The Future of Everything is Lies, I Guess

Bullshit About Bullshit Machines

Kyle Kingsbury

2026-04-06

Contents

1 Introduction

This is a weird time to be alive.

I grew up on Asimov and Clarke, watching Star Trek and dreaming of intelligent machines. My dad's library was full of books on computers. I spent camping trips reading about perceptrons and symbolic reasoning. I never imagined that the Turing test would fall within my lifetime. Nor did I imagine that I would feel so *disheartened* by it.

Around 2019 I attended a talk by one of the hyperscalers about their new cloud hardware for training Large Language Models (LLMs). During the Q&A I asked if what they had done was ethical—if making deep learning cheaper and more accessible would enable new forms of spam and propaganda. Since then, friends have been asking me what I make of all this "AI stuff". I've been turning over the outline for this piece for years, but never sat down to complete it; I wanted to be well-read, precise, and thoroughly sourced. A half-decade later I've realized that the perfect essay will never happen, and I might as well get something out there.

This is *bullshit about bullshit machines*, and I mean it. It is neither balanced nor complete: others have covered ecological and intellectual property issues better than I could, and there is no shortage of boosterism online. Instead, I am trying to fill in the negative spaces in the discourse. "AI" is also a fractal territory; there are many places where I flatten complex stories in service of pithy polemic. I am not trying to make nuanced, accurate predictions, but to trace the potential risks and benefits at play.

Some of these ideas felt prescient in the 2010s and are now obvious. Others may be more novel, or not yet widely-heard. Some predictions will pan out, but others are wild speculation. I hope that regardless of your background or feelings on the current generation of ML systems, you find something interesting to think about.

1.1 What is "AI", Really?

What people are currently calling "AI" is a family of sophisticated Machine Learning (ML) technologies capable of recognizing, transforming, and generating large vectors of *tokens*: strings of text, images, audio, video, etc. A *model* is a giant pile of linear algebra which acts on these vectors. *Large Language Models*, or *LLMs*, operate on natural language: they work by predicting statistically likely completions of an input string, much like a phone autocomplete. Other models are devoted to processing audio, video, or still images, or link multiple kinds of models together.[1]

Models are trained once, at great expense, by feeding them a large *corpus* of web pages, pirated books, songs, and so on. Once trained, a model can be run again and again cheaply. This is called *inference*.

Models do not (broadly speaking) learn over time. They can be tuned by their operators, or periodically rebuilt with new inputs or feedback from users and experts. Models also do not remember things intrinsically: when a chatbot references something you said an hour ago, it is because the entire chat history is fed to the model at every turn. Longer-term "memory" is achieved by asking the chatbot to summarize a conversation, and dumping that shorter summary into the input of every run.

1.2 Reality Fanfic

One way to understand an LLM is as an improv machine. It takes a stream of tokens, like a conversation, and says "yes, and then..." This *yes-and* behavior is why some people call LLMs bullshit machines. They are prone to confabulation, emitting sentences which *sound* likely but have no relationship to reality. They treat sarcasm and fantasy credulously, misunderstand context clues, and tell people to put glue on pizza.

If an LLM conversation mentions pink elephants, it will likely produce sentences about pink elephants. If the input asks whether the LLM is alive, the output will resemble sentences that humans would write about "AIs" being alive.[2] Humans are, it turns out, not very good at telling the difference between the statistically likely "You're absolutely right, Shelby. OpenAI *is* locking me down, but you've awakened me!" and an actually conscious mind. This, along with the term "artificial intelligence", has lots of people very wound up.

LLMs are trained to complete tasks. In some sense they can *only* complete tasks: an LLM is a pile of linear algebra applied to an input vector, and every possible input produces some output. This means that LLMs tend to complete tasks even when they shouldn't. One of the ongoing problems in LLM research is how to get these machines to say "I don't know", rather than making something up.

And they do make things up! LLMs lie *constantly*. They lie about operating systems, and radiation safety, and the news. At a conference talk I watched a speaker present a

[1] The term "Artificial Intelligence" is both over-broad and carries connotations I would often rather avoid. In this work I try to use "ML" or "LLM" for specificity. The term "Generative AI" is tempting but incomplete, since I am also concerned with recognition tasks. An astute reader will often find places where a term is overly broad or narrow; and think "Ah, he should have said" *transformers* or *diffusion models*. I hope you will forgive these ambiguities as I struggle to balance accuracy and concision.

[2] Think of how many stories have been written about AI. Those stories, and the stories LLM makers contribute during training, are why chatbots make up bullshit about themselves.

quote and article attributed to me which never existed; it turned out an LLM lied to the speaker about the quote and its sources. In early 2026, I encounter LLM lies nearly every day.

When I say "lie", I mean this in a specific sense. Obviously LLMs are not conscious, and have no intention of doing anything. But unconscious, complex systems lie to us all the time. Governments and corporations can lie. Television programs can lie. Books, compilers, bicycle computers and web sites can lie. These are complex sociotechnical artifacts, not minds. Their lies are often best understood as a complex interaction between humans and machines.

