AI SNAKE OIL

Chapter 1

INTRODUCTION

IMAGINE AN ALTERNATE universe in which people don't have words for different forms of transportation—only the collective noun "vehicle." They use that word to refer to cars, buses, bikes, spacecraft, and all other ways of getting from place A to place B. Conversations in this world are confusing. There are furious debates about whether or not vehicles are environmentally friendly, even though no one realizes that one side of the debate is talking about bikes and the other side is talking about trucks. There is a breakthrough in rocketry, but the media focuses on how vehicles have gotten faster—so people call their car dealer (oops, vehicle dealer) to ask when faster models will be available. Meanwhile, fraudsters have capitalized on the fact that consumers don't know what to believe when it comes to vehicle technology, so scams are rampant in the vehicle sector.

Now replace the word "vehicle" with "artificial intelligence," and we have a pretty good description of the world we live in.

Artificial intelligence, AI for short, is an umbrella term for a set of loosely related technologies. ChatGPT has little in common with, say, software that banks use to evaluate loan applicants. Both are referred to as AI, but in all the ways that matter—how

they work, what they're used for and by whom, and how they fail—they couldn't be more different.

Chatbots, as well as image generators like Dall-E, Stable Diffusion, and Midjourney, fall under the banner of what's called generative AI. Generative AI can generate many types of content in seconds: chatbots generate often-realistic answers to human prompts, and image generators produce photorealistic images matching almost any description, say "a cow in a kitchen wearing a pink sweater." Other apps can generate speech or even music.

Generative AI technology has been rapidly advancing, its progress genuine and remarkable. But as a product, it is still immature, unreliable, and prone to misuse. At the same time, its popularization has been accompanied by hype, fear, and misinformation.

In contrast to generative AI is predictive AI, which makes predictions about the future in order to guide decision-making in the present. In policing, AI might predict "How many crimes will occur tomorrow in this area?" In inventory management, "How likely is this piece of machinery to fail in the next month?" In hiring, "How well will this candidate perform if hired for this job?"

Predictive AI is currently used by both companies and governments, but that doesn't mean it works. It's hard to predict the future, and AI doesn't change this fact. Sure, AI can be used to pore over data to identify broad statistical patterns—for instance, people who have jobs are more likely to pay back loans—and that can be useful. The problem is that predictive AI is often sold as far more than that, and it is used to make decisions about people's lives and careers. It is in this arena that most AI snake oil is concentrated.

AI snake oil is AI that does not and cannot work as advertised. Since AI refers to a vast array of technologies and applications,

most people cannot yet fluently distinguish which types of AI are actually capable of functioning as promised and which types are simply snake oil. This is a major societal problem: we need to be able to separate the wheat from the chaff if we are to make full use of what AI has to offer while protecting ourselves from its possible harms, harms which in many cases are already occurring.

This book is a guide to identifying AI snake oil and AI hype. In it, we'll give you essential vocabulary to tease apart generative AI, predictive AI, and other types of AI. We'll share commonsense ways of assessing whether or not a purported advance is plausible. This will make you read news about AI much more skeptically and with an eye toward details that often get buried. A deeper understanding of AI will both satisfy your scientific curiosity and translate into practical ideas on how to use—and when not to use—AI in your life and career. And we will make the argument that predictive AI not only does not work today but will likely never work, because of the inherent difficulties in predicting human behavior. Finally, we hope that this book will get you thinking about your own responsibilities—and opportunities for change—with respect to the harmful implications of these tools.

The Dawn of AI as a Consumer Product

Generative AI burst into public view with the release of ChatGPT in November 2022. Ironically, the developer OpenAI released it without any fanfare and internally thought of it as a "research preview."[1] But it went viral overnight as people shared amusing and amazing examples of its capabilities. For example, software developer Thomas Ptacek asked it to write a biblical verse in the style of the King James Bible, explaining how to

remove a peanut butter sandwich from a VCR.[2] This is what ChatGPT spat out:

> And it came to pass that a man was troubled by a peanut butter sandwich, for it had been placed within his VCR, and he knew not how to remove it.
>
> And he cried out to the Lord, saying, "Oh, Lord, how can I remove this sandwich from my VCR, for it is stuck fast and will not budge?"
>
> And the Lord spoke unto him, saying, "Fear not, my child, for I shall guide thy hand and show thee the way. Take thy butter knife, and carefully insert it between the sandwich and the VCR, and gently pry them apart. And with patience and perseverance, the sandwich shall be removed, and thy VCR shall be saved."
>
> And the man did as the Lord commanded, and lo and behold, the sandwich was removed from the VCR, and the man was saved.
>
> And the Lord said, "Verily I say unto thee, seek not to put thy peanut butter sandwiches in thy VCR, for it is not a suitable place for such things. Rather, keep thy sandwiches in thy refrigerator or on thy plate, where they belong."
>
> And the man heeded the Lord's words, and from that day forth, he kept his sandwiches in their proper place, and was saved from trouble and woe. Amen.

Two months later, the app reportedly had over one hundred million users.[3] OpenAI was caught off guard by the explosion of interest and hadn't even procured enough computing power to handle the traffic that it generated.

Computer programmers soon started using it, because it turned out that ChatGPT was pretty good at generating snippets of programming code given only a description of what the

code should do. In fact, programmers had been using a previous product called GitHub Copilot that used similar technology, but their use of AI accelerated after ChatGPT's release. It decreased the time it took to create apps. Even nonprogrammers could now build some simple apps.

Microsoft soon licensed the technology from OpenAI, and its Bing search engine rolled out a chatbot that could answer questions based on search results. Google had built its own chatbot in 2021 but had not yet released it or integrated it into its products.[4] Bing's move was seen as an existential threat to Google, and Google hurriedly announced its own search chatbot called Bard (later renamed Gemini).

That's when things started to go wrong. In the promotional video for Bard, the bot said that the James Webb Space Telescope took the first picture of a planet outside the solar system. An astrophysicist pointed out that this was wrong.[5] Apparently Google couldn't get even a cherry-picked example right. Its market value instantly took a hundred-*billion*-dollar dip. That's because investors were spooked by the prospect of a search engine that would get much worse at answering simple factual queries if Google were to integrate Bard into search, as it had promised.[6]

Google's embarrassment, while expensive, was only a ripple that portended the wave of problems that arose from chatbots' difficulties with factual information. Their weakness is a consequence of the way they are built. They learn statistical patterns from their training data—which comes largely from the web—and then generate remixed text based on those patterns. But they don't necessarily remember what's in their training data. We'll dive into this in chapter 4.

Misuse of the technology is rampant. News websites have been caught publishing error-filled AI-generated stories on

important topics such as financial advice, and then refusing to stop using the technology even after the errors came to light.[7] Amazon is overrun with AI-generated books, including a few mushroom foraging guides, where errors can be fatal if a reader trusts the book.[8]

It's easy to look at all the flaws and misuses of chatbots and conclude that the world has gone mad for being so gaga about a technology that is so failure prone. But that conclusion would be too simplistic.

We think most knowledge industries can benefit from chatbots in some way. We use them ourselves for research assistance, for tasks ranging from mundane ones such as formatting citations correctly, to things we wouldn't otherwise be able to do such as understanding a jargon-filled paper in a research area we aren't familiar with.

The catch is that it takes effort and practice to use chatbots while avoiding their ever-present pitfalls. But *inappropriate* uses are much easier, because someone trying to make a quick buck, say by selling an AI-generated book, doesn't often care if the contents are garbage. That's what makes chatbots so conducive to misuse.

There are thornier questions about power. Suppose web search companies replace their traditional list of ten links with AI-generated ready answers. Even assuming that accuracy problems are fixed, the result is basically a machine for rewriting content found on other websites and passing it off as original, without having to send traffic or revenue to those websites. If search engines simply presented others' content as their own, they would run afoul of copyright law. But AI-generated answers seem to skirt this issue, although there are many lawsuits seeking to change this as of 2024.[9]

AI Shakes Up Entertainment

Another generative AI technology that has captivated people is text-to-image generation. In mid-2023, it was estimated that over a billion images had been created using Dall-E 2 by OpenAI, Firefly by Adobe, and Midjourney (by a company of the same name).[10] Another widely used image generator is Stable Diffusion by Stability AI, which is openly available, meaning that anyone can modify it to their liking. Stable Diffusion–based tools have been *downloaded* over two hundred million times. Since users run it on their own devices, there is no central tally of how many images have been generated using it, but it is likely to be several billion.

