CHAPTER 3

# The Nature of Bullshit

SO WHAT IS BULLSHIT, EXACTLY? PEOPLE SOMETIMES USE THE term the same way that little kids use the word "unfair": to describe anything they don't like. Used in this way, the term can apply to a close but correct call at home plate, or a parking ticket written two minutes after your meter expires. Other times the term applies to injustice: a corrupt process of granting government contracts, an unwarranted acquittal, or legislation denying people basic civil liberties.

Then there is language to soften bad news and more generally grease the wheels of social interaction. "I just loved your turnip lasagna. You absolutely must give me the recipe." "What a beautiful baby!" "It's not you, it's me." "Your haircut looks great." "Your call is important to us." "I had a lovely evening." These are often bullshit, but they're not really the kind of bullshit we're concerned with here.

Insincere promises and outright lies get a bit closer to the mark. "Honey, I'm going to have to work late again tonight"; "I didn't inhale"; "I'd love to give you a better deal if I could"; "I have read and agreed to the above terms and conditions"; "Read my lips: No new taxes." Still, we tend to think of these claims as outright lies rather than bullshit.

But lies are often most persuasive when dressed in superfluous details; these details come quite close to what we mean by "bullshit." A friend leaves you waiting for twenty-five minutes at the local café because he got distracted watching YouTube videos and left his house

half an hour late. "I'm sorry I'm late, traffic was terrible. You know that bridge down by Fifteenth? A bus had stalled right where it goes down to two lanes each way, and some idiot had rear-ended it, blocking the other lane. So by the time I got through all that mess . . ." The first sentence is a lie; the subsequent ones are bullshit.

Harry Frankfurt, the philosopher we introduced in the preface, refined this notion a bit further. He described bullshit as what people create when they try to impress you or persuade you, without any concern for whether what they are saying is true or false, correct or incorrect. Think about a high school English essay you wrote without actually reading the book, a wannabe modernist painter's description of his artistic vision, or a Silicon Valley tech bro co-opting a TED Talk invitation to launch his latest startup venue. The intention may be to mislead, but it need not be. Sometimes we are put on the spot and yet have nothing to say. The bullshit we produce under those circumstances is little more than the "filling of space with the inconsequential."

Bullshit can be total nonsense. Another philosopher to take up the issue of bullshit, G. A. Cohen, notes that a lot of bullshit—particularly of the academic variety—is meaningless and so cloaked in rhetoric and convoluted language that no one can even critique it. Thus for Cohen, bullshit is "unclarifiable unclarity." Not only is the bullshitter's prose unclear, but the ideas underlying it are so ill-formed that it cannot possibly be clarified. Cohen suggests a test for unclarity: If you can negate a sentence and its meaning doesn't change, it's bullshit. "Shakespeare's Prospero is ultimately the fulcrum of an epistemic tragedy, precisely because of his failure to embrace the hermeneutics of the transfinite."

What these notions of bullshit have in common is that the speaker aims to persuade or impress, rather than to lead the listener toward the truth. The speaker may do this with active obfuscation, or simply by talking nonsense to conceal the fact that he or she doesn't actually know anything about the subject at hand. Some authors distinguish between *persuasive bullshit* and *evasive bullshit*. The former aims to convey an exaggerated sense of competence or authority, while the latter avoids directly answering a question that the speaker would prefer not to address.

When someone is bullshitting, he and his audience are not allies in the process of communication. Rather, the speaker is aiming to manipulate the listener with rhetorical flair, superfluous detail, or statistical snake oil. For us:

> *Bullshit* involves language, statistical figures, data graphics, and other forms of presentation intended to persuade or impress an audience by distracting, overwhelming, or intimidating them with a blatant disregard for truth, logical coherence, or what information is actually being conveyed.

The key elements of this definition are that bullshit bears no allegiance to conveying the truth, and that the bullshitter attempts to conceal this fact behind some type of rhetorical veil. Sigmund Freud illustrated the concept about as well as one could imagine in a letter he wrote his fiancée, Martha Bernays, in 1884:

> So I gave my lecture yesterday. Despite a lack of preparation, I spoke quite well and without hesitation, which I ascribe to the cocaine I had taken beforehand. I told about my discoveries in brain anatomy, all very difficult things that the audience certainly didn't understand, but all that matters is that they get the impression that I understand it.

