one of the first researchers to provide laboratory evidence of confirmation bias. In one of several innovative experiments conducted in the 1960s and 1970s, he gave participants a sequence of numbers, such as 2-4-6, and asked them to figure out the rule that produced it (in this case: *three numbers in increasing order of magnitude*).[6] Having formed a hypothesis, participants were then allowed to write down their own sequence, after which they were told whether their sequence was consistent or inconsistent with the actual rule. Wason found that participants showed a strong bias to test various hypotheses by *confirming* them, even when the outcome of doing so failed to eliminate plausible alternatives (such as *three even numbers*). Wason's participants used this strategy despite being told in advance that "your aim is not simply to find numbers which conform to the rule, but to discover the rule itself."

Since then, many studies have explored the basis of confirmation bias in a range of laboratory-controlled situations. Perhaps the most famous of these is the ingenious Selection Task, which was also developed by Wason in 1968.[7] The Selection Task works like this. Suppose I were to show you four cards on a table, labeled D, B, 3, and 7 (see figure 1.1). I tell you that if the card shows a letter on one side then it will have a number on the other side, and I provide you with a more specific rule (hypothesis) that may be true or false: *"If there is a D on one side of any card, then there is a 3 on its other side."* Finally, I ask you to tell me which cards you would need to turn over in order to determine whether this rule is true or false. Leaving an informative card unturned or turning over an uninformative card (i.e., one that

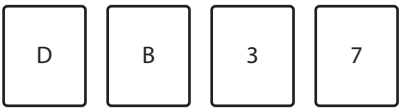

FIGURE 1.1. Peter Wason's Selection Task for measuring confirmation bias. Four cards are placed face down on a table. You're told that if there is letter on one side then there will always be a number on the other side. Then you are given a specific hypothesis: *If there is a D on one side then there is a 3 on its other side*. Which cards would you turn over to test whether this hypothesis is true or false?

doesn't test the rule) would be considered an incorrect response. Before reading further, take a moment and ask yourself, which cards would you choose and which would you avoid?

If you chose D and avoided B then you're in good company. Both responses are correct and are made by the majority of participants. Selecting D seeks to test the rule by confirming it, whereas avoiding B is correct because the flip side would be uninformative regardless of the outcome.

Did you choose 3? Wason found that most participants did, even though 3 should be avoided. This is because if the flip side *isn't* a D, we learn nothing—the rule states that cards with D on one side are paired a 3 on the other, not that D is the *only* letter to be paired with a 3 (drawing such a conclusion would be a logical fallacy known as "affirming the consequent"). And even if the flip side *is* a D then the outcome would be consistent with the rule but wouldn't confirm it, for exactly the same reason.

Finally, did you choose 7 or avoid it? Interestingly, Wason found that few participants selected 7, even though doing so is correct—in fact, it is just as correct as selecting D. If the flip side to 7 were discovered to be a D then the rule would be categorically disproven—a logical test of what's known as the "contrapositive." And herein lies the key result: the fact that most participants correctly select D but fail to select 7 provides evidence that people seek to test rules or hypotheses by confirming them rather than by falsifying them.

Wason's findings provided the first laboratory-controlled evidence of confirmation bias, but centuries of informal observations already pointed strongly to its existence. In a landmark review, psychologist Raymond Nickerson noted how confirmation bias dominated in the witchcraft trials of the middle ages.[8] Many of these proceedings were a foregone conclusion, seeking only to obtain evidence that confirmed the guilt of the accused. For

instance, to test whether a person was a witch, the suspect would often be plunged into water with stones tied to her feet. If she rose then she would be proven a witch and burned at the stake. If she drowned then she was usually considered innocent or a witch of lesser power. Either way, being suspected of witchcraft was tantamount to a death sentence within a legal framework that sought only to confirm accusations. Similar biases are apparent in many aspects of modern life. Popular TV programs such as *CSI* fuel the impression that forensic science is bias-free and infallible, but in reality the field is plagued by confirmation bias.[9] Even at the most highly regarded agencies in the world, forensic examiners can be biased toward interpreting evidence that confirms existing suspicions. Doing so can lead to wrongful convictions, even when evidence is based on harder data such as fingerprints and DNA tests.

