THE AI CON

How to Fight Big Tech's Hype and Create the Future We Want

EMILY M. BENDER
&
ALEX HANNA

THE AI CON

THE AI CON

How to Fight Big Tech's Hype and Create the Future We Want

EMILY M. BENDER
& ALEX HANNA

BH

THE BODLEY HEAD
LONDON

1 3 5 7 9 10 8 6 4 2

The Bodley Head, an imprint of Vintage, is part of
the Penguin Random House group of companies

Vintage, Penguin Random House UK, One Embassy Gardens,
8 Viaduct Gardens, London SW11 7BW

penguin.co.uk/vintage
global.penguinrandomhouse.com

Penguin
Random House
UK

First published in Great Britain by The Bodley Head in 2025
First published in the United States of America by HarperCollins in 2025

Designed by Bonni Leon-Berman

Printed and bound in Great Britain by Clays Ltd, Elcograf S.p.A.

The authorised representative in the EEA is Penguin Random House Ireland,
Morrison Chambers, 32 Nassau Street, Dublin D02 YH68

A CIP catalogue record for this book is available from the British Library

HB ISBN 9781847928610
TPB ISBN 9781847928627

Penguin Random House is committed to a sustainable future for
our business, our readers and our planet. This book is made
from Forest Stewardship Council® certified paper.

MIX
Paper | Supporting
responsible forestry
FSC® C018179
FSC
www.fsc.org

For my kids, and the world I'd like them to have. —Emily

For my dad, who knew what work was. —Alex

CONTENTS

PREFACE

This book is the result of a joyful collaboration between a linguist (Emily) and a sociologist (Alex), rooted in a shared drive to deflate AI hype through analysis grounded in our academic fields and pointed humor.

We met online in 2020, participating in broader discussions about the impacts of technologies sold as "AI". In 2020 and 2021, we collaborated (with three other scholars, led by the inimitable Deb Raji) on some academic papers critiquing the dismal evaluation and data handling practices of the field.

Since then, we've continued to use social media to take down ridiculous claims of tech boosters as well as bad journalism that fawns over them. This work can be trying, since speaking out against people's favorite toys, tech leaders who appeal to a particular type of nerdy masculinity, and exploitative practices draws all kinds of negative pushback. Many people, especially in Silicon Valley and computer science departments, are willing to grant tech companies a good deal of grace for technologies that don't live up to the hype. Beyond that, this work frequently involves contending with racism, sexism, and white supremacy and their associated violence. What kept us going was community, largely in the form of group chats, where we poked fun at all the hype, as well as processing all the nasty replies we received.

In one of those group chats, in April 2022, together with our colleagues Timnit Gebru and Margaret "Meg" Mitchell, we debated how to most effectively respond, especially when the hype takes the form of longer artifacts, such as scientific papers, podcast interviews, and long-form journalism. Meg suggested something in the style of the TV show *Mystery Science Theater 3000* (MST3K), in which the characters take terrible sci-fi movies and make them enjoyable with running commentary.

A few months later, Emily came across a blog post that was too long for a tweet-thread takedown (a Medium post by Blaise Agüera y Arcas of Google titled "Can Machines Learn How to Behave?"). At an estimated sixty-minute read, every single paragraph was oozing AI hype. But the thought of writing a tweet thread or a blog post to counter each part of that seemed exhausting. So Emily asked in the group chat if anyone was up for giving it the *Mystery Science Theater* treatment.

Alex, a big fan of MST3K, jumped on board, and an accidental podcast was born: we livestreamed our takedown of the blog post on the platform Twitch, figuring it would take an hour or so. One hour wasn't enough, so we scheduled a second, and then a third, and then just kept going. A few more streams in, we heeded many calls to turn our series into a podcast and brought Christie Taylor on board as a producer.

Our show, *Mystery AI Hype Theater 3000*, is billed as a space in which to seek catharsis in this age of AI hype: we find the worst of it and pop it with the sharpest needles we can find! Those needles are strong because they're built from linguistic and sociological analysis, but sharp because they're honed in humor.