While he has not written about bullshit directly, the sociologist of science Bruno Latour has had a formative effect on our thinking about how people bullshit their audiences. Latour looks at the power dynamics between an author and a reader. In Latour's worldview, a primary objective of nonfiction authors is to appear authoritative. One good way to do this is to be correct, but that is neither necessary nor sufficient. Correct or not, an author can adopt a number of tactics to make her claims unassailable by her readers—who in turn strive not to be duped. For example, the author can line up a phalanx of allies by citing other writers who support her point, or whose work she builds upon. If you question me, she implies, you have to question all of us. She can also deploy sophisticated jargon. Jargon may facilitate technical communication within a field, but it also serves to exclude those who have not been initiated into the inner circle of a discipline.

## BULLSHIT AND BLACK BOXES

According to Latour, scientific claims are typically built upon the output of metaphorical "black boxes," which are difficult if not impossible for the reader to penetrate. These black boxes often involve the use of specialized and often expensive equipment and techniques that are time-consuming and unavailable, or are so broadly accepted that to question them represents a sort of scientific heresy.* If I were to write a paper claiming that specific genetic variants are associated with susceptibility to bullshit, a skeptic might reasonably argue my choice of sample population, the way I measure bullshit susceptibility, or the statistical method I use to quantify associations. But the biotechnology used to derive the DNA sequences from blood samples would typically be treated as a black box. In principle a skeptic could question this as well, but to do so she would be challenging the scientific establishment and, more important for our purposes, she would need access to advanced equipment and extensive technical expertise in molecular genetics.

Latour is not saying that these aspects of academic science make the entire enterprise bullshit, and neither are we. He is saying only that science is more than a dispassionate search for the truth, a theme to which we will return in chapter 9. The important thing about Latour's black box idea is that we see a powerful analogy here to what speakers do when they bullshit effectively. Outright lies are often straightforward to catch and refute. But effective bullshit is difficult to fact-check. Bullshit can act like one of Latour's black boxes, shielding a claim from further investigation.

Suppose a friend tells you, "You know, on average, cat people earn higher salaries than dog people." It's easy to call bullshit on that statement when it stands by itself. And when you do so, perhaps your friend will simply laugh and admit, "Yeah, I made that up."

* For Latour, aspects of technology or experimental procedure are "black-boxed" when they are so thoroughly accepted by the relevant scientific community that they become generally agreed upon standards and are no longer open to question. In Latour's *Pandora's Hope: Essays on the Reality of Science*, he explains: "When a machine runs efficiently, when a matter of fact is settled, one need focus only on its inputs and outputs and not on its internal complexity. Thus, paradoxically, the more science and technology succeed, the more opaque and obscure they become."

Here our analogy begins to break down somewhat. For the purposes of thinking about quantitative bullshit, consider a technique to be black-boxed once it requires expertise beyond that of the typical reader, irrespective of whether it is accepted by the scientific community.

But suppose instead she doubles down and starts filling out—or making up—details to support her claim. "No, really, it's true. I saw this TED Talk about it. They explained how cat owners value independence whereas dog owners value loyalty. People who value independence are more likely to have NVT . . . no . . . NVS . . . I can't remember, but some kind of personality. And that makes them better able to rise in the workplace."

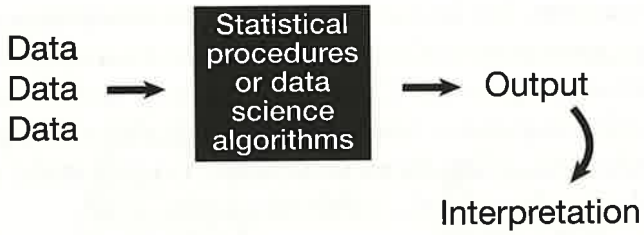
This is full-on bullshit, and it functions like one of Latour's black boxes. If you want to dispute your friend's claims, you now have substantial work to do. This is where lies and bullshit come together: In our view, a lie becomes bullshit when the speaker attempts to conceal it using various rhetorical artifices.

Now imagine that she points you to a research study that makes this claim. Suppose you track down the study, and read something like the following:

> We observe a statistically significant difference in cat- and dog-lovers' earnings, based on an ANCOVA using log-transformed earnings data ($F = 3.86$).