Confirmation bias also crops up in the world of science communication. For many years it was assumed that the key to more effective public communication of science was to fill the public's lack of knowledge with facts—the so-called deficit model.[10] More recently, however, this idea has been discredited because it fails to take into account the prior beliefs of the audience. The extent to which we assimilate new information about popular issues such as climate change, vaccines, or genetically modified foods is susceptible to a confirmation bias in which evidence that is consistent with our preconceptions is favored, while evidence that flies in the face of them is ignored or attacked. Because of this bias, simply handing people more facts doesn't lead to more rational beliefs. The same problem is reflected in politics. In his landmark 2012 book, the *Geek Manifesto*, Mark Henderson laments the cherry-picking of evidence by politicians in order to reinforce a predetermined agenda. The resulting "policy-based evidence" is a perfect example of confirmation bias in practice and represents the antithesis of how science should be used in the formulation of evidence-based policy.

If confirmation bias is so irrational and counterproductive, then why does it exist? Many different explanations have been suggested based on cognitive or motivational factors. Some researchers have argued that it reflects a fundamental limit of human cognition. According to this view, the fact that we have incomplete information about the world forces us to rely on the memories that are most easily retrieved (the so-called availability heuristic), and this reliance could fuel a bias toward what we think we already know. On the other hand, others have argued that confirmation bias

is the consequence of an innate "positive-test strategy"—a term coined in 1987 by psychologists Joshua Klayman and Young-Won Ha.[11] We already know that people find it easier to judge whether a positive statement is true or false (e.g., "there are apples in the basket") compared to a negative one ("there are no apples in the basket"). Because judgments of presence are easier than judgments of absence, it could be that we prefer positive tests of reality over negative ones. By taking the easy road, this bias toward positive thoughts could lead us to wrongly accept evidence that agrees positively with our prior beliefs.

Against this backdrop of explanations for why an irrational bias is so pervasive, psychologists Hugo Mercier and Dan Sperber have suggested that confirmation bias is in fact perfectly rational in a society where winning arguments is more important than establishing truths.[12] Throughout our upbringing, we are taught to defend and justify the beliefs we hold, and less so to challenge them. By interpreting new information according to our existing preconceptions we boost our self-confidence and can argue more convincingly, which in turn increases our chances of being regarded as powerful and socially persuasive. This observation leads us to an obvious proposition: If human society is constructed so as to reward the act of winning rather than being correct, who would be surprised to find such incentives mirrored in scientific practices?

Neophilia: When the Positive and New Trumps the Negative but True

The core of any research psychologist's career—and indeed many scientists in general—is the rate at which they publish empirical articles in high-quality peer-reviewed journals. Since the peer-review process is competitive (and sometimes extremely so), publishing in the most prominent journals equates to a form of "winning" in the academic game of life.

Journal editors and reviewers assess submitted manuscripts on many grounds. They look for flaws in the experimental logic, the research methodology, and the analyses. They study the introduction to determine whether the hypotheses are appropriately grounded in previous research. They scrutinize the discussion to decide whether the paper's conclusions are justified by the evidence. But reviewers do more than merely critique

the rationale, methodology, and interpretation of a paper. They also study the results themselves. How important are they? How exciting? How much have we learned from this study? Is it a breakthrough? One of the central (and as we will see, lamentable) truths in psychology is that exciting positive results are a key factor in publishing—and often a requirement. The message to researchers is simple: if you want to win in academia, publish as many papers as possible in which you provide positive, novel results.

What does it mean to find "positive" results? Positivity in this context doesn't mean that the results are uplifting or good news—it refers to whether the researchers found a reliable difference in measurements, or a reliable relationship, between two or more study variables. For example, suppose you wanted to test the effect of a cognitive training intervention on the success of dieting in people trying to lose weight. First you conduct a literature review, and, based on previous studies, you decide that boosting people's self-control might help. Armed with a good understanding of existing work, you design a study that includes two groups. The experimental group perform a computer task in which they are trained to respond to images of foods, but crucially, to refrain from responding to images of particular junk foods. They perform this task every day for six weeks, and you measure how much weight they lose by the end of the experiment. The control group does a similar task with the same images but responds to all of them—and you measure weight loss in that group as well.