Image generators have enabled a deluge of entertainment.[11] Unlike traditional entertainment, these images are endlessly customizable to each user's interests. Some people delight in fantastic landscapes or cityscapes. Others enjoy images of historical figures in modern situations, or famous people doing things they wouldn't normally do, such as the Pope wearing a puffer jacket, dubbed "Balenciaga Pope." Fake trailers for various movies such as *Star Wars* in the highly recognizable style of Wes Anderson—symmetrical framing, pastel colors, whimsical sets—have proven popular.

It's not only hobbyists who are excited about image generators: entertainment apps are big business. Video game companies have created in-game characters that players can have a natural conversation with.[12] Many photo editing apps now have generative AI functionality. So, for example, you can ask such an app to add balloons to a picture of a birthday party.

AI was a major point of contention in the 2023 Hollywood strikes.[13] Actors worried that studios would be able to use

existing footage of them to train AI tools capable of generating new videos based on a script—videos that looked like they featured the real actors whose images and videos the AI tools were trained on. In other words, studios would be able to capitalize on actors' likenesses and past labor in perpetuity, but without compensation.

While the strikes have ended, the underlying tensions between labor and capital are sure to resurface, especially as the technology advances.[14] Many companies are working on text-to-video generators, while others are working on automating script writing. The end result might not be as artistically complex or valuable, but that might not matter to studios looking to crank out a summer blockbuster.

In the long run, we think that a combination of technology and law can alleviate most of the problems we've described, as well as amplify the benefits. For example, there are many promising technical ideas to make chatbots less likely to fabricate information, while regulation can curb intentional misuses. But in the short term, adjusting to a world with generative AI is proving to be painful, as these tools are highly capable but unreliable. It's as if everyone in the world has been given the equivalent of a free buzzsaw.

It will take work to integrate AI appropriately into our lives. A good example is what's happening in schools and colleges, given that AI can generate essays and pass college exams. Let's be clear—AI is no threat to education, any more than the introduction of the calculator was.[15] With the right oversight, it can be a valuable learning tool. But to get there, teachers will have to overhaul their curricula, their teaching strategies, and their exams. At a well-funded institution such as Princeton, where we teach, this is an opportunity rather than a challenge. In fact, we encourage our students to use AI. But many others have

been left scrambling as ChatGPT suddenly put a potential cheating tool in the hands of millions of students.

Will society be left perpetually reacting to new developments in generative AI? Or do we have the collective will to make structural changes that would allow us to spread out the highly uneven benefits and costs of new innovations, whatever they may be?

Predictive AI: An Extraordinary Claim That Requires Extraordinary Evidence

Generative AI creates many social costs and risks, especially in the short term. But we're cautiously optimistic about the potential of this type of AI to make people's lives better in the long run. Predictive AI is a different story.

In the last few years, applications of predictive AI to predict social outcomes have proliferated. Developers of these applications claim to be able to predict future outcomes about people, such as whether a defendant would go on to commit a future crime or whether a job applicant would do well at a job. In contrast to generative AI, predictive AI often does not work at all.[16]

People in the United States over the age of sixty-five are eligible to enroll in Medicare, a state-subsidized health insurance plan. To cut costs, Medicare providers have started using AI to predict how much time a patient will need to spend in a hospital.[17] These estimates are often incorrect. In one case, an eighty-five-year-old was evaluated as being ready to leave in seventeen days. But when the seventeen days passed, she was still in severe pain, and couldn't even push a walker without help. Still, based on the AI assessment, her insurance payments stopped. In cases like this, AI technology is often deployed with sensible

intentions. For example, without predictive AI, nursing homes would be logically incentivized to house patients forever. But in many cases, the goals of the system as well as how it's deployed change over time. One can easily imagine how Medicare providers' use of AI may have started as a way to create a modicum of accountability for nursing homes, but then morphed into a way to squeeze pennies out of the system regardless of the human cost.

Similar stories are prevalent across domains. In hiring, many AI companies claim to be able to judge how warm, open, or kind someone is based on their body language, speech patterns, and other superficial features in a thirty-second video clip. Does this really work? And do these judgments actually predict job performance? Unfortunately, the companies making these claims have failed to release any verifiable evidence that their products are effective. And we have lots of evidence to the contrary, showing that it is extremely hard to predict individuals' life outcomes, as we'll see in chapter 3.

In 2013, Allstate, an insurance company, wanted to use predictive AI to determine insurance rates in the U.S. state of Maryland—so that the company could make more money without losing too many customers. It resulted in a "suckers list"—a list of people whose insurance rates increased dramatically compared to their earlier rates.[18] Seniors over the age of sixty-two were drastically overrepresented in this list, an example of automated discrimination. It is possible that seniors are less likely to shop around for better prices and that AI picked up on that pattern in the data. The new pricing would likely increase revenue for the insurance company, yet it is morally reprehensible. While Maryland refused Allstate's proposal to use this AI tool on the grounds that it was

discriminatory, the company does use it in at least ten other U.S. states.*

If individuals object to AI in hiring, they can simply choose not to apply for jobs that engage AI to judge résumés. When predictive AI is used by governments, however, individuals have no choice but to comply. (That said, similar concerns also arise if many companies were to use the same AI to decide who to hire.) Many jurisdictions across the world use criminal risk prediction tools to decide whether defendants arrested for a crime should be released before their trial. Various biases of these systems have been documented: racial bias, gender bias, and ageism. But there's an even deeper problem: evidence suggests that these tools are only slightly more accurate than randomly guessing whether or not a defendant is "risky."

One reason for the low accuracy of these tools could be that data about certain important factors is not available. Consider three defendants who are identical in terms of the features that might be used by predictive AI to judge them: age, the number of past offenses, and the number of family members with criminal histories. These three defendants would be assigned the same risk score. However, in this example, one defendant is deeply remorseful, another has been wrongly arrested by the police, and the third is itching to finish the job. There is no good way for an AI tool to take these differences into account.

Another downside of predictive AI is that decision subjects have strong incentives to game the system. For example, AI was used to estimate how long the recipient of a kidney transplant

* Many of the examples in this book, like this one, are from the United States, simply because that is where we are based. However, the lessons we draw from these examples are intended to be broadly applicable.

would live after their transplant.[19] The logic was that people who had the longest to live after a transplant should be prioritized to receive kidneys. But the use of this prediction system would *disincentivize* patients with kidney issues to take care of their kidney function. That's because if their kidneys failed at a younger age, they would be more likely to get a transplant! Fortunately, the development of this system involved a deliberative process with participation by patients, doctors, and other stakeholders. So, the incentive misalignment was recognized and the use of predictive AI for kidney transplant matching was abandoned.

We'll see many more failures of predictive AI in chapters 2 and 3. Are things likely to improve over time? Unfortunately, we don't think so. Many of its flaws are inherent. For example, predictive AI is attractive because automation makes decision-making more efficient, but efficiency is exactly what results in a lack of accountability. We should be wary of predictive AI companies' claims unless they are accompanied by strong evidence.

Painting AI with a Single Brush Is Tempting but Flawed

Generative and predictive AI are two of the main types of AI. How many other types of AI are there? There is no way to answer that question, since there is no consensus about what is and isn't AI.

Here are three questions about how a computer system performs a task that may help us determine whether the label AI is appropriate. Each of these questions captures something about what we mean by AI, but none is a complete definition. First, does the task require creative effort or training for a human to perform? If yes, and the computer can perform it, it might be AI. This would explain why image generation, for example, qualifies

as AI. To produce an image, humans need a certain amount of skill and practice, perhaps in the creative arts or in graphic design. But even *recognizing* what's in an image, say a cat or a teapot—a task that is trivial and automatic for humans—proved daunting to automate until the 2010s, yet object recognition has generally been labeled AI. Clearly, comparison to human intelligence is not the only relevant criterion.