One of our taglines on the podcast is "Along the way, we learn to always read the footnotes." That's because checking the sources for all of the hype-tastic claims often gives us a good vista on the house of cards (that is, thin research methods, shoddy argumentation, and questionable citation practice) supporting the flashy façade. In that same spirit, it is important to us to cite our sources: we care both about the provenance of the information we are sharing and about giving credit where credit is due. You'll find those sources, along with further details and analysis we deemed too in-the-weeds for the main text, in the endnotes.

Our goal is to help the public at large as well as decision-makers at all levels become resistant to hype. Think of us as your guides to navigating a glitzy technology expo hall, full of salespeople trying to get you to buy a new product or fork over your data. We don't need that energy, and neither do you.

We are writing now, in late 2024, from the inside of what feels like the height of the AI hype bubble. As we say on the podcast: each time we think we've reached peak AI hype—the summit of bullshit mountain—we discover there's worse to come. We're using what we see in this bubble to document the contours of hype about AI, its causes and its short- and long-term effects. Our primary goal is to inhibit the next tech bubble. We hope that by pulling back the curtain, we'll help you to be able to spot the hype now and the next time around, while honing your own needles.

CHAPTER 1

AN INTRODUCTION TO AI HYPE

In late 2023, inside the grand halls of the United States capital of Washington, DC, Senator Charles "Chuck" Schumer, Senate majority leader, led the eighth of a set of forums he had been convening around artificial intelligence, or AI. These "insight forums" were intended to provide the august body of the U.S. Senate with information on how to handle this "brand-new" technology of AI. At this particular meeting, a number of notables were in attendance: researcher Yoshua Bengio, who received one of computer science's highest honors for his work on AI; Jared Kaplan, cofounder of the influential AI startup Anthropic; Aleksander Mądry, OpenAI's "Head of Preparedness"; and Stuart Russell, an influential professor of computer science. Also in the room were people from civil society (including civil rights and nonprofit research groups), policy institutes, and venture capital firms.

Schumer began the conversation with an unusual prompt: What was everyone's p(doom) and p(hope)? Pronounced *pee-doom* (and *pee-hope*), this phrase references notation from statistics and is short for "probability of doom/hope", referring to a popular trope that machines with minds of their own will, perhaps, kill us all, intentionally or unintentionally. Estimates from those in the room ranged from 0 up to 90 percent, according to reporting on the event. Schumer tweeted afterward: "If managed properly, AI promises unimaginable potential. If left unchecked, AI poses both immediate and long-term risks." These risks have been deemed "existential" by those who have a p(doom) around the high end—risks that, if left unchecked, would threaten the whole of humankind.

Probability of doom, especially when written in the mathy-looking format *p(doom)*, sounds like an important and sophisticated metric. Or at least the probability or *p()* part does. But these stark terms are meant to grab headlines and grant an inflated sense of self-importance to those in the room. *Doom* reminds us of titanic, cartoonish fictional battles of good versus evil. And the cartoonish connotations are apt: just like such fictional battles, p(doom) estimates are based in fantasy rather than data or empirical fact. But that hasn't prevented this imaginary metric from becoming a fixation of lawmakers, venture capitalists, and Silicon Valley's managerial class. We imagine part of the appeal is that it allows people in power to imagine themselves as heroes out to save humanity, while actually turning away from the very real threats to actual people.

For example, the probability of techno-enabled doom brought about through automated state violence is very high for some citizens of Detroit. In January 2020, Robert Williams was arrested in front of his two young daughters, when Detroit police trusted the result of a database search of 49 million photos that matched his driver's license photo to a freeze frame from a surveillance video of a theft, committed by someone else, two years earlier. The detectives didn't acknowledge their error until Williams held the printed freeze frame next to his face. In February 2023, Porcha Woodruff was arrested and detained for eleven hours based also on the output of an automated facial recognition system. At the time, Woodruff was eight months pregnant and began to experience contractions while in police custody. The facial recognition system matched her image to footage of a (not visibly pregnant) person stealing a car. Both Williams and Woodruff are Black, and most known false positives for facial recognition tools have involved Black individuals. The probability is quite high that the lives of these people—and a number of other Black residents who have been mistakenly marked as criminal by facial recognition systems—have been irrevocably altered for the worse.