The null hypothesis (called "H0") in this case is that there should be no difference in weight loss—your training intervention has no effect on whether people gain or lose weight. The alternative hypothesis (called "H1") is that the training intervention should boost people's ability to refrain from eating junk foods, and so the amount of weight loss should be greater in the treatment group compared with the control group. A positive result would be finding a statistically significant difference in weight loss between the groups (or in technical terms, "rejecting H0"), and a negative result would be failing to show any significant difference (or in other words, "failing to reject H0"). Note how I use the term "failing." This language is key because, in our current academic culture, journals indeed regard such outcomes as scientific failures. Regardless of the fact that the rationale and methods are identical in each outcome, psychologists find negative results much harder to publish than positive results. This is because positive results are regarded by journals as reflecting a greater degree of scientific advance and interest

to readers. As one journal editor said to me, "Some results are just more interesting and important than others. If I do a randomized trial on a novel intervention based on a long-shot and find no effect that is not a great leap forward. However, if the same study shows a huge benefit that is a more important finding."

This publication bias toward positive results also arises because of the nature of conventional statistical analyses in psychology. Using standard methods developed by Neyman and Pearson, positive results reject H0 in favor of the alternative hypothesis (H1). This statistical approach—called null hypothesis significance testing—estimates the probability (p) of an effect of the same or greater size being obtained if the null hypothesis were true. Crucially, it doesn't estimate the probability of the null hypothesis *itself* being true: p values estimate the probability of a given effect or more extreme arising given the hypothesis, rather than the probability of a particular hypothesis given the effect. This means that while a statistically significant result (by convention, $p<.05$) allows the researcher to *reject* H0, a statistically nonsignificant result ($p>.05$) doesn't allow the researcher to *accept* H0. All the researcher can conclude from a statistically nonsignificant outcome is that H0 might be true, or that the data might be insensitive. The interpretation of statistically nonsignificant effects is therefore inherently inconclusive.

Consider the thought process this creates in the minds of researchers. If we can't test directly whether there is *no* difference between experimental conditions, then it makes little sense to design an experiment in which the null hypothesis would ever be the focus of interest. Instead, psychologists are trained to design experiments in which findings of interest would always be *positive*. This bias in experimental design, in turn, means that students in psychology enter their research careers reciting the mantra "Never predict the null hypothesis." If researchers can never predict the null hypothesis, and if positive results are considered more interesting to journals than negative results, then the inevitable outcome is a bias in which the peer-reviewed literature is dominated by positive findings that reject H0 in favor of H1, and in which most of the negative or nonsignificant results remain unpublished. To ensure that they keep winning in the academic game, researchers are thus pushed into finding positive results that agree with their expectations—a mechanism that incentivizes and rewards confirmation bias.

All this might sound possible in theory, but is it true? Psychologists have known since the 1950s that journals are predisposed toward publishing positive results, but, historically, it has been difficult to quantify how much publication bias there really is in psychology.[13] One of the most compelling analyses was reported in 2010 by psychologist Daniele Fanelli from the University of Edinburgh.[14] Fanelli reasoned, as above, that any domain of the scientific literature that suffers from publication bias should be dominated by positive results that support the stated hypothesis (H1). To test this idea, he collected a random sample of more than 2,000 published journal articles from across the full spectrum of science, ranging from the space sciences to physics and chemistry, through to biology, psychology, and psychiatry. The results were striking. Across all sciences, positive outcomes were more common than negative ones. Even for space science, which published the highest percentage of negative findings, 70 percent of the sampled articles supported the stated hypothesis. Crucially, this bias was highest in psychology, topping out at 91 percent. It is ironic that psychology—the discipline that produced the first empirical evidence of confirmation bias—is at the same time one of the most vulnerable *to* confirmation bias.