If you don't have a professional background in statistics, you've just slammed head-on into a particularly opaque black box. You probably don't know what an ANCOVA is or what the $F$ value means or what a log transformation is or why someone would use it. If you do know some of these things, you still probably don't remember all of the details. We, the authors, use statistics on a daily basis, but we still have to look up this sort of stuff all the time. As a result, you can't unpack the black box; you can't go into the details of the analysis in order to pick apart possible problems. Unless you're a data scientist, and probably even then, you run into the same kind of problem you encounter when you read about a paper that uses the newest ResNet algorithm to reveal differences in the facial features of dog and cat owners. Whether or not this is intentional on the part of the author, this kind of black box shields the claim against scrutiny.

But it doesn't need to. The central theme of this book is that you usually don't have to open the analytic black box in order to call bullshit on the claims that come out of it. Any black box used to generate bullshit has to take in data and spit results out, like the diagram on page 43.




Most often, bullshit arises either because there are biases in the data that get fed into the black box, or because there are obvious problems with the results that come out. Occasionally the technical details of the black box matter, but in our experience such cases are uncommon. This is fortunate, because you don't need a lot of technical expertise to spot problems with the data or results. You just need to think clearly and practice spotting the sort of thing that can go wrong. In the pages that follow, we will show you how to do precisely this.

**RETURNING TO OUR CAT** and dog example, instead of digging into the details of the statistical analysis, you might ask how the samples were collected. Maybe the study looked at pet ownership and merged data from people in living in New York City—where salaries are high and keeping a dog is difficult—with data from upstate New York, where salaries are lower and dogs are far more practical. Maybe dog-lover salaries are assumed to be the US average whereas cat-lover salaries were solicited from visitors to a website for startup founders who cohabitate with cats.

If the data that go into the analysis are flawed, *the specific technical details of the analysis don't matter*. One can obtain stupid results from bad data without any statistical trickery. And this is often how bullshit arguments are created, deliberately or otherwise. To catch this sort of bullshit, you don't have to unpack the black box. All you have to do is think carefully about the data that went into the black box and the results that came out. Are the data unbiased, reasonable, and relevant to the problem at hand? Do the results pass basic plausibility checks? Do they support whatever conclusions are drawn?

Being able to spot bullshit based on data is a critical skill. Decades ago, fancy language and superfluous detail might have served a bullshitter's needs. Today, we are accustomed to receiving information in

quantitative form, but hesitant to question that information once we receive it. Quantitative evidence generally seems to carry more weight than qualitative arguments. This weight is largely undeserved—only modest skill is required to construct specious quantitative arguments. But we defer to such arguments nonetheless. Consequently, numbers offer the biggest bang for the bullshitting buck.

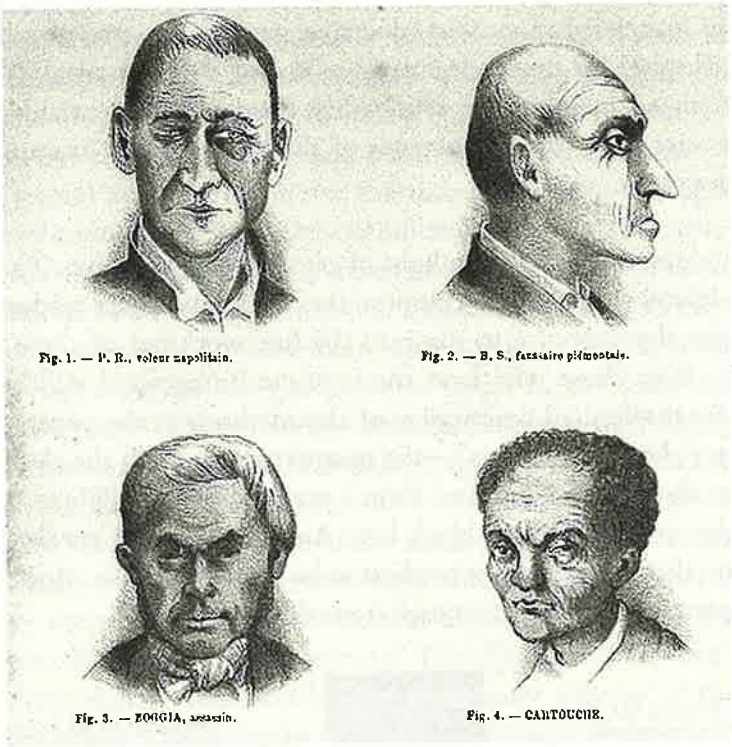
## CRIMINAL MACHINE LEARNING

Let's take a moment to illustrate how we can call bullshit on a fancy algorithm without ever delving into the details of how it works, i.e., without opening the black box.