Second, we can ask: Was the behavior of the system directly specified in code by the developer, or did it indirectly emerge, say by learning from examples or searching through a database? If the system's behavior emerged indirectly, it might qualify as AI. Learning from examples is called machine learning, which is a form of AI. This criterion helps explain why an insurance pricing formula, for example, might be considered AI if it was developed by having the computer analyze past claims data, but not if it was a direct result of an expert's knowledge, even if the actual rule was identical in both cases. Still, many manually programmed systems are nonetheless considered AI, such as some robot vacuum cleaners that avoid obstacles and walls.

A third criterion is whether the system makes decisions more or less autonomously and possesses some degree of flexibility and adaptability to the environment. If the answer is yes, the system might be considered AI. Autonomous driving is a good example—it is considered AI. But like the previous criteria, this criterion alone can't be considered a complete definition—we wouldn't call a traditional thermostat AI, one that contains no electronics. Its behavior rather arises from the simple principle of a metal expanding or contracting in response to changes in temperature and turning the flow of current on or off.

In the end, whether an application gets labeled AI is heavily influenced by historical usage, marketing, and other factors. We won't fret about the fact that there's no consistent definition.

That might seem surprising for a book about AI. But recall our overarching message: there's almost nothing one can say in one breath that applies to all types of AI. Most of our discussion in the book will be about specific types of AI, and as long as each type is clearly defined, we'll be on the same page.

There's a humorous AI definition that's worth mentioning, because it reveals an important point: "AI is whatever hasn't been done yet." In other words, once an application starts working reliably, it fades into the background and people take it for granted, so it's no longer thought of as AI. There are many examples: Robot vacuum cleaners like the Roomba. Autopilot in planes. Autocomplete on our phones. Handwriting recognition. Speech recognition. Spam filtering. Spell-check. Yes, there was a time when spell-check was considered a hard problem!

We think these tools are all wonderful. They quietly make our lives better. These are the kinds of AI we want more of. This book is about the types of AI that are problematic in some way, because you wouldn't want to read three hundred pages on the virtues of spell-check. But it's important to recognize that not all AI is problematic—far from it.

Some new AI technologies will hopefully one day come to be seen as mundane. Today, self-driving cars often make the news for accidents and fatalities.[20] But safe automated driving is ultimately a solvable problem, although one whose difficulty has repeatedly been underestimated. The bigger challenge for society might be the massive labor displacement that the technology will cause if it becomes widespread—millions of people drive trucks, taxis, or rideshare vehicles. Still, if the safety problem is solved and the necessary social and political adjustments are made, we may one day take self-driving cars for granted, like we do elevators today.

However, we think other types of AI, notably predictive AI, are unlikely to become normalized. Accurately predicting people's social behavior is not a solvable technology problem, and determining people's life chances on the basis of inherently faulty predictions will always be morally problematic.

For a more in-depth case study of why we must avoid sweeping generalizations about AI, consider facial recognition, an AI technology that has civil liberties advocates concerned. It has led to many false arrests in the United States—six, as we write this—all Black people. Should the use of facial recognition by police be discontinued because it is error prone and misidentifies Black people more often?

One fact that's easy to miss in this debate is that all the false arrests involved a cascading set of police failures, most of them human errors rather than technological. Robert Williams was arrested for shoplifting in part based on the testimony of a security contractor who wasn't even present at the time of the theft.[21] Randall Reid was arrested in Georgia for a shoplifting crime in Louisiana—a state he had never set foot in.[22] Porcha Woodruff was arrested based on a 2015 photo, despite the fact that a 2021 driver's license photo was available.[23] And so on.

Policing errors leading to the arrest of the wrong person happen every day, and will probably continue whether or not facial recognition is used.

Besides, police have made hundreds of thousands of facial recognition searches, so the error rate of the technology is minuscule.[24] In fact, the error rate dropped to 0.08 percent—a fifty-fold decrease between 2014 and 2020—according to studies by the National Institute of Standards and Technology.[25]

Facial recognition AI, if used correctly, tends to be accurate because there is little uncertainty or ambiguity in the task. Such AI is trained using vast databases of photos and labels that tell it

whether or not any two photos represent the same person. So, given enough data and computational resources, it will learn the patterns that distinguish one face from another. Facial recognition is different from other facial analysis tasks such as gender identification or emotion recognition, which are far more error prone.[26,27] The crucial difference is that the information required to identify faces is present in the images themselves. Those other tasks involve guessing something about a person—their gender identity or emotional state—based on their face, which puts an inherent limit on their accuracy.

Civil rights advocates have often lumped together facial recognition with other error-prone technologies used in the criminal justice system, like those that predict the risk of crime—despite the fact that the two technologies have nothing in common and the fact that error rates differ by many orders of magnitude. (The majority of people who are labeled "high risk" by predictive AI do not in fact go on to commit another crime.)

The biggest danger of facial recognition arises from the fact that *it works really well,* so it can cause great harm in the hands of the wrong people. Kashmir Hill, in her book *Your Face Belongs to Us,* details many harmful ways in which it has been used.[28] For example, oppressive governments can and do use it to identify people in peaceful protests and retaliate against them.[29]

Facial recognition can also be abused by private companies. Madison Square Garden is a famous venue for sports events and concerts in New York City. In 2022, lawyer Nicolette Landi was denied entry to a Mariah Carey concert at the venue.[30] Her boyfriend had bought the nearly $400 tickets for her birthday. She was one of many lawyers turned away from various events at Madison Square Garden. The reason? The company that

operates the venue had banned all lawyers who worked at firms that had sued it—even if they weren't responsible for the lawsuit, and even if they were longtime visitors with season tickets. The ban was enforced using facial recognition.

When critics oppose facial recognition on the basis that it doesn't work, they may simply try to shut it down or shame researchers who work on it. This approach misses out on the benefits that facial recognition has brought. For example, the Department of Homeland Security used it in a three-week operation to solve child exploitation cold cases based on photos or videos posted by abusers on social media.[31] It reportedly led to hundreds of identifications of children and abusers. Of course, there are more mundane benefits of facial recognition as well: unlocking our smartphones or easily organizing photos into albums based on who appears in them.

To be clear, even though facial recognition can be highly accurate when used correctly, it can easily fail in practice. For example, if used on grainy surveillance footage instead of clear photos, false matches are more likely. U.S. pharmacy chain Rite Aid used a flawed facial recognition system that led to employees wrongly accusing customers of theft. False matches happened thousands of times. The company tried its best to keep the system a secret. Fortunately, law enforcement agencies were paying attention. The Federal Trade Commission banned Rite Aid from using facial recognition for surveillance purposes for five years.[32]

To summarize, a nuanced approach to the double-edged nature of facial recognition would be to engage in vigorous democratic debate to identify which applications are appropriate, to resist inappropriate uses, and to develop guardrails to prevent abuse or misuse, whether by governments or private actors.

A Series of Curious Circumstances Led to This Book

In late 2019, a former researcher from an AI company reached out to Arvind out of the blue. The company is in the lucrative business of hiring automation—a business that is filled with snake oil, as we described above. The researcher explained that people at the company knew the tool wasn't very effective, in contrast to the company's marketing claims, but the company had suppressed internal efforts to investigate its accuracy.

Coincidentally, around the same time, Arvind was invited to give a public lecture at MIT. The meeting with the researcher fresh in his mind, he spoke about AI snake oil, showcasing the sketchiness of hiring automation. Encouraged by the audience's reaction, he shared his presentation slides online, thinking that a few scholars and activists might find them interesting. But the slides unexpectedly went viral. They were downloaded tens of thousands of times and his tweets about them were viewed two million times.

Once the shock wore off, it was clear to Arvind why the topic had touched a nerve. Most of us suspect that a lot of the AI around us is fake, but we don't have the vocabulary or the authority to question it.[33] After all, it's being peddled by supposed geniuses and trillion-dollar companies. But a computer science professor calling bullshit gave legitimacy to those doubts. It turned out to be the impetus that people needed to share their own skepticism.