The drive to publish positive results is a key cause of publication bias, but it still explains only half the problem. The other half is the quest for novelty. To compete for publication at many journals, articles must either adopt a novel methodology or produce a novel finding—and preferably both. Most journals that publish psychological research judge the merit of manuscripts, in part, according to novelty. Some even refer explicitly to novelty as a policy for publication. The journal *Nature* states that to be considered for peer review, results must be "novel" and "arresting,"[15] while the journal *Cortex* notes that empirical Research Reports must "report important and novel material."[16] The journal *Brain* warns authors that "some [manuscripts] are rejected without peer review owing to lack of novelty,"[17] and *Cerebral Cortex* goes one step further, noting that even after peer review, "final acceptance of papers depends not just on technical merit, but also on subjective ratings of novelty."[18] Within psychology proper, *Psychological Science*, a journal that claims to be the highest-ranked in psychology, prioritizes papers that produce "breathtaking" findings.[19]

At this point, you might well ask: what's wrong with novelty? After all, in order for something to be marked as discovered, surely it can't have been observed already (so it must be a novel result), and isn't it also reasonable

to assume that researchers seeking to produce novel results might need to adopt new methods? In other words, by valuing novelty aren't journals simply valuing discovery? The problem with this argument is the underlying assumption that every observation in psychological research can be called a discovery—that every paper reports a clear and definitive fact. As with all scientific disciplines, this is far from the truth. Most research findings in psychology are probabilistic rather than deterministic: conventional statistical tests talk to us in terms of probabilities rather than proofs. This in turn means that no single study and no one paper can lay claim to a discovery. Discovery depends wholly and without exception on the extent to which the original results can be repeated or *replicated* by other scientists, and not just once but over and over again. For example, it would not be enough to report only once that a particular cognitive therapy was effective at reducing depression; the result would need to be repeated many times in different groups of patients, and by different groups of researchers, for it be widely adopted as a public health intervention. Once a result has been replicated a satisfactory number of times using the same experimental method, it can then be considered *replicable* and, in combination with other replicable evidence, can contribute meaningfully to the theoretical or applied framework in which it resides. Over time, this mass accumulation of replicable evidence within different fields can allow theories to become accepted through consensus and in some cases can even become laws.

In science, prioritizing novelty hinders rather than helps discovery because it dismisses the value of direct (or close) replication. As we have seen, journals are the gatekeepers to an academic career, so if they value findings that are positive and novel, why would scientists ever attempt to replicate each other? Under a neophilic incentive structure, direct replication is discarded as boring, uncreative, and lacking in intellectual prowess.

Yet even in a research system dominated by positive bias and neophilia, psychologists have retained some realization that reproducibility matters. So, in place of unattractive direct replication, the community has reached for an alternative form of validation in which one experiment can be said to replicate the key concept or theme of another by following a different (novel) experimental method—a process known as *conceptual* replication. On its face, this redefinition of replication appears to satisfy the need to validate previous findings while also preserving novelty. Unfortunately, all it really does is introduce an entirely new and pernicious form of confirmation bias.

Replicating Concepts Instead of Experiments

In early 2012, a professor of psychology at Yale University named John Bargh launched a stinging public attack on a group of researchers who failed to replicate one of his previous findings.[20] The study in question, published by Bargh and colleagues in 1996, reported that priming participants unconsciously to think about concepts related to elderly people (e.g., words such as "retired," "wrinkle," and "old") caused them to walk more slowly when leaving the lab at the end of the experiment.[21] Based on these findings, Bargh claimed that people are remarkably susceptible to automatic effects of being primed by social constructs.

Bargh's paper was an instant hit and to date has been cited more than 3,800 times. Within social psychology it spawned a whole generation of research on social priming, which has since been applied in a variety of different contexts. Because of the impact the paper achieved, it would be reasonable to expect that the central finding must have been replicated many times and confirmed as being sound. Appearances, however, can be deceiving.

Several researchers had reported failures to replicate Bargh's original study, but few of these nonreplications have been published, owing to the fact that journals (and reviewers) disapprove of negative findings and often refuse to publish direct replications. One such attempted replication in 2008 by Hal Pashler and colleagues from the University of California San Diego was never published in an academic journal and instead resides at an online repository called PsychFileDrawer.[22] Despite more than doubling the sample size reported in the original study, Pashler and his team found no evidence of such priming effects—if anything they found the opposite result.

Does this mean Bargh was wrong? Not necessarily. As psychologist Dan Simons from the University of Illinois has noted, failing to replicate an effect does not necessarily mean the original finding was in error.[23] Nonreplications can emerge by chance, can be due to subtle changes in experimental methods between studies, or can be caused by the poor methodology of the researchers attempting the replication. Thus, nonreplications are themselves subject to the same tests of replicability as the studies they seek to replicate.