In late 2016, engineering researchers Xiaolin Wu and Xi Zhang submitted an article titled "Automated Inference on Criminality Using Face Images" to a widely used online repository of research papers known as the arXiv. In their article, Wu and Zhang explore the use of machine learning to detect features of the human face that are associated with "criminality." They claim that their algorithm can use simple headshots to distinguish criminals from noncriminals with high accuracy. If this strikes you as frighteningly close to Philip K. Dick's Precrime police in *Minority Report,* and to other dystopian science fiction, you're not alone. The media thought so too. A number of technology-focused press outlets picked up on the story and explored the algorithm's ethical implications. If an algorithm could really detect criminality from the structure of a person's face, we would face an enormous ethical challenge. How would we have to adjust our notions of innocence and guilt if we could identify people as criminals even before they committed a crime?

The notion that criminals are betrayed by their physiognomy is not a new one. In the nineteenth century, an Italian doctor named Cesare Lombroso studied the anatomy of hundreds of criminals. His aim was to develop a scientific theory of criminality. He proposed that people were born to be criminals or to be upstanding citizens. Born criminals, he postulated, exhibit different psychological drives and physical features. Lombroso saw these features as hearkening back to our subhuman evolutionary past. He was particularly interested in what could be learned from faces. In his view, the shape of the jaw, the slope of the forehead, the size of the eyes, and the structure of the ear

all contained important clues about a man's moral composition. Below, a figure from Cesare Lombroso's 1876 book *Criminal Man*.

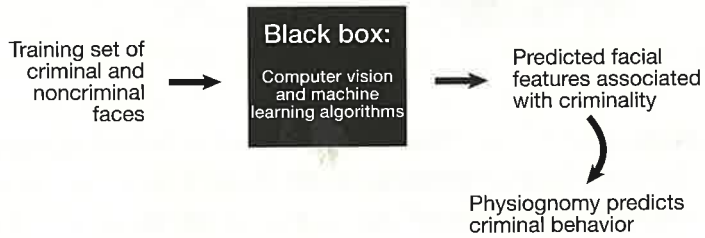



Lombroso was wrong. None of his theories linking anatomy to moral character have a sound scientific basis. His ideas—many of which wrapped racist ideas of the time in a thin veneer of scientific language—were debunked in the first half of the twentieth century and disappeared from the field of criminology.

But in the 2016 arXiv paper, Wu and Zhang revisit Lombroso's program. Essentially, they aim to determine whether advanced computer vision can reveal subtle cues and patterns that Lombroso and his followers might have missed. To test this hypothesis, the authors use machine learning algorithms to determine what features of the human face are associated with "criminality." Wu and Zhang claim that based on a simple headshot, their programs can distinguish criminal from noncriminal faces with nearly 90 percent accuracy. Moreover, they argue that their computer algorithms are free from the myriad biases and prejudices that cloud human judgment:

> Unlike a human examiner/judge, a computer vision algorithm or classifier has absolutely no subjective baggages [sic], having no emotions, no biases whatsoever due to past experience, race, religion, political doctrine, gender, age, etc., no mental fatigue, no preconditioning of a bad sleep or meal. The automated inference on criminality eliminates the variable of meta-accuracy (the competence of the human judge/examiner) all together.

Let's look at all of this in light of our black box schema. The machine learning algorithms compose the black box. Most readers will not have the expertise to dig into the fine workings of these algorithms. Even those who have the requisite background will be stymied by the limited description of the methods in the paper. Then there are the "training sets"—the images used to teach the algorithm how to distinguish a criminal from a noncriminal face. These are the data that are fed into the black box. And finally, there are the facial features that the algorithm predicts to be associated with criminality. These are the results that emerge from the black box.