A doomsday scenario has also arrived for teenagers, especially teenage girls, in the form of apps that purport to "undress" a person in an image. These image generation apps automate the task of making deepfake porn, allowing high school students to sexually harass and bully their classmates with a few clicks. The vast majority (99 percent) of deepfakes are of women. The apps can produce such outputs because they are trained on indiscriminately collected troves of images from the internet, datasets that are so enormous no one could possibly verify each individual image in them. The datasets contain a lot of porn, meaning deepfake apps also create nonconsensual images of a sex worker's body. Distressingly, these datasets also include child sexual abuse material.

In 2023, the Israel Defense Forces (IDF) and Prime Minister Benjamin Netanyahu's war-driven cabinet leveraged a system, again marketed as AI, in carrying out their assault on the Gaza Strip, in which tens of thousands of civilians were killed in just the first three months. While the AI system was far from the only ingredient in what the head of the International Committee of the Red Cross and a spokesperson for a United Nations office both called "hell on Earth," it served the purpose of rapidly scaling (and justifying) target selection: using a system called "The Gospel", the IDF dramatically expanded the scope of possible targets to include so-called "power targets", which includes high-rise residential blocks where a single Hamas member may live.

In the words of one former officer, the system facilitates a "mass assassination factory."

But harms befalling real people are not what p(doom) refers to. Despite what many leaders in DC, New York, and Silicon Valley say, p(doom) is the wrong metric and the wrong framing. It serves to obfuscate what's really going on. Artificial intelligence, if we're being frank, is a *con*: a bill of goods you are being sold to line someone's pockets. A few major well-placed players are poised to accumulate significant wealth by extracting value from other people's creative work, personal data, or labor, and replacing quality services with artificial facsimiles. The language of p(doom) is a ruse to keep us focused on imaginary scenarios, filled with awe at modern robber barons' allegedly potentially world-ending technology, and too distracted to see the daily harms being done in its name.

We call this type of con "AI hype". Hype is not particularly new, and in fact we've been through AI hype cycles before. A characteristic of our current hype cycle is that the con men are taking a series of tropes from science fiction—of artificial minds hell-bent on turning us into paper clips or Terminators waging wars for their right to exist (and to look cool on motorcycles)—and injecting them into discussions at the highest echelons of business and government. This framing is useful to those creating the technology because it makes them appear powerful—if not godlike—in their technical creation. But this belies what these technologies are doing to the rest of us: threatening stable careers and replacing them with gig work, slashing personnel in government, cheapening our social services, and degrading creativity.

To successfully navigate this technological moment, make wise choices as individual consumers and institutional decision-makers, and encourage lawmakers to enact smart policy, we argue that this framing needs to be discarded altogether. Not only does the rhetoric around p(doom) distract from actual harms, but the very terminology of "artificial intelligence" impedes clear understanding of the technologies in question, what they can and should be used for, and how to evaluate them. So, in our exploration of AI hype, we must first take a closer look at what people are talking about when they talk about "artificial intelligence".

WHAT IS "AI"?

To put it bluntly, "AI" is a marketing term. It doesn't refer to a coherent set of technologies. Instead, the phrase "artificial intelligence" is deployed when the people building or selling a particular set of technologies will profit from getting others to believe that their technology is similar to humans, able to do things that, in fact, intrinsically require human judgment, perception, or creativity. But even in this case, there has to be a claim to similarity: calculators are far better than people at doing arithmetic, but they aren't sold as "AI". Sometimes the people selling these tools seem to believe their own marketing (we'll meet several examples in later chapters), but what really matters is that they can sell it that way.

Throughout this book, we're going to use the terms "artificial intelligence" or "AI" to refer to technologies sold as such. When speaking about a particular technology, we aim to be as precise as possible. But when referring to these technologies in general, we will sometimes use the shorthand abbreviation of "AI". We want to keep a critical distance from the term: every time we write "AI", imagine we have a set of scare quotes around it. Or if you prefer, replace it with a ridiculous phrase. Some of our favorites include "mathy maths", "a racist pile of linear algebra", "stochastic parrots" (referring to large language models specifically), or Systematic Approaches to Learning Algorithms and Machine Inferences (aka SALAMI).

The set of technologies that get sold as AI is diverse, in both application and construction—in fact, we wouldn't be surprised if some of the tech being sold this way is actually just a fancy wrapper around some spreadsheets. The term serves to obscure that diversity, however, so the conversation becomes clearer if one speaks in terms of "automation" rather than "AI" and looks at precisely *what is being automated*. In doing so, we find several types of automation.