Nevertheless, the failed replication by Pashler and colleagues—themselves an experienced research team—raised a question mark over the status of Bargh's original study and hinted at the existence of an invisible file drawer

of unpublished failed replications. In 2012, another of these attempted replications came to light when Stéphane Doyen and colleagues from the University of Cambridge and Université Libre de Bruxelles also failed to replicate the elderly priming effect.[24] Their article appeared prominently in the peer-reviewed journal *PLOS ONE*, one of the few outlets worldwide that explicitly renounces neophilia and publication bias. The ethos of *PLOS ONE* is to publish any methodologically sound scientific research, regardless of subjective judgments as to its perceived importance or originality. In their study, Doyen and colleagues not only failed to replicate Bargh's original finding but also provided an alternative explanation for the original effect—rather than being due to a priming manipulation, it was the experimenters themselves who unwittingly induced the participants to walk more slowly by behaving differently or even revealing the hypothesis.

The response from Bargh was swift and contemptuous. In a highly publicized blogpost at psychologytoday.com entitled "Nothing in Their Heads,"[25] he attacked not only Doyen and colleagues as "incompetent or ill-informed," but also science writer Ed Yong (who covered the story)[26] for engaging in "superficial online science journalism," and *PLOS ONE* as a journal that "quite obviously does not receive the usual high scientific journal standards of peer-review scrutiny." Amid a widespread backlash against Bargh, his blogpost was swiftly (and silently) deleted but not before igniting a fierce debate about the reliability of social priming research and the status of replication in psychology more generally.

Doyen's article, and the response it generated, didn't just question the authenticity of the elderly priming effect; it also exposed a crucial disagreement about the definition of replication. Some psychologists, including Bargh himself, claimed that the original 1996 study had been replicated at length, while others claimed that it had *never* been replicated. How is this possible?

The answer, it turned out, was that different researchers were defining replication differently. Those who argued that the elderly priming effect had never been replicated were referring to *direct* replications: studies that repeat the method of a previous experiment as exactly as possible in order to reproduce the finding. At the time of writing, Bargh's central finding has been directly replicated just twice, and in each case with only partial success. In the first attempt, published six years after the original study,[27] the researchers showed the same effect but only in a subgroup of participants who scored

high on self-consciousness. In the second attempt, published another four years later, a different group of authors showed that priming elderly concepts slowed walking only in participants who held positive attitudes about elderly people; those who harbored negative attitudes showed the opposite effect.[28] Whether these partial replications are themselves replicable is unknown, but as we will see in chapter 2, hidden flexibility in the choices researchers make when analyzing their data (particularly concerning subgroup analyses) can produce spurious differences where none truly exist.

In contrast, those who argued that the elderly priming effect had been replicated many times were referring to the notion of "conceptual replication": the idea that the *principle* of unconscious social priming demonstrated in Bargh's 1996 study has been extended and applied in many different contexts. In a later blog post at psychologytoday.com called "Priming Effects Replicate Just Fine, Thanks," Bargh referred to some of these conceptual replications in variety of social behaviors, including attitudes and stereotypes unrelated to the elderly.[29]

The logic of "conceptual replication" is that if an experiment shows evidence for a particular phenomenon, you can replicate it by using a different method that the experimenter believes measures the same class of phenomenon. Psychologist Rolf Zwaan argues that conceptual replication has a legitimate role in psychology (and indeed all sciences) to test the extent to which particular phenomena depend on specific laboratory conditions, and to determine whether they can be generalized to new contexts.[30] The current academic culture, however, has gone further than merely valuing conceptual replication—it has allowed it to usurp direct replication. As much as we all agree about the importance of converging evidence, should we be seeking it out at the expense of knowing whether the phenomenon being generalized exists in the first place?

A reliance on conceptual replication is dangerous for three reasons.[31] The first is the problem of subjectivity. A conceptual replication can hold only if the different methods used in two different studies are measuring the same phenomenon. For this to be the case, some evidence must exist that they are. Even if we meet this standard, this raises the question of how similar the methods must be for a study to qualify as being conceptually replicated. Who decides and by what criteria?