Within two days, Arvind's inbox had forty to fifty invitations to turn the talk into an article or even a book. But he didn't think he understood the topic well enough to write a book. He didn't want to do it unless he had a book's worth of things to say, and he didn't want to simply trade on the popularity of the talk.

The second best way to understand a topic in a university is to take a course on it. The best way is to teach a course on it. So that's what Arvind did, teaming up with Princeton sociology professor Matthew Salganik. Matt had published many foundational pieces of research showing why it's hard to predict the future with AI. We'll see two of them in chapter 3. The course was called Limits to Prediction. Matt and Arvind invited the students in the course to conduct research. One of the students in the course was Sayash.

Sayash had just joined Princeton, having previously worked at Facebook. He ultimately decided to leave Facebook to obtain a PhD and pursue public-interest technology outside a tech company. He was accepted to a few computer science PhD programs. Accepted students are invited to visit the departments in person, to meet prospective collaborators and ask questions to judge whether they would be a good fit.

When visiting departments, PhD students are advised to ask questions of this sort: What is your style of advising? How much time do your students take off? What is your approach to work-life balance? These questions are important, and they can tell you how an advisor works, but not what they value and how they think. A far more revealing question is "What would you do if a tech company files a lawsuit against you?" The answer can tell you the advisor's stance on Big Tech, how they view the impact of their research, and what they would do in a crunch. It is also unusual enough that potential advisors wouldn't have prepared their answers in advance.

Sayash asked every potential advisor this question. It carried the element of surprise, yet the scenario it described was not completely unthinkable. When Arvind answered, "I would be glad if a company threatened to sue me for my research, because

that means my work is having an impact," Sayash knew he had found the right program.

In the course on limits to prediction, students in the class were interested in predictive AI: in any and all attempts to predict the future using data, especially in social settings, ranging from civilizations to social media. Some interesting questions we looked at were: Can we predict geopolitical events such as election outcomes, recessions, or social movements? Can we predict which videos will go viral?

What we found was a graveyard of ambitious attempts to predict the future. The same fundamental roadblocks seemed to come up over and over, but since researchers in different disciplines rarely talk to each other, many scientific fields had independently rediscovered these limits. We were alarmed by the contrast between the weight of the evidence and the widespread perception that machine learning is a good tool for predicting the future.

The course included many case studies, including Google Flu Trends. This was a project that Google launched in 2008 to predict flu outbreaks by analyzing the search queries that its millions of users make every day. An increase in searches for flu-related terms could be indicative of an imminent outbreak. Google heavily promoted it as an example of AI and mass data collection used for social good. But within a few years, the accuracy of the predictions dropped precipitously. One reason was that it is hard to distinguish between media-driven panic searches and actual increases in flu activity. Another was that Google's own changes to its app changed people's search patterns in ways that weren't accounted for by the AI. Google Flu Trends ultimately ended up as a cautionary tale.[34] The lesson is that even in cases where it is possible to make somewhat accurate forecasts, it is very easy to get the details wrong.

Sayash found that the course confirmed his previous experiences at Facebook, where he saw how easy it was to make errors when building AI and to be overoptimistic about its efficacy. Errors could arise due to many subtle reasons and often weren't caught in testing, but only when AI was actually deployed to real users.[35] Sayash decided to choose the limits of AI as his research topic.

After four years of research, separately and together, we're ready to share what we've learned. But this book isn't just about sharing knowledge. AI is being used to make impactful decisions about us every day, so broken AI can and does wreck lives and careers. Of course, not all AI is snake oil—far from it—so the ability to distinguish genuine progress from hype is critical for all of us. Perhaps our book can help.

The AI Hype Vortex

Since we started working together, we've come to better appreciate why there is so much misinformation, misunderstanding, and mythology about AI. In short, we realized that the problem is so persistent because researchers, companies, and the media all contribute to it.

Let's start with an example from the research world. A 2023 paper claimed that machine learning could predict hit songs with 97 percent accuracy.[36] Music producers are always looking out for the next hit, so this finding would have been music to their ears. News outlets, including *Scientific American* and Axios, published pieces about how this "frightening accuracy" could revolutionize the music industry.[37,38] Earlier studies had found that it is hard to predict if a song will be successful in advance, so this paper seemed to describe a dramatic achievement.

Unfortunately for music producers, we found that the study's results were bogus.

The method presented in the paper exhibits one of the most common pitfalls in machine learning: data leakage. This means roughly that the tool is evaluated on the same, or similar, data that it is trained on, which leads to exaggerated estimates of accuracy. This is like teaching to the test—or worse, giving away the answers before an exam. We redid the analysis after fixing the error and found that machine learning performed no better than random guessing.

This is not an isolated example. Textbook errors in machine learning papers are shockingly common, especially when machine learning is used as an off-the-shelf tool by researchers not trained in computer science. For example, medical researchers may use it to predict diseases, social scientists to predict people's life outcomes, and political scientists to predict civil wars.

Systematic reviews of published research in many areas have found that the *majority* of machine-learning-based research that was re-examined turned out to be flawed.[39] The reason is not always nefarious; machine learning is inherently tricky, and it is extremely easy for researchers to fool themselves. Overall, research teams in more than a dozen fields have compiled evidence of widespread flaws in their own arenas, unaware that they were all part of a far-reaching credibility crisis in machine learning.

The more buzzy the research topic, the worse the quality seems to be. There are *thousands* of studies claiming to detect COVID-19 from chest x-rays and other imaging data. One systematic review looked at over four hundred papers, and concluded that *none* of them were of any clinical use because of flawed methods.[40] In over a dozen cases, the researchers used a training dataset where all the images of people with COVID-19 were from adults, and all the images of people without COVID-19

were from children. As a result, the AI they developed had merely learned to distinguish between adults and children, but the researchers mistakenly concluded that they had developed a COVID-19 detector.

We ourselves discovered flaws in many studies, mainly in the field of trying to predict civil wars (in short: it doesn't work). When we tried to publish a paper about an entire body of research being flawed, no journal was interested. It is notoriously hard to correct flaws in the scientific record. We eventually published our paper, but only after reframing it to be more palatable, as a guide to future researchers to avoid these pitfalls.

These days, when we find flawed machine learning papers, we don't even try to correct the record. The system doesn't work. In fact, in many fields, studies that fail attempts at replication by other research groups are cited *more* than those that replicate successfully.[41] The party line among scientists is that science "self-corrects," meaning that the normal process of science is sufficient to root out flawed research, but everything we've seen about the process suggests otherwise.

To be clear, incorrect machine learning claims in research papers usually don't result in broken AI products on the market. If a music producer tried to predict hits using a flawed method, they would quickly find out that it doesn't work. (Commercial AI snake oil usually results from companies knowingly selling AI that doesn't work, rather than they themselves being fooled.) Still, the ocean of scientific misinformation damages the public understanding of AI, because the media tends to trumpet every purported breakthrough.

There are rays of hope, though. In summer 2022, we organized a day-long online workshop to discuss the spate of flawed machine-learning-based science. To our surprise, hundreds of scientists showed up. Based on the workshop, we led a team of

about twenty researchers across many disciplines to devise an intervention: a simple checklist that helps scientists better document how they use machine learning, which can help minimize errors and make them easier to spot when they do creep in.[42] It's still early days, and it remains to be seen if our intervention will be adopted. At any rate, scientific practice changes glacially, and it is likely that things will continue to get worse for a while before they get better.

Let's turn to companies. While overhyped research misleads the public, overhyped products lead to direct harm. To study this, we teamed up with colleagues Angelina Wang and Solon Barocas and investigated uses of predictive AI in industry and government.[43] We documented about fifty applications spanning criminal justice, healthcare, welfare allocation, finance, education, worker management, and marketing. Most of these deployments are recent. During the 2010s, predictive AI extended its tentacles into many spheres of life, judging us and determining our opportunities in life based on data covertly collected about us.