Decision making. The first group involves using computers to automate consequential decisions. These are called automatic decision systems and they are often used, for example, in the process of setting bail, approving loans, screening résumés, or allocating social benefits. These uses are contentious, and rightfully so, because they have extreme ramifications for people who are subject to the system's recommendations.

Classification. The second kind of automation involves classification of inputs of different types. For example, image classification can be used to help consumers organize their photos (where are all the photos of Grandma?), or can be used by governments for surveillance (matching a security footage frame to a database of driver's license photos). The classification of web users for targeted advertising also fits into this group.

Recommendation. A third type selects information to present to someone, based on their own search or purchase history, or searches performed by someone else with a similar profile to them. These systems are called recommender systems. They're behind the ordering of your feed in social media websites, Amazon product recommendations, or movie suggestions on Netflix.

Transcription/Translation. The fourth type is the automatic translation of information from one format to another: automatic transcription (sometimes called "automatic speech recognition" or "speech to text"), finding words and characters in images (like automatically reading license plates), machine translation of one language to another, or something like image style transfer (taking a selfie and making it look like an anime character).

Text and Image Generation. Then finally there's a type that's been very much in everyone's mind recently: so-called generative AI or, more aptly, synthetic media machines. These are systems like ChatGPT, Gemini, or DALL-E that allow users to generate images or plausible-sounding text based on textual prompts. A "prompt", in generative AI terminology, is the words used to describe the desired output.

Lumping all of these different technologies under the label of "AI" creates the illusion of "intelligent" technology: if our photo software's sharpening tool is imagined to be the same thing as the system that appears to cheerfully answer questions on any topic, then both are perceived as even more "intelligent" or even "magical" than each alone, and we're more likely to accept automation in other domains, like deciding who gets social benefits or who is classified as a possible repeat offender. They are all supposedly driven by the same "intelligence". Text synthesis machines have an outsized role here: language is so central to our understanding of each other that when we encounter language that doesn't actually reflect the thoughts, ideas, or communicative intent of another person, it's difficult not to imagine some humanlike mind behind it.

"AI" has always been a marketing term, but it hasn't always been the marketing term of choice. In fact, up until fairly recently, the field was experiencing an "AI winter", a time during which research funding was scarce, and the overall project of building computer systems that mimic human cognition was fairly marginalized within computer science. The companies building and selling such technologies as speech synthesis, automatic transcription, machine translation, image processing, and robotics did not label them as "AI". That all changed in the 2010s, when one particular approach to pattern matching at scale—called "deep learning"—became practical for the first time. This wasn't because of any magic or quantum leap in technology, but for the most part followed from innovation predicated on the falling costs of microchips and the abundance of digitized data on the web, easily accessible through a small set of platforms that centralized data sharing (Flickr, Tumblr, Google, and the like).

Even researchers working on these very approaches were surprised by the rapid switch from "AI winter" to seemingly unlimited venture capital funds. A research conference called Neural Information Processing Systems (NeurIPS, for short) grew from 1,354 attendees in 2010 to 13,000 attendees in 2019 and 22,000 attendees in 2020 (virtual due to COVID). In December 2012, when the conference was held outside of snowy Lake Tahoe (with a relatively sparse attendance of 1,676 people), a researcher named Geoff Hinton, along with his graduate students Alex Krizhevsky and Ilya Sutskever, held a secret auction for their company, DNNresearch. The company had no product, nor any content on its website beyond its name. All it had was a paper that demonstrated their success in deep learning. Four companies—Microsoft, Google, the London-based AI startup DeepMind (later acquired by Google), and the Chinese search engine Baidu—made bids. The day went to Google, however, when Hinton stopped the auction at $44 million. Hinton went on to join Google as a Distinguished Researcher for over a decade, and Sutskever later went on to become a cofounder and chief scientist at another startup, OpenAI. The deep learning era started with a bang, powered by immense amounts of money, capital, and, of course, hype.

WHAT IS HYPE?