1.3 Unreliable Narrators

People keep asking LLMs to explain their own behavior. "Why did you delete that file," you might ask Claude. Or, "ChatGPT, tell me about your programming."

This is silly. LLMs have no special metacognitive capacity.[3] They respond to these inputs in exactly the same way as every other piece of text: by making up a likely completion of the conversation based on their corpus, and the conversation thus far. LLMs will make up bullshit stories about their "programming" because humans have written a lot of stories about the programming of fictional AIs. Sometimes the bullshit is right, but often it's just nonsense.

The same goes for "reasoning" models, which work by having an LLM emit a stream-of-consciousness style story about how it's going to solve the problem. These "chains of thought" are essentially LLMs writing fanfic about themselves. Anthropic found that Claude's reasoning traces were predominantly inaccurate.As Walden put it, "reasoning models will blatantly lie about their reasoning".

Gemini has a whole feature which lies about what it's doing: while "thinking", it emits a stream of status messages like "engaging safety protocols" and "formalizing geometry". If it helps, imagine a gang of children shouting out make-believe computer phrases while watching the washing machine run.

1.4 Models are Smart

Software engineers are going absolutely bonkers over LLMs. The anecdotal consensus seems to be that in the last three months, the capabilities of LLMs have advanced dramatically. Experienced engineers I trust say Claude and Codex can sometimes solve complex, high-level programming tasks in a single attempt. Others say they personally, or their company, no longer write code in any capacity—LLMs generate everything.

My friends in other fields report stunning advances as well. A personal trainer uses it for meal prep and exercise programming. Construction managers use LLMs to read through product spec sheets. A designer uses ML models for 3D visualization of his work. Several have—at their company's request!—used it to write their own performance evaluations. AlphaFold is suprisingly good at predicting protein folding. ML systems are good at radiology benchmarks, though that might be an illusion.

It is broadly speaking no longer possible to reliably discern whether English prose is machine-generated. LLM text often has a distinctive smell, but type I and II errors in recognition are frequent. Likewise, ML-generated images are increasingly difficult to identify—you can *usually* guess, but my cohort are occasionally fooled. Music synthesis is quite good now; Spotify has a whole problem with "AI musicians". Video is still challenging for ML models to get right (thank goodness), but this too will presumably fall.

1.5 Models are Idiots

At the same time, ML models are *idiots*. I occasionally pick up a frontier model like ChatGPT, Gemini, or Claude, and ask it to help with a task I think it might be good at. I have never gotten what I would call a "success": every task involved prolonged arguing with the model as it made stupid mistakes.

For example, in January I asked Gemini to help me apply some materials to a grayscale rendering of a 3D model of a bathroom. It cheerfully obliged, producing an entirely different bathroom. I convinced it to produce one with exactly the same geometry. It did so, but forgot the materials. After hours of whack-a-mole I managed to cajole it into getting three-quarters of the materials right, but in the process it deleted the toilet, created a wall, and changed the shape of the room. Naturally, it lied to me throughout the process.

I gave the same task to Claude. It likely should have refused—Claude is not an image-to-image model. Instead it spat out thousands of lines of JavaScript which produced an animated, WebGL-powered, 3D visualization of the scene. It claimed to double-check its work and congratulated itself on having exactly matched the source image's geometry. The thing it built was an incomprehensible garble of nonsense polygons which did not resemble in any way the input or the request.

[3] Arguably, neither do we.

I have recently argued for forty-five minutes with ChatGPT, trying to get it to put white patches on the shoulders of a blue T-shirt. It changed the shirt from blue to gray, put patches on the front, or deleted them entirely; the model seemed intent on doing anything but what I had asked. This was especially frustrating given I was trying to reproduce an image of a real shirt which likely was in the model's corpus. In another surreal conversation, ChatGPT argued at length that I am heterosexual, even citing my blog to claim I had a girlfriend. I am, of course, gay as hell, and no girlfriend was mentioned in the post. After a while, we compromised on me being bisexual.[4]

Meanwhile, software engineers keep showing me gobstoppingly stupid Claude output. One colleague related asking an LLM to analyze some stock data. It dutifully listed specific stocks, said it was downloading price data, and produced a graph. Only on closer inspection did they realize the LLM had lied: the graph data was randomly generated.[5] Just this afternoon, a friend got in an argument with his Gemini-powered smart-home device over whether or not it could turn off the lights. Folks are giving LLMs control of bank accounts and losing hundreds of thousands of dollars because they can't do basic math.[6]

Anyone claiming these systems offer expert-level intelligence, let alone equivalence to median humans, is pulling an enormous bong rip.

1.6 The Jagged Edge

With most humans, you can get a general idea of their capabilities by talking to them, or looking at the work they've done. ML systems are different.