We realized that while vendors of these tools aggressively chase clients, they are rarely transparent about how well their products work, or if they work at all. Notably, as far as we know, no hiring automation company has ever published a peer-reviewed paper validating its predictive AI, or even allowed an external researcher to evaluate it. Two of the leading companies made a show of external audits: Pymetrics contracted with a leading research group from Northeastern University, and HireVue contracted a noted independent auditor. But in both cases, the researchers were allowed to analyze only whether the AI was biased with respect to race or gender, and not whether it worked. The companies cleverly used a concern about discrimination to their advantage. If what you have is an

elaborate random number generator that works equally poorly for everyone, it's easy to make it unbiased!

Here, too, there are slivers of good news. Regulators are wising up to the fact that many predictive AI products don't work. In 2023, the U.S. Federal Trade Commission (FTC) warned companies that "we're not yet living in the realm of science fiction, where computers can generally make trustworthy predictions of human behavior. Your performance claims would be deceptive if they lack scientific support or if they apply only to certain types of users or under certain conditions."[44] The key word here is "deceptive"; the FTC is authorized by Congress to police deceptive practices by companies. We hope companies will heed that warning.

If researchers and companies kindle the sparks of hype, the media fans the flames. Every day we are bombarded with stories about purported AI breakthroughs. Many articles are just reworded press releases laundered as news.

Of course, with the media so reliant on clicks and newsrooms so cash strapped, this is no surprise. Still, there are systemic problems in the industry besides crumbling revenue. Many AI reporters practice what's called access journalism. They rely on maintaining good relationships with AI companies so that they can get access to interview subjects and advance product releases. That means not asking too many skeptical questions.

Claims of AI being conscious have proven particularly irresistible to the media. When a Google engineer claimed in June 2022 that the company's internal chatbot had become sentient (and faced "bigotry"), just about every publication ran with that headline.[45] The same thing happened when Bing's chatbot claimed to be sentient in early 2023. That's despite the fact that most AI researchers don't think there is any scientific basis for these claims.

There are many AI journalists who rise above the fray and do excellent investigative work. But so far they are a handful, constantly swimming against the tide. We've had the opportunity to discuss the hype problem with journalists and speak at a few journalism conferences. We learned about many ongoing efforts to improve the quality of tech journalism.

For example, the Pulitzer Center funds a network of journalists to work on "in-depth AI accountability stories that examine governments' and corporations' uses of predictive and surveillance technologies to guide decisions in policing, medicine, social welfare, the criminal justice system, hiring, and more."[46] Many notable investigations have resulted from this program, including one by Ari Sen and Derêka K. Bennett for the *Dallas Morning News*. Sen and Bennett looked into Social Sentinel, an AI product used by schools across the United States to scan students' social media posts, purportedly to identify safety threats, but often misused to surveil student protests.[47]

The Pulitzer Center fellowships support only ten journalists per year. In the long run, whether or not journalism can serve as a check against Big Tech's power will depend on whether funding models like these—that don't rely on clicks—can be scaled up.

AI experts have a responsibility to speak up against hype, whether it comes from researchers, companies, or the media. We are trying to do our part. In our newsletter, at AISnakeOil.com, we comment on new developments in AI and help readers separate the milk from the froth.[48]

What Is AI Snake Oil?

In the late nineteenth and early twentieth centuries, snake oil peddlers were rampant in America, selling miracle cures and health tonics under false pretenses. Figure 1.1 shows a typical

FIGURE 1.1. A 1905 advertisement for snake oil. (*Sources:* https://www.nlm.nih.gov/exhibition/ephemera/medshow.html, attributed to Clark Stanley's Snake Oil Liniment, *True Life in the Far West*, 200 page pamphlet, illus., Worcester, Massachusetts, c. 1905, 23 × 14.8 cm. https://commons.wikimedia.org/w/index.php?curid=47338529.)

advertisement. Snake oil sellers exploited people's unscientific belief that oil from snakes had various health benefits, and their inability to tell effective treatments from useless ones. Besides, most of the concoctions being sold as snake oil didn't in fact contain any. In some cases, these medicines were ineffective but harmless. In others, they led to the loss of life or health. Until

the Food and Drug Administration (FDA) was established in 1906, there was no good way to keep snake oil salesmen accountable to their promises regarding the contents, the efficacy, or the safety of their products.

AI snake oil is AI that does not and cannot work, like the hiring video analysis software that originally motivated the research that led to this book. The goal of this book is to identify AI snake oil—and to distinguish it from AI that can work well if used in the right ways. While some cases of snake oil are clear cut, the boundaries are a bit fuzzy. In many cases, AI works to some extent but is accompanied by exaggerated claims by the companies selling it. That hype leads to overreliance, such as using AI as a replacement for human expertise instead of as a way to augment it.

Just as important: even when AI works well, it can be harmful, as we saw in the example of facial recognition technology being abused for mass surveillance. To identify what the harm is and how to remedy it, it is vital to understand whether the problem has arisen due to AI failing to work, or being overhyped, or in fact working exactly as intended. Harm and truthfulness are the two axes in figure 1.2. In this book, we're interested in everything except the bottom left part of the figure, which is AI that both works and is benign.

With this picture in mind, here's a roadmap of the rest of the book.

Chapter 2 is about automated decision-making, which is one area where AI, specifically predictive AI, is increasingly used: predicting who will commit a crime, who will drop out of school, and so forth. We'll look at many examples of systems that have failed and caused great harm. In our research, we've identified a recurring set of reasons these failures keep happening—reasons that are intrinsic to the use of predictive

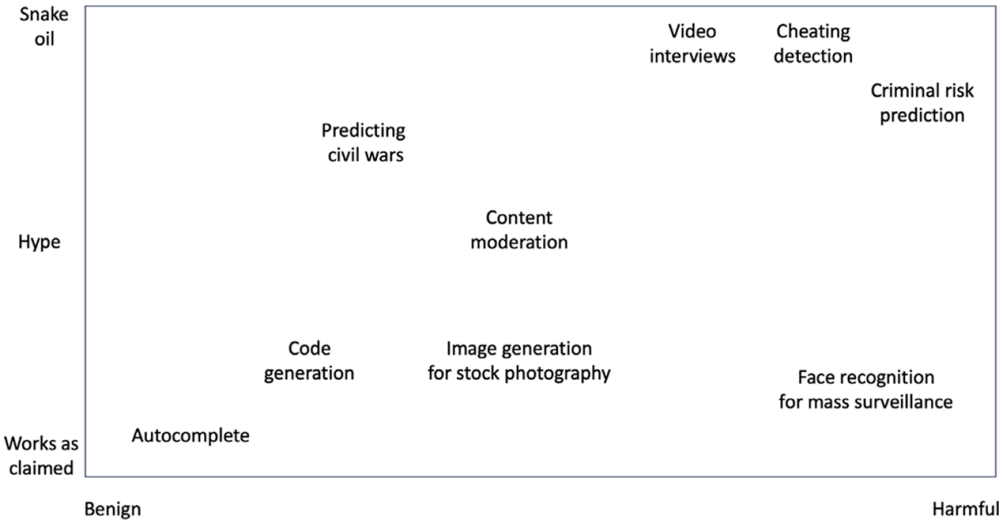

FIGURE 1.2. The landscape of AI snake oil, hype, and harms, showing a few illustrative applications.

logic in these high-impact systems. We'll end the chapter by asking if it is possible to reimagine decision-making without predictive AI, and we'll discuss what sorts of organizational and cultural adaptations we'll need in order to embrace the unpredictability inherent to consequential decisions.

In chapter 3 we'll take a step back to understand why predicting the future is so hard. Our answer is that its challenges are ultimately not about AI, but rather the nature of social processes; it is inherently hard to predict human behavior, and we'll see many reasons for this. We'll review evidence from many efforts to predict the future, from crime to children's life outcomes. We'll draw from academic studies as well as the rare cases where commercial products have been subjected to independent scrutiny. We'll look at prediction of both positive outcomes, such as succeeding at a job or publishing a bestseller, and negative outcomes, such as failing to pay back a loan; all of these turn out to be hard to predict. We'll also look at less

consequential but more easily analyzed prediction tasks such as identifying which social media posts will go viral. And finally, in addition to outcomes about individuals, we'll look at macro-level predictions such as the evolution of pandemics. Across all of these domains, strikingly common patterns emerge, which lead us to conclude that the limitations of predictive AI won't go away in the foreseeable future.