The second problem is that a reliance on conceptual replications risks findings becoming *unreplicated* in the future. To illustrate how this could

happen, suppose we have three researchers, Smith, Jones, and Brown, who publish three scientific papers in sequence. Smith publishes the first paper, showing evidence for a particular phenomenon. Jones then uses a different method to show evidence for a phenomenon that appears similar to the one that Smith discovered. The psychological community decide that the similarity crosses some subjective threshold and so conclude that Jones "conceptually replicates" Smith. Now enter Brown. Brown isn't convinced that Smith and Jones are measuring the same phenomenon and suspects they are in fact describing *different* phenomena. Brown obtains evidence suggesting that this is indeed the case. In this way, Smith's finding that was previously considered replicated by Jones now assumes the bizarre status of becoming *un*replicated.

Finally, conceptual replication fuels an obvious confirmation bias. When two studies draw similar conclusions using different methods, the second study can be said to conceptually replicate the first. But what if the second study draws a very different conclusion—would it be claimed to conceptually *falsify* the first study? Of course not. Believers of the original finding would immediately (and correctly) point to the multitude of differences in methodology to explain the different results. Conceptual replications thus force science down a one-way street in which it is possible to confirm but never disconfirm previous findings. Through a reliance on conceptual replication, psychology has found yet another way to become enslaved to confirmation bias.

Reinventing History

So far we have seen how confirmation bias influences psychological science in two ways: through the pressure to publish results that are novel and positive, and by ousting direct replication in favor of bias-prone conceptual replication. A third, and especially insidious, manifestation of confirmation bias can be found in the phenomenon of hindsight bias. Hindsight bias is a form of creeping determinism in which we fool ourselves (and others) into believing that an observation was expected even though it actually came as a surprise.

It may seem extraordinary that any scientific discipline should be vulnerable to a fallacy that attempts to reinvent history. Indeed, under the classic

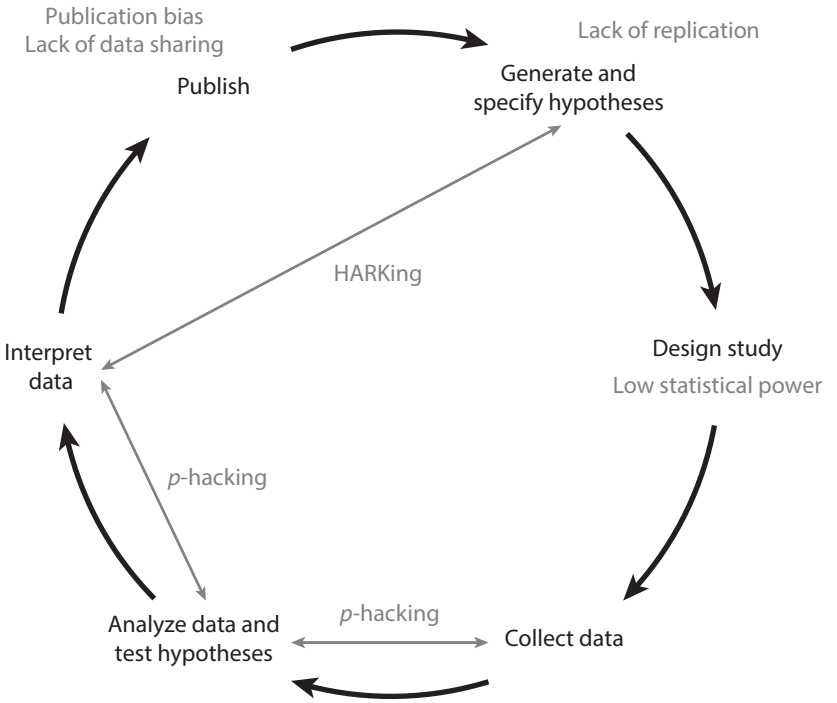

FIGURE 1.2. The hypothetico-deductive model of the scientific method is compromised by a range of questionable research practices. **Lack of replication** impedes the elimination of false discoveries and weakens the evidence base underpinning theory. **Low statistical power** (to be discussed in chapter 3) increases the chances of missing true discoveries and reduces the probability that obtained positive effects are real. Exploiting researcher degrees of freedom (***p*-hacking**—to be discussed in chapter 2) manifests in two general forms: collecting data until analyses return statistically significant effects, and selectively reporting analyses that reveal desirable outcomes. **HARKing**, or Hypothesizing After Results are Known, involves generating a hypothesis from the data and then presenting it as a priori. **Publication bias** occurs when journals reject manuscripts on the basis that they report negative or otherwise unattractive findings. Finally, **lack of data sharing** (to be discussed in chapter 4) prevents detailed meta-analysis and hinders the detection of data fabrication.