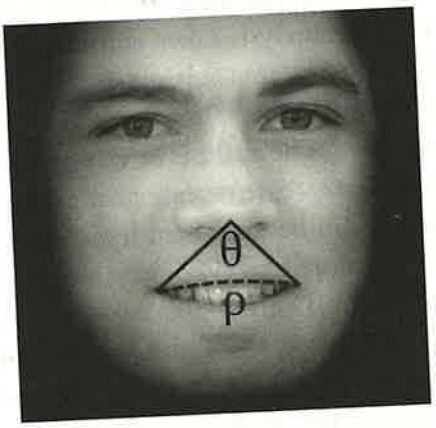



To understand the paper, we need to look at the training set. A machine learning algorithm can be only as good as the training data that it is provided. Wu and Zhang collected over 1,800 photos of Chinese men aged 18 to 55, with no distinguishing facial hair, scars, or tattoos. About 1,100 of these were noncriminals. Their photographs were taken from a variety of sources on the World Wide Web, including job-based social networking sites and staff listings from professional firms. Just over 700 of the subjects were convicted criminals. Their photos were provided by police departments and taken from official pieces of identification, not from mugshots.

We see two massive problems. The first is that the images of non-criminals were selected to cast the individuals in a positive light. By contrast, the images from the set of criminals are official ID photographs. While it is unclear exactly what this means, it's safe to guess that these have been selected neither by the person depicted, nor with the aim of casting him in a favorable light. Thank goodness no one judges our characters based upon our driver's license photos!

A second source of bias is that the authors are using photographs of *convicted* criminals. If there are facial differences between the two groups, we won't know whether these differences are associated with committing crimes or with being convicted. Indeed, appearance seems to matter for convictions. A recent study reports that in the US, unattractive individuals are more likely to be found guilty in jury trials than their attractive peers.* Thus while the authors claim that their algorithm is free of human biases, it could be picking up on *nothing but* these biases.

Having identified some potential problems with the data that go into the black box, we turn to the black box's output. As we mentioned, the authors find that their algorithm can classify criminal faces within their data set with 90 percent accuracy. What are the facial features that it uses to discriminate? The algorithm finds that criminals have shorter distances between the inner corners of the eyes, smaller angles $\theta$ between the nose and the corners of the mouth, and higher curvature $\rho$ to the upper lip.




* While the Chinese criminal system is structured differently from that of the US and jury trials are rarer, the judges and occasional jurors in Chinese trials may suffer from similar biases.

Why would this possibly be?

There's a glaringly obvious explanation for the nose-mouth angle and the lip curvature. As one smiles, the corners of the mouth spread out and the upper lip straightens. Try it yourself in the mirror.

If you look at the original research paper, you can see six example images from the training set. The criminals are frowning or scowling. The noncriminals are faintly smiling. Now we have an alternative—and far more plausible—hypothesis for the authors' findings. There are not important differences in facial structure between criminals and noncriminals. Rather, noncriminals are smiling in their professional headshots, whereas criminals are not smiling in their government ID photographs. It appears that the authors have confused innate *facial features* with labile *facial expressions*. If so, their claims about detecting criminality are bullshit. They have not invented a criminality detector; they have invented a smile detector.

We can see further evidence of this in the output of their black box. To illustrate the purported facial differences between criminals and noncriminals, the authors produced composite images by aggregating features from each group. The criminal composites are frowning, whereas the noncriminal composites are smiling. This supports our hypothesis that the machine learning algorithm is picking up on situation-dependent facial expressions (whether a person is smiling or not) rather than underlying facial structure.

You might find yourself thinking that even without opening the black box, this type of analysis takes considerable time and focus. That's true. But fortunately, some claims should more readily trigger our bullshit detectors than others. In particular, *extraordinary claims require extraordinary evidence*. The authors of this paper make the extraordinary claim that facial structure reveals criminal tendencies. Here we see that their findings can be explained by a much more reasonable hypothesis: People are more likely to be smiling in professional headshots than in government ID photographs.

Notice that we established all of this without opening the black box. We didn't have to look at the details of the machine learning algorithms at all, because the problem didn't arise there. A machine learning algorithm is only as good as its training data, and these training data are fundamentally flawed. As is often the case, one does not need technical expertise in machine learning to call bullshit. A non-

specialist can do so by thinking carefully about what any generic learning system would conclude from the same data. The algorithm in this paper is not picking up some underlying physical structures associated with criminality. And it doesn't look like we have to worry about the ethical minefield of precrime just yet.