Hype is the aggrandizement of some person, artifact, technology, or technique that you, the consumer, absolutely need to buy or invest in as early as possible, lest you miss out on entertainment or pleasure, monetary reward, return on investment, or market share. In the hip-hop world, the hype man is an accessory to the main act, the person who amps up the crowd for their employer. Software developer conferences might seem like the antithesis of hip-hop concerts, but then–Microsoft CEO Steve Ballmer played the hype man at a 1999 Microsoft event. Voice hoarse, visibly sweaty, he pranced around the stage chanting "Developers, developers, developers!" and managed to get his audience of software engineers and managers to pick up that chant, buying into his hype about a mundane software framework.

Hype drives fashion trends, new musical artists, and car purchases. But more critically for this book, it drives investment in startups, technologies, and particular people.

Like other kinds of hype, AI hype plays on FOMO (the fear of missing out): it is the repeated message that a set of technologies—currently, a set of statistical methods developed within computer science and engineering—will change the world and you, the consumer or corporate manager, absolutely must use it, lest you be left in the dust. As a consumer, if you don't get in on the hyped product, you'll be seen as a regressive Luddite, lacking in modern skills, and/or about to have your job automated away. If you're a corporate manager, you have to get on board, or competitors will eat your lunch. If you're a computer programmer, you have to use new tools, otherwise you will be wasting time and won't meet product deadlines. If you're a teacher, you have to incorporate it into your curriculum, lest your students not be prepared for the AI-enhanced workplace. And if you're a student, you have to thoroughly understand AI to take on today's modern workplace, or else you'll get passed over for job opportunities.

The commercial function of tech hype is to boost sales of a product. In other words, marketing. Sam Altman, the CEO of OpenAI, is, like all the great tech barons of our era, an adman. But while all tech hype plays a commercial function, AI hype in particular plays a cultural function as well. It connects a commercial goal with a popular fantasy of sentient machines.

When selling rosy scenarios, AI hype promises us a life of ease: jobs deemed menial like data entry, writing ad copy, and making basic graphics will become a thing of the past. AI "companions" will take notes for you in online meetings or, even better, become your stand-in while you address more pressing matters. Surely technologies of today are just a few rounds of "progress" away from the onboard computer that Captain John-Luc Picard can confidently command to provide "Tea, Earl Grey, Hot" or the caring, competent "operating system" voiced by Scarlett Johansson in *Her*. Altman made this implicit fantasy explicit when he tweeted the single word "her" in advance of a product demo, a voice assistant that sounded suspiciously like Johansson—created without her consent.

But AI hype also depends on promulgating worst-case scenarios. Here, AI hype invokes visions of robots that disobey Isaac Asimov's First Law of Robotics: "A robot may not injure a human being or, through inaction, allow a human being to come to harm." These examples are rife throughout science fiction and myth: the robot HAL 9000, which disobeys the commands of humans in *2001: A Space Odyssey* in order to complete its mission; the machine race that takes over the face of the earth and uses humans as a power source in *The Matrix*; a rogue "Entity" in *Mission Impossible: Dead Reckoning*, which, after being developed by the U.S. government, turns on its masters. The tale is as old as Mary Shelley's *Frankenstein*, about the monster that turns on its creator, or even older, in the Judaic figure of the golem, which in some iterations of the story goes rogue after its human handlers forget to deactivate it.

Claims that we're but a step away from living in a science fiction world have little basis in reality. But just because the hype is ungrounded in the real world doesn't mean the hype itself doesn't impact the world, culturally, economically, and environmentally. And while AI hype has reached a fever pitch in recent years, it has been with us for decades, back to the founding of the field. We can expect AI hype to accompany AI research as long as such research is pursued. A quick tour of the original AI hype will help us see through today's, and comparing AI hype—old and new—will help you identify it in the future, too.

A BRIEF HISTORY OF AI (AND AI HYPE)

As long as there's been research on AI, there's been AI hype. In the most commonly told narrative about the research field's development, mathematician John McCarthy and computer scientist Marvin Minsky organized a summer-long workshop in 1956 at Dartmouth College in Hanover, New Hampshire, to discuss a set of methods around "thinking machines". The term "artificial intelligence" is attributed to McCarthy, who was trying to find a name suitable for a workshop that concerned a diverse set of existing knowledge communities. He was also trying to find a way to exclude Norbert Wiener—the pioneer of a proximate field, cybernetics, a field that has to do with communication and control of machines—due to personal differences.