hypothetico-deductive (H-D) model of the scientific method, the research process is supposed to be protected against such bias (see figure 1.2). According to the H-D method, to which psychology at least nominally adheres, a scientist begins by formulating a hypothesis that addresses some aspect of a relevant theory. With the hypothesis decided, the scientist then conducts an experiment and allows the data to determine whether or not the

hypothesis was supported. This outcome then feeds into revision (and possible rejection) of the theory, stimulating an iterative cycle of hypothesis generation, hypothesis testing, and theoretical advance. A central feature of the H-D method is that the hypothesis is decided *before* the scientist collects and analyzes the data. By separating in time the prediction (hypothesis) from the estimate of reality (data), this method is designed to protect scientists from their own hindsight bias.

Unfortunately, much psychological research seems to pay little heed to this aspect of the scientific method. Since the hypothesis of an experiment is only rarely published in advance, researchers can covertly alter their predictions after the data have been analyzed in the interests of narrative flair. In psychology this practice is referred to as Hypothesizing After Results are Known (HARKing), a term coined in 1998 by psychologist Norbert Kerr.[32] HARKing is a form of academic deception in which the experimental hypothesis (H1) of a study is altered after analyzing the data in order to pretend that the authors predicted results that, in reality, were unexpected. By engaging in HARKing, authors are able to present results that seem neat and consistent with (at least some) existing research or their own previously published findings. This flexibility allows the research community to produce the kind of clean and confirmatory papers that psychology journals prefer while also maintaining the illusion that the research is hypothesis driven and thus consistent with the H-D method.

HARKing can take many forms, but one simple approach involves reversing the predictions after inspecting the data. Suppose that a researcher formulates the hypothesis that, based on the associations we form across our lifetime between the color red and various behavioral acts of stopping (e.g., traffic lights; stop signs; hazard signs), people should become more cautious in a gambling task when the stimuli used are red rather than white. After running the experiment, however, the researcher finds the opposite result: people gambled *more* when exposed to red stimuli. According to the H-D method, the correct approach here would be to report that the hypothesis was unsupported, admitting that additional experiments may be required to understand how this unexpected result arose and its theoretical implications. However, the researcher realizes that this conclusion may be difficult to publish without conducting those additional experiments, and he or she also knows that nobody reviewing the paper would be aware that

the original hypothesis was unsupported. So, to create a more compelling narrative, the researcher returns to the literature and searches for studies suggesting that being exposed to the color red can lead people to "see red," losing control and becoming *more* impulsive. Armed with a small number of cherry-picked findings, the researcher ignores the original (better grounded) rationale and rewrites the hypothesis to predict that people will actually gamble *more* when exposed to red stimuli. In the final published paper, the introduction section is written with this post hoc hypothesis presented as a priori.

Just how prevalent is this kind of HARKing? Norbert Kerr's survey of 156 psychologists in 1998 suggested that about 40 percent of respondents had observed HARKing by other researchers; strikingly, the surveyed psychologists also suspected that HARKing was about 20 percent more prevalent than the classic H-D method.[33] A more recent survey of 2,155 psychologists by Leslie John and colleagues estimated the true prevalence rate to be as high as 90 percent despite a self-admission rate of just 35 percent.[34]

Remarkably, not all psychologists agree that HARKing is a problem. Nearly 25 years before suggesting the existence of precognition, Daryl Bem claimed that if data are "strong enough" then researchers are justified in "subordinating or even ignoring [their] original hypotheses."[35] In other words, Bem argued that it is legitimate to subvert the H-D method, and to do so covertly, in order to preserve the narrative structure of a scientific paper.