It's simple to state the primary limitation of predictive AI: it's hard to predict the future. But with generative AI, to which we turn next, things are more complicated. The technology is remarkably capable, yet it struggles with many things a toddler can do.[49] It is also improving quickly. So, to understand what the limitations are and have some sense of where things might be going, it's important to understand the technology. In chapter 4, we hope to demystify how generative AI works.

We'll also discuss the many harms that arise from generative AI. In some cases, harms arise because the product is flawed. For instance, software that claims to detect AI-generated essays doesn't work, which can lead to false accusations of AI-based cheating. In other cases, harms arise because the product works well. Image generators are putting stock photographers out of jobs even as AI companies use their work without compensation to build the technology. Of course, there are many applications of generative AI that both work well and are broadly beneficial, such as automating some parts of computer programming (although, even here, there are minor risks that programmers should watch out for, such as the possibility of bugs in AI-generated code that might give hackers an advantage). Given the focus of the book, we won't spend much time on these beneficial applications. But we should emphasize that we are excited about them and about the potential of generative AI in general.

In chapter 5 we turn to questions of existential risk that have gripped the public discourse on AI. The fear is that once future AI systems are sufficiently advanced, they will be too hard to control. Our key perspective is that these fears rest on a binary notion of AI that crosses some critical threshold of autonomy or superhuman intelligence. But this idea is contradicted by the history of AI. The technology has gradually been increasing in flexibility and capability, which we explain by introducing the concept of a "ladder of generality." Current technology is already on the seventh rung of this ladder, each step being more general and powerful than the ones below it.

We have every reason to think that this pattern of step-by-step progress will continue. This means we don't have to speculate about the future but can instead learn from history. And what this grounded analysis shows is that claims of out-of-control AI rest on a series of flawed premises. Of course, we must take risks concerning powerful AI seriously. But we'll show that we already have the means to address them calmly and collectively.

In chapter 6, we turn to social media, where so-called recommendation algorithms are used for creating the personalized feeds we scroll through. AI is also used for determining which content violates policies and must be taken down; this process is called content moderation. The chapter is primarily about content moderation AI, with a brief discussion of recommendation algorithms. The central question we examine is whether AI has the potential to remove harmful content such as hate speech from social media without curbing free expression, as tech companies have often promised.

In this debate, much attention has been paid to the inevitable errors of enforcement, such as a piece of content being mistakenly flagged as unacceptable and taken down. But even if these

errors are fixed, the more fundamental issue is that platforms have this power to regulate speech in the first place, with little accountability. We lack a democratic process to decide the rules by which online speech should be governed and to find a balance between values such as free speech and safety. Given this reality, AI will remain impotent at easing our frustrations with social media.

We have placed content moderation AI in the middle of figure 1.2. It works well enough that social media companies have come to rely on it, but it is often misleadingly portrayed as a way out of the moral and political quandaries that beset social media governance, which it is not. As for its harmfulness, while content moderation has often failed badly and even enabled large-scale real-world violence, we argue that these failures are not fundamentally about the technology but rather intrinsic consequences of handing over the digital public square to unaccountable private entities.

Predictive AI, generative AI, and content moderation AI are the three main types of AI that we'll discuss in this book. That's not an exhaustive list. As discussed earlier, there are many applications such as autocomplete or spell-check that work well and fly under the radar. There are also applications such as robotics and self-driving cars that are worth discussing but didn't make the cut, in part because they don't yet have societal consequences on the same scale as the applications we discuss. Still, the conceptual understanding we provide in this book will help you evaluate other AI applications as well—which ones are likely to work, and which ones might be snake oil.

In chapter 7 we consider the question of why myths about AI are so pervasive. Companies not only hype up their tech but also use their enormous wealth and power in ways that make academia and the press less effective as counterweights

to their self-serving claims. In fact, academic researchers are as often a source of hype in AI as they are a voice of reason. In many fields, researchers have falsely come to a consensus that AI is highly accurate in their respective domains, all based on flawed and irreproducible research. We'll look at civil war prediction as one example. While faulty AI research usually doesn't lead to the deployment of faulty products—which is why we've put it in the top left quadrant of figure 1.2—it is nonetheless harmful and wasteful because it misleads the public. Turning to the media, we discuss ways in which journalists knowingly or unknowingly contribute to AI hype, and give you a guide for reading the news skeptically.

In the final chapter, we look at directions for change. We identify three broad paths. The first is to set ground rules for companies to govern how they build and advertise their products. We think there is an important role for regulation here, while we acknowledge that regulation shouldn't go overboard. The second path is the way in which we integrate AI into society. For example, what is the role of AI in education and in children's lives in general? And in the workplace, will we use AI to replace jobs or augment them? We see many of these questions as social and political choices rather than inevitable consequences of the technology itself.

Our third suggested path is to focus on the demand for AI snake oil rather than its supply. We show that over and over again, AI snake oil is appealing because those buying it are in broken institutions and are desperate for a quick fix. For example, schoolteachers, already overworked, have reacted badly to the disruption caused by students using AI to help with their homework. Unable to carry out the overhaul of their teaching and assessment strategies that AI necessitates, they have turned to cheating detection software. But these products don't work

and have led to a spate of false accusations of academic dishonesty, often with disastrous consequences for students.

We can't fix these problems by fixing AI. If anything, AI snake oil does us a favor by shining a spotlight on these underlying problems. More broadly, we show how concerns about AI, especially in the labor market, are often really about capitalism. We must urgently figure out how to strengthen existing safety nets and develop new ones so that we can better absorb the shocks caused by rapid technological progress and reap its benefits.

Who This Book Is For

We hope this book will be interesting to three kinds of readers. Maybe you simply want to make sense of what's going on. You've seen the headlines, and you're wondering if AI can really predict earthquakes or pass the bar exam. And if so, how? Which jobs will still be around in twenty years? What will our children's lives be like?

What we offer to sate your curiosity are not philosophical musings about what it means to be human in the age of AI. Reasonable people can have different opinions on whether we're living through something profound or just the next step in the march of automation. Rather, what we hope to impart is a nuts-and-bolts understanding of what's going on behind the screen.

Or you may be interested in AI because you need to make decisions about AI at your workplace. We hope that this book will help you understand which types of AI work, which ones don't, and what the gotchas are. Throughout the history of AI, computer scientists have attempted to classify which problems are "easy" for AI and which ones are "hard." None of these

sweeping generalizations has stood the test of time as the technology evolved. Our approach, instead, is to treat each kind of AI individually.

Finally, you may be interested in AI because you want to take action against the harms being perpetrated in the name of AI. Public-interest advocates have built up effective movements to resist harmful predictive AI. But with generative AI, the battle lines are still forming. If the technology leads to an economic transformation, as generative AI companies hope, it probably won't be good news for labor, whether or not it eliminates jobs. That's because this type of AI relies on the invisible, drudging, low-wage work of millions of people to create training data, as well as the use of data found on the web without credit or compensation to the writers, artists, and photographers who created them.

In the wake of the Industrial Revolution, millions of new jobs were created in factories and mines, with horrific working conditions. It took many decades to secure labor rights and improve workers' wages and safety. Similarly, there is a movement today to secure labor rights, human creativity, and dignity in the face of encroaching automation.[50] It is far from clear that this movement will succeed. It's up to all of us.

Finally, a quick note to instructors and students using this book in a course: we have exercises and other pedagogical materials on our website, AISnakeOil.com.[47]

Chapter 2

HOW PREDICTIVE AI GOES WRONG

IN 2015, administrators at Mount St. Mary's University, a private university in Maryland, USA, wanted to increase their university's retention rate—the proportion of admitted students who go on to graduate. The school conducted a survey to find which students were struggling. This might sound like a worthy goal; once the administration knows who is struggling, they can offer additional support to those students to help them adjust to college. But instead, the president suggested dismissing students who weren't doing well. He reasoned that if these students dropped out in the first few weeks of the semester, as opposed to later, they wouldn't count as "enrolled" and therefore wouldn't decrease the university's retention rate.