The way the origin story is told, Minsky and McCarthy convened the two-month working group at Dartmouth, consisting of a group of ten mathematicians, physicists, and engineers, which would make "a significant advance" in this area of research. Just as it is today, the term "artificial intelligence" did not have much coherence. It did include something similar to today's "neural networks" (also called "neuron nets" or "nerve nets" in those early documents), but also covered topics that included "automatic computers" and human-computer language interfaces (what we would today consider to be "programming languages").

Fundamentally, the forerunners of this new field were concerned with translating dynamics of power and control into machine-readable formulations. McCarthy, Minsky, Herbert Simon (political scientist, economist, computer scientist, and eventual Nobel laureate), and Frank Rosenblatt (one of the originators of the "neural network" metaphor) were concerned with developing tools that could be used for the guidance of administrative—and ultimately—military systems. In an environment where the battle for American supremacy in the Cold War was being fought on all fronts—military, technological, engineering, and ideological—these men sought to gain favor and funding in the eyes of a defense apparatus trying to edge out the Soviets. They relied on huge claims with little to no empirical support, bad citation practices, and moving goalposts to justify their projects, which found purchase in Cold War America. These are the same set of practices that we see from today's AI boosters, although they are now primarily chasing market valuations, in addition to government defense contracts.

The first move in the original AI hype playbook was foregrounding the fight with the Soviets. The second was to argue that computers were likely to match human capabilities by arguing that humans weren't really all that complex. In 1956, Minsky claimed in an influential paper that "[h]uman beings are instances of certain kinds of very complicated machines." If that were indeed the case, we could use more controllable electronic circuits in place of people in military and industrial contexts.

In the late 1960s, Joseph Weizenbaum, a German émigré, professor at the Massachusetts Institute of Technology, and contemporary of Minsky, was alarmed by how quickly people attributed agency to automated systems. Weizenbaum developed a chatbot called ELIZA, named for the working-class character in George Bernard Shaw's *Pygmalion* who learns to mimic upper-class speech. ELIZA was designed to carry on a conversation in the style of a Rogerian psychotherapist; that is, the program primarily repeated what its users said, reframing their thoughts into questions. Weizenbaum used this form for ELIZA, not because he thought it would be useful as a therapist, but rather because it was a convenient setup for the chatbot: this kind of psychotherapy is one of the few conversational situations where it wouldn't matter if the machine didn't have access to other data about the world.

Despite its grave limitations, computer scientists used ELIZA to celebrate how thoroughly computers could replace human labor and heralded the entry into the artificial intelligence age. A shocked Weizenbaum spent the rest of his life as a critic of AI, noting that humans were not meat machines, while Minsky went on to found MIT's AI laboratory and rake in funding from the Pentagon unhindered.

The murky, unethical funding networks—through unfettered weapons manufacturing then, and with the addition of ballooning speculative venture capital investments now—around AI continue to this day. So does the drawing of false equivalences between the human brain and the calculating capabilities of machines. Claiming such false equivalences inspires awe, which, it turns out, can be used to reel in boatloads of money from investors whipped into a FOMO frenzy.

When we say boatloads, think megayachts: in January 2023, Microsoft announced that it intended to invest $10 billion in OpenAI. This is after Mustafa Suleyman (former CEO of DeepMind, made CEO of Microsoft AI in March 2024) and LinkedIn cofounder Reid Hoffman received a cool $1.3 billion from Microsoft and chipmaker Nvidia in a funding round to their young startup, Inflection.AI. OpenAI alums cofounded Anthropic, a company solely focused on creating generative AI tools, and received $580 million in an investment round led by crypto-scammer Sam Bankman-Fried. These startups, and a slew of others, have been chasing a gold mine of investment from venture capitalists and Big Tech companies, frequently without any clear path to robust monetization. By the second quarter of 2024, venture capital was dedicating $27.1 billion, or nearly half of their quarterly investments, to AI and machine learning companies.

The incentives to ride the AI hype train are clear and widespread—dress something up as AI and investments flow. But both the technologies and the hype around them are causing harm in the here and now.