Norbert Kerr and others have objected to this point of view, as well they might. First and foremost, because HARKing relies on deception, it violates the fundamental ethical principle that research should be reported honestly and completely. Deliberate HARKing may therefore lie on the same continuum of malpractice as research fraud. Secondly, the act of deception in HARKing leads the reader to believe that an obtained finding was more expected, and hence more reliable, than it truly is—this, in turn, risks distorting the scientific record to place undue certainty in particular findings and theories. Finally, in cases where a post hoc hypothesis is pitted against an alternative account that the author already knows was unsupported, HARKing creates the illusion of competitive hypothesis testing. Since a HARKed hypothesis can, by definition, never be disconfirmed, this contrived scenario further exacerbates confirmation bias.

The Battle against Bias

If confirmation bias is so much a part of human nature then what hope can we have of defeating it in science? In an academic culture that prizes novel results that confirm our expectations, is there any real chance of reform? We have known about the various manifestations of bias in psychology since the 1950s—and have done little to counteract them—so it is easy to see why many psychologists are cynical about the prospect of change. However, the tide is turning. Chapter 8 will address the set of changes we must make—and are already launching—to protect psychological science against bias and the other "deadly sins" that have become part of our academic landscape. Some of these reforms are already bearing fruit.

Our starting point for any program of reform must be the acceptance that we can never completely eliminate confirmation bias—in Nietzsche's words we are *human, all too human*. Decades of psychological research shows how bias is woven into the fabric of cognition and, in many situations, operates unconsciously. So, rather than waging a fruitless war on our own nature, we would do better to accept imperfection and implement measures that protect the outcome of science as much as possible from our inherent flaws as human practitioners.

One such protection against bias is study preregistration. We will return to the details of preregistration in chapter 8, but for now it is useful to consider how publicly registering our research intentions *before* we collect data can help neutralize bias. Consider the three main manifestations of confirmation bias in psychology: publication bias, conceptual replication, and HARKing. In each case, a strong motivation for engaging in these practices is not to generate high-quality, replicable science, but to produce results that are publishable and perceived to be of interest to other scientists. Journals enforce publication bias because they believe that novel, positive results are more likely to indicate discoveries that their readers will want to see; by comparison, replications and negative findings are considered boring and relatively lacking in intellectual merit. To fit with the demands of journals, psychologists have thus replaced direct replication with conceptual replication, maintaining the comfortable but futile delusion that our science values replication while still satisfying the demands of novelty and originality. Finally, as we have seen, many researchers engage in HARKing because

they realize that failing to confirm their own hypothesis is regarded as a form of intellectual failure.

Study preregistration helps overcome these problems by changing the incentive structure to value "good science" over and above "good results." The essence of preregistration is that the study rationale, hypotheses, experimental methods, and analysis plan are stated publicly in advance of collecting data. When this process is undertaken through a peer-reviewed journal, it forces journal editors to make publishing decisions before results exist. This, in turn, prevents publication bias by ensuring that whether results are positive or negative, novel or familiar, groundbreaking or incremental, is irrelevant to whether the science will be published. Similarly, since authors will have stated their hypotheses in advance, preregistration prevents HARKing and ensures adherence to the H-D model of the scientific method. As we will see in chapter 2, preregistration also prevents researchers from cherry-picking results that they believe generate a desirable narrative.

In addition to study preregistration, bias can be reduced by reforming statistical practice. As discussed earlier, one reason negative findings are regarded as less interesting is our cultural reliance on null hypothesis significance testing (NHST). NHST can only ever tell us whether the null hypothesis is rejected, and never whether it is supported. Our reliance on this one-sided statistical approach inherently places greater weight on positive findings. However, by shifting to alternative Bayesian statistical methods, we can test all potential hypotheses (H0, H1 . . . Hn) fairly as legitimate possible outcomes. We will explore this alternative method in more detail in chapter 3.

As we take this journey it is crucial that individual scientists from every level feel empowered to promote reform without damaging their careers. Confirmation bias is closely allied with "groupthink"—a pernicious social phenomenon in which a consensus of behavior is mistaken for a convergence of informed evidence. The herd doesn't always make the most rational or intelligent decisions, and groupthink can stifle innovation and critical reflection. To ensure the future of psychological science, it is incumbent on us as psychologists to recognize and challenge our own biases.