In a faculty meeting, the president said: "My short-term goal is to have 20–25 people leave by the 25th [of September]. This one thing will boost our retention 4–5%."[1] Professors raised objections by pointing out that it is hard to judge who will be successful in the first few weeks after entering college. "This is hard for you because you think of the students as cuddly

bunnies, but you can't," the president responded. "You just have to drown the bunnies . . . put a Glock to their heads."

This is a startling example, but the fact is that many schools would like to predict which students are at risk of dropping out—some of them for reasons more aligned with students' well-being. An AI-based product called EAB Navigate could automate this process. In its marketing pitch to schools, EAB claimed: "The model will provide your school and its advisors with invaluable and otherwise unobtainable insight into your students' likelihood of academic success." Even if some schools might use this insight to pressure students to leave, others could conceivably use it to design interventions that might help students stay in school. But interventions that seem helpful could also be questionable. For example, the tool helps by recommending alternative majors in which a student would be more likely to succeed. This might have the effect of driving out poorer and Black students—whom the tool is more likely to flag—from more lucrative but more challenging STEM majors.[2] And throughout this process, students may have no idea that they are being evaluated using AI.

EAB Navigate is an example of an automated decision-making system that uses predictive AI. In this realm, there is a vast amount of AI snake oil.

Companies selling these tools make strong claims about their utility.[3] They are deployed widely by governments and private companies. Still, unlike generative AI applications like ChatGPT, predictive AI has largely flown under the radar when it comes to public interrogation. In many cases, including EAB Navigate, the decision subjects don't even know they are being automatically evaluated.

In this chapter, we will see how predictive AI goes wrong. While a full account of these failures would take up more than

the entire length of this book,[4,5]* we will highlight common failures that are nevertheless hard to fix, underscoring the difficulty of making predictive AI work.

Before we dive in, let's look more closely at how these automated decisions are made.

Predictive AI Makes Life-Altering Decisions

Algorithms like EAB Navigate are everywhere, and they are used to automate consequential decisions about you—with or without your knowledge.[6] When you show up at a hospital, an algorithm could determine if you need to be admitted overnight or released the same day. When you apply for public benefits such as child welfare, an algorithm could determine if your application is valid or if you are trying to commit fraud. When you apply for a job, an algorithm could determine if your application will be considered by a recruiter or discarded. When you visit a beach, an algorithm could determine if the water is safe to swim in.

An algorithm is a set of steps or rules used to make a decision. Sometimes these rules are developed by people or institutions. For example, during the COVID-19 pandemic, the U.S. government sent stimulus checks to help citizens deal with the economic hardships they faced. Adults were paid USD 1,200; children were paid USD 500. Policymakers decided these rules. But once the rules were established, people's eligibility was automatically determined based on their past tax records.

* Two popular incident databases record failures of AI, including predictive AI, in the real world. The AI Incident Database has over 600 reports, and the AI, Algorithmic, and Automation Incidents and Controversies Repository has over 1,400 reports, as of 2024.

1. If the applicant is a U.S. citizen and eighteen or older, send a check for USD 1,200.
2. If they are a U.S. citizen and under eighteen: send a check for USD 500.
3. If they're not a U.S. citizen or earn more than USD 75,000 per year*: don't send a check.

This type of algorithm, in which the rules are developed manually but applied automatically, is commonly employed in the public sector, such as in allocating funds for public housing or welfare.

But increasingly, algorithms are used to *develop* the rules automatically from patterns in past data. For instance, while watching Netflix, if you rated *Forrest Gump* highly but not *The Shining*, a recommendation algorithm might predict that you'll enjoy drama films over horror films. Here, a Netflix employee is not manually creating a rule stating that users who liked *Forrest Gump* should be recommended other drama films. Instead, your ratings and the time you spend watching different types of films determine what films you're recommended next. Unlike the first kind of algorithm, the rules are both developed and applied automatically.

This kind of algorithm is called a model, which is a term you might encounter frequently. A model is usually a set of numbers that mathematically specify how the system should behave. These numbers may not be intelligible to a human, even to the developer of the system, unless the model is explicitly designed with that goal in mind. Models are created from data, or "trained," using a set of statistical techniques called machine learning.

* The actual algorithm was slightly more complex: the amount paid if a person earned more than USD 75,000 tapered off. And there were different rules for people with children.

These models can be much more consequential than deciding your next Friday night movie. They are used for allocating scarce resources, such as jobs or loans. They provide certain opportunities to people and foreclose other opportunities. This is what we call predictive AI: models used for decision-making based on predictions about the future, such as who will do well at a job or who will pay back a loan.

Take the example of criminal justice, where predictive AI is used for many kinds of decisions: Should an inmate get parole? What should happen to people who are arrested? Before trial, a judge needs to decide whether a defendant should be detained in jail, released on bail (and if so, what the bail amount should be), or released without any bail, but perhaps with other restrictions such as ankle monitors.

A lot rides on the answer.[7] Spending time in jail can ruin someone's life. They can lose their income in the short term—and even if they are released, especially if they have restrictions such as ankle monitors, it will still be much harder to find a job. They suffer from increased physical and mental illnesses, both due to stigmatization and because of poor conditions in jail. And a large fraction of the people in jail are simply there because they can't afford high bail amounts.

As a result, the criminal justice system disproportionately burdens the poor and leads to cycles of poverty and racial inequality. Almost half a million people are in U.S. jails at any given time without having been convicted of a crime.[8] Despite violent crime in the country going down by almost 50 percent, the number of people jailed in the last four decades has nearly doubled.[9,10]

Many U.S. states mandate the use of risk assessment tools to decide whether a person should be released or detained before trial, and predictive AI is often used. The tools produce two

main risk scores. One is the risk that the defendant will commit a crime, especially a violent crime, if released. The other is the risk that they will fail to appear in court on the appointed date. In each case, defendants are labeled low, medium, or high risk. The tools try to calculate these scores based on certain features of the defendants.

Let's zoom in on one predictive AI product: COMPAS, which stands for Correctional Offender Management Profiling for Alternative Sanctions. COMPAS relies on defendants' answers to 137 questions to make decisions about them.[11] It includes questions about their past history of crime or failure to appear for court cases. It also includes questions about things individuals have little or no control over, such as how often their family members have been arrested, or whether their friends or acquaintances have been arrested. And some questions seem to punish people for their personality or for poverty, such as "How often do you feel bored?" and "How often do you have barely enough money to get by?"

The developers of COMPAS aim to predict if a defendant will fail to appear in court or be arrested for a crime within two years.[12] It is trained using data on defendants' behavior in the past. COMPAS uses this past data to find patterns in the characteristics of defendants who failed to appear in court, such as their age, number of past offenses, and the criminal history of their peers, and tries to distinguish them from those who did appear for their trial. This highlights an assumption built into much of predictive AI: people with similar characteristics will behave similarly in the future.

Predictive AI is quickly gaining in popularity. Hospitals, employers, insurance providers, and many other types of organizations use it. A major selling point is that it allows them to reuse existing datasets that have already been collected for

other purposes, such as for bureaucratic reasons and record keeping, to make automated decisions.

Yet, predicting the future is hard. People could face unexpected setbacks such as getting evicted, or surprises such as winning a lottery, that no model can predict. Small changes in someone's life, such as a visit to the emergency room, could have large effects on their future—say, due to compounding medical bills.

Claims about the virtues of predictive AI are pervasive. Upstart's model decides whether an applicant should be approved for a loan.[13] The company claims that its model is significantly more accurate than traditional lending models. It also claims to be a leader in fair lending practices and promises that future versions of the model will continue to be fair. Finally, it claims to be highly efficient: three-quarters of its loan decisions require no input from humans. Other companies make similar promises. HireVue sells tools to automate hiring decisions. It claims to predict how well a candidate will perform at their job once hired. HireVue's home page announces its selling points: "Fast. Fair. Flexible. Finally, hiring technology that works how you want it to."

Despite these statements, human decisions still come in at many stages of the development of predictive AI, often hidden from sight. And since the data used to train predictive AI is itself created using human discretion, we cannot guarantee that the decisions will be unbiased or fair.[14] In other words, decisions made using predictive AI may still be very human.