OF HYPE AND HARM

There *are* applications of machine learning that are well scoped, well tested, and involve appropriate training data such that they deserve their place among the tools we use on a regular basis. These include such everyday things as spell-checkers (no longer simple dictionary look-ups, but able to flag real words used incorrectly) and other more specialized technologies like image processing used by radiologists to determine which parts of a scan or X-ray require the most scrutiny. But in the cacophony of marketing and startup pitches, these sensible use cases are swamped by promises of machines that can effectively do magic, leading users to rely on them for information, decision-making, or cost savings—often to their detriment or to the detriment of others.

As investor interest pushes AI hype to new heights, tech boosters have been promoting AI "solutions" in nearly every domain of human activity. We're told that AI can shore up threadbare spots in social services, providing medical care and therapy to those who aren't fortunate enough to have good access to health care, education to those who don't live in a wealthy school district, and legal services for people who can't afford a licensed attorney. We're told that AI will provide individualized versions of all of these things, flexibly meeting user needs. We're told that AI will "democratize" creative activity by allowing *anyone* to become an artist. We're told that AI is on the verge of doing science for us, finally providing us with answers to urgent problems from medical breakthroughs (discovering a cure for cancer!) to the climate crisis (discovering a solution for global warming!). And self-driving cars are perpetually just around the corner (watch out: that means they're about to run into you). But as you may have surmised from our snarky tone, these solutions are, by and large, AI hype. There are myriad cases in which AI solutions have been posed but fall short of their stated goals.

In 2017, a Palestinian man was arrested by Israeli authorities over a Facebook post in which he posed next to a bulldozer with the caption (in Arabic) of "good morning." Facebook's machine translation software rendered that as "hurt them" in English and "attack them" in Hebrew—and the Israeli authorities just took that at face value, never checking with any Arabic speakers to see if it was correct. Machine translation has also become a weak stopgap in other critical situations, such as in handling asylum cases. Here, the problem to solve is one of communication, between people fleeing violence in their home countries and immigration officials. Machine translation systems, which can work well in cases like translating newspapers written in standard varieties of a handful of dominant languages, can fail drastically in translating asylum claims written or spoken in minority languages or dialects.

In August 2020, thousands of British students, unable to take their A-level exams due to the COVID-19 pandemic, received grades calculated based on an algorithm that took as input, among other things, the grades that other students at their schools received in previous years. After massive public outcry, in which hundreds of students gathered outside the prime minister's residence at 10 Downing Street in London, chanting "Fuck the algorithm!" the grades were retracted and replaced with grades based on teachers' assessment of student work. In May 2023, Jared Mumm, a professor at Texas A&M University, suspected his students of cheating by using ChatGPT to write their final essays—so he input the essays into ChatGPT and asked it whether it wrote them. After reading ChatGPT's affirmative output, he assigned the whole class incomplete grades, and some seniors were (temporarily) denied their diplomas.

On our roads, promises of self-driving cars have led to death and destruction. A Tesla employee died after engaging the so-called "Full Self-Driving" mode in his Tesla Model 3, which ran the car off the road. (We know this partially because his passenger survived the crash.) A few months later, on Thanksgiving Day 2022, Tesla CEO Elon Musk announced the availability of Tesla's "Full Self-Driving" mode. Hours later, it was involved in an eight-car pileup on the San Francisco–Oakland Bay Bridge.

In 2023, lawyer Steven A. Schwartz, representing a plaintiff in a lawsuit against an airline, submitted a legal brief citing legal precedents that he found by querying ChatGPT. When the lawyers defending the airline said they couldn't find some of the cases cited and the judge asked Schwartz to submit them, he submitted excerpts, rather than the traditional full opinions. Ultimately, Schwartz had to own up to having trusted the output of ChatGPT to be accurate, and he and his cocounsel were sanctioned and fined by the court.

In November 2022, Meta released Galactica, a large language model trained on scientific text, and promoted it as able to "summarize academic papers, solve math problems, generate Wiki articles, write scientific code, annotate molecules and proteins, and more." The demo stayed up for all of three days, while the worldwide science community traded examples of how it output pure fabrications, including fake citations, and could easily be prompted into outputting toxic content relayed in academic-looking prose.

What all of these stories have in common is that someone oversold an automated system, people used it based on what they were told it could do, and then they or others got hurt. Not all stories of AI hype fit this mold, but for those that don't, it's largely the case that the harm is either diffuse or undocumented. Sometimes,