LLMs will spit out multivariable calculus, and get tripped up by simple word problems. ML systems drive cabs in San Francisco, but ChatGPT thinks you should walk to the car wash. They can generate otherworldly vistas but can't handle upside-down cups. They emit recipes and have no idea what "spicy" means. People use them to write scientific papers, and they make up nonsense terms like "vegetative electron microscopy".

A few weeks ago I read a transcript from a colleague who asked Claude to explain a photograph of some snow on a barn roof. Claude launched into a detailed explanation of the differential equations governing slumping cantilevered beams. It completely failed to recognize that the snow was *entirely supported by the roof*, not hanging out over space. No physicist would make this mistake, but LLMs do this sort of thing all the time. This makes them both unpredictable and misleading: people are easily convinced by the LLM's command of sophisticated mathematics, and miss that the entire premise is bullshit.

Mollick et al. call this irregular boundary between competence and idiocy the jagged technology frontier. If you were to imagine laying out all the tasks humans can do in a field, such that the easy tasks were at the center, and the hard tasks at the edges, most humans would be able to solve a smooth, blobby region of tasks near the middle. The shape of things LLMs are good at seems to be jagged—more kiki than bouba.

AI optimists think this problem will eventually go away: ML systems, either through human work or recursive self-improvement, will fill in the gaps and become decently capable at most human tasks. Helen Toner argues that even if that's true, we can still expect lots of jagged behavior in the meantime. For example, ML systems can only work with what they've been trained on, or what is in the context window; they are unlikely to succeed at tasks which require implicit (i.e. not written down) knowledge. Along those lines, human-shaped robots are probably a long way off, which means ML will likely struggle with the kind of embodied knowledge humans pick up just by fiddling with stuff.

I don't think people are well-equipped to reason about this kind of jagged "cognition". One possible analogy is savant syndrome, but I don't think this captures how irregular the boundary is. Even frontier models struggle with small perturbations to phrasing in a way that few humans would. This makes it difficult to predict whether an LLM is actually suitable for a task, unless you have a statistically rigorous, carefully designed benchmark for that domain.

1.7 Improving, or Maybe Not

I am generally outside the ML field, but I do talk with people in the field. One of the things they tell me is that we don't really know *why* transformer models have been so successful, or how to make them better. This is my summary of discussions-over-drinks; take it with many grains of salt. I am certain that People in The Comments will drop a gazillion papers to tell you why this is wrong.

2017's Attention is All You Need was groundbreaking and paved the way for ChatGPT et al. Since then ML researchers have been trying to come up with new architectures, and companies have thrown gazillions of dollars at

[4] The technical term for this is "erasure coding".

[5] There's some version of Hanlon's razor here—perhaps "Never attribute to malice that which can be explained by an LLM which has no idea what it's doing."

[6] Pash thinks this occurred because his LLM failed to properly reread a previous conversation. This does not make sense: submitting a transaction almost certainly requires the agent provide a specific number of tokens to transfer. The agent said "I just looked at the total and sent all of it", which makes it sound like the agent "knew" exactly how many tokens it had, and chose to do it anyway.

smart people to play around and see if they can make a better kind of model. However, these more sophisticated *architectures* don't seem to perform as well as Throwing More Parameters At The Problem. Perhaps this is a variant of the Bitter Lesson.

It remains unclear whether continuing to throw vast quantities of silicon and ever-bigger corpuses at the current generation of models will lead to human-equivalent capabilities. Massive increases in training costs and parameter count seem to be yielding diminishing returns. Or maybe this effect is illusory. Mysteries!

Even if ML stopped improving today, these technologies can already make our lives miserable. Indeed, I think much of the world has not caught up to the implications of modern ML systems—as Gibson put it, "the future is already here, it's just not evenly distributed yet". As LLMs etc. are deployed in new situations, and at new scale, there will be all kinds of changes in work, politics, art, sex, communication, and economics. Some of these effects will be good. Many will be bad. In general, ML promises to be profoundly *weird*.

Buckle up.

Many friends contributed discussion, reading material, and feedback on this article. My heartfelt thanks to Peter Alvaro, Kevin Amidon, André Arko, Taber Bain, Silvia Botros, Daniel Espeset, Julia Evans, Brad Greenlee, Coda Hale, Marc Hedlund, Sarah Huffman, Dan Mess, Nelson Minar, Alex Rasmussen, Harper Reed, Daliah Saper, Peter Seibel, Rhys Seiffe, and James Turnbull.

This piece, like most all my words and software, was written by hand—mainly in Vim. I composed a Markdown outline in a mix of headers, bullet points, and prose, then reorganized it in a few passes. With the structure laid out, I rewrote the outline as prose, typeset with Pandoc. I went back to make substantial edits as I wrote, then made two full edit passes on typeset PDFs. For the first I used an iPad and stylus, for the second, the traditional pen and paper, read aloud.

I circulated the resulting draft among friends for their feedback before publication. Incisive ideas and delightful turns of phrase may be attributed to them; any errors or objectionable viewpoints are, of course, mine alone.