We were skeptical of the promises made by developers of predictive AI, and so we decided to dig deeper. Together with researchers Angelina Wang and Solon Barocas, we spent over a year reading hundreds of research papers, news articles, and

reports about automated decision-making systems. We were surprised to find that the same set of shortcomings plague many applications of predictive AI.[15] In the next few sections, we will take a closer look at these shortcomings, through real-world case studies.

Spoiler: We think predictive AI falls far short of the claims made by its developers.

A Good Prediction Is Not a Good Decision

Healthcare workers make a critical decision when a patient comes to the hospital with symptoms of pneumonia: whether to send the patient home after treatment or admit them overnight. They consider a patient's age as well as preexisting medical conditions, such as asthma, which place them at high risk if they get pneumonia. High-risk patients with pneumonia symptoms are typically sent straight to the ICU to avoid complications.[16]

In a 1997 study, researchers investigated whether AI could make better decisions than healthcare workers in predicting outcomes of patients with pneumonia.[17] Like countless other AI researchers, they thought a model trained with a sufficiently large amount of data would make better decisions than a human, decisions which would help to prioritize high-risk patients.

The researchers trained an AI model and found that it had fairly good accuracy in predicting which pneumonia patients would face complications or death. Surprisingly, the model found that having asthma leads to *lower* risk of complications due to pneumonia. If this model were to be used in a hospital, it would be more likely to send an asthmatic patient home than a non-asthmatic one (let alone sending asthmatic patients to the ICU). How is this possible?

The researchers looked more closely. They found that in the dataset, asthmatic patients were indeed at a lower risk for serious pneumonia or death—but this was only because the training data was collected under the hospital's preexisting decision-making system.[15] Asthmatic patients were sent straight to the ICU as soon as they arrived, thereby receiving more intensive care than non-asthmatic patients and thus becoming less likely to develop complications.

So the model's predictions were correct—but they were correct under the hospital's existing system. Ironically, the model was meant to replace this system.

Deploying this model would have led to disastrous decisions. Asthmatic patients would have been perceived as lower risk by the model and sent home. Thankfully, the researchers realized what had happened and did not use the model in this particular hospital.

This error shines a light on a fundamental limitation of much of predictive AI: AI can make good predictions *if nothing else changes*. But correlation is not causation (and having asthma does not reduce the risk of complications from pneumonia). In other words, predictive AI does not account for the impact of its own decisions. That is, it can't tell us what would happen if something changed in a system—if, for instance, the model started sending asthma patients home.

Let's look at another example from healthcare. A 2018 study claimed to accurately predict hypertension using machine learning, with impressive results.[18] On closer look, however, it became clear that the model was evaluated on people who had already been seen by clinicians.[19] This meant that a critical input to the model was whether a patient already used drugs for controlling hypertension. But if a patient is using antihypertensive drugs, they have obviously already been diagnosed with hyper-

tension. These cases were still counted as successes when evaluating the model, vastly inflating its accuracy numbers.

This issue arose in part because the researchers relied on existing data rather than new data collected specifically for the task. Generally, collecting data is expensive and time-consuming. So, some AI developers claim that existing data is enough to make good decisions.

The importance of collecting new data is widely recognized in medicine, which relies on randomized controlled trials (RCTs) to test the impact of introducing a new drug or vaccine. To test if a drug performs well, half the people in the trial are randomly assigned to the experimental group and are given the drug. The other half are assigned to the "control" group and are given a placebo. The groups are carefully balanced with respect to demographics like age and gender. A drug's effectiveness is evaluated based on whether it leads to lower disease rates in the experimental group compared to the control. Medical researchers perform RCTs despite their slow pace and high expense for a simple reason—easier, faster methods don't work. The same is true in many areas where AI is used for automated decision-making.

Unfortunately, predictive AI companies don't seem to have realized the importance of collecting data on the impact of their tools' decisions. And such data collection is also expensive and time-consuming, so it would negate their claims of cutting costs and increasing efficiency.

As a result, even if AI can make accurate *predictions* based on past data, we can't know how good the resulting *decisions* will be before AI is deployed on a new dataset or in a new setting. So, when you encounter claims about the effectiveness of predictive AI, it's important to find out if developers evaluated the impact of the decisions or only the accuracy on past data.

Opaque AI Incentivizes Gaming

We've seen that we can't fully determine the impact of AI before deployment, in part because AI makes its predictions assuming that the system will remain the same as it was during testing. But systems change, and people are one obvious wild card. Things become even more complicated when people behave strategically.

The colonial British government in India wanted to reduce the cobra population, so it decided to offer a reward to people who brought in dead cobras. But instead of killing cobras in the wild, people started breeding them to claim the reward—leading to an *increase* in the cobra population. This is an apocryphal example of a mismatch between what is specified and what is desirable. A similar problem occurs in AI. When developers build AI, they specify the exact outcome they want to predict. But because of the reliance on past data, they are forced to choose this outcome based on the data that is already available. As a result, what AI predicts can be a poor proxy for what we want it to predict.

A stark example is AI used for hiring. In the United States, three-fourths of employers use automated tools to screen job candidates.[20] Automated hiring tools come in various forms. Some are used to filter candidates based on their résumés. Others test candidates using an automated video interview. Yet others require them to solve puzzles. These tools are used as the first step in selecting candidates. If someone doesn't pass, their résumé is discarded without a human ever looking at it. But the process is opaque; companies don't share how their software is built, and candidates are left in the dark about what criteria they are judged on.[21]

In response, candidates have developed strategies to work around opaque hiring AI. They stuff their résumés with keywords

from the job application and add the names of top universities in white text (which a human reader can't see, but a computer can recognize).[22] In video interviews that they know will be judged by AI, they try using fancy words such as "conglomerate."[23]

It is not clear whether these strategies work. So a group of journalists decided to investigate. They looked at Retorio, a Munich-based startup that offers hiring tools based on video interviews.[24] Their findings were surprising: simple changes to a person's appearance, such as wearing a scarf or glasses, were enough to drastically change the score that the AI tool produced. Adding a bookshelf or a painting in the background increased scores, while making a video darker (without changing its content) led to lower scores. In another study, researchers looked at personality tests used in hiring and found that merely changing the format of candidates' résumé from PDF to plain text changed their personality scores.[25]

Changing the background or format of a résumé obviously does not change someone's capability to perform well at a job. So why do candidates' scores change? One reason could be that in the data used to train the model, people who had bookshelves in the background performed better compared to those who had a plain background.

This is precisely the kind of correlation that candidates banked on when they added keywords to their résumé. They assumed people with those keywords in their résumé performed better in the past. They wanted to avoid getting rejected just for omitting the right keywords or not using enough fancy words.

Acting strategically—gaming—to avoid negative consequences is common. This is what teachers do when they teach to the test, and what consumers do when they try to increase

their credit scores without changing their spending habits, such as by getting a retail credit card or filling out a prequalification form before applying for credit. But with AI-based hiring, people don't even know what actions influence their chances. Instead of guiding candidates toward decisions that actually improve their skills, AI-based hiring encourages candidates to experiment with superficial changes to their résumé and application.

We are not taking a stance on whether candidates should or shouldn't be attempting these tricks. The ethics of gaming are interesting but tangential to our point. Rather, our point is that when AI companies make claims about accuracy, they fail to account for the effects of people behaving strategically. When the model's output can be easily manipulated using superficial changes, we cannot take claims of accuracy at face value. Opaque models also impose a cost on decision subjects in terms of wasted time, such as when candidates experiment with ways to add invisible fake qualifications to their résumés.

Overautomation

In 2013, the Netherlands deployed an algorithm to flag welfare fraud, replacing an earlier system in which humans reviewed each decision.[26] The algorithm was used to make serious accusations of guilt using only statistical correlations in data, without any other evidence.[27]

This shift to an automated system had many adverse effects. For one, people lost the ability to challenge decisions. Inaccurate or outdated government data often leads to erroneous fraud allegations, which under the new system could not be reversed. Further, the data used to make these allegations was not publicly available, so people had no way to find out why they'd been accused of fraud.