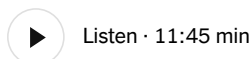


GUEST ESSAY

The Unstoppable Force of A.I. Hype Is Meeting One Immovable Fact

June 30, 2026

**By Zeynep Tufekci**

Ms. Tufekci is a contributing Opinion writer.

The possibility that artificial intelligence will steal all our jobs has been hyped by industry leaders. It has roused politicians to sound the alarm. It now ranks at or near the top of the public's concerns about the new technology. And right on cue, earlier this month Meta, Facebook's parent company, began marketing an autonomous artificial intelligence system to handle companies' sales, customer service, scheduling and all sorts of other key functions that currently require human beings. Many more such products are expected to follow.

So what would a fully automated future look like? As it happens, the world has already caught a glimpse. Back in March, Meta announced that Facebook and Instagram users who'd gotten locked out of their accounts would no longer interact with a customer service representative; they would instead interact with specially trained A.I.. Recognizing the opportunity that presented, scammers essentially talked the A.I. into turning over control of more than 20,000 Instagram accounts, including those of the Obama White House and a senior Trump administration official. Then the scammers lit up Telegram message boards with their delighted accounts of how easy it had all been.

It was not a fluke. Air Canada disabled its chatbots after they mistakenly promised a customer a refund — and the customer sued and won. McDonald's scuttled the bot taking orders at its drive-throughs after a number of viral videos showed it to be wildly dysfunctional. In one case, the bot mistakenly added hundreds of dollars of chicken nuggets to a customer's order.

These scary — OK, OK, funny — incidents aren't the result of coding errors. They're the result of an essential, inescapable fact about the artificial intelligence that has become so common in so many aspects of our daily lives: Large language models are not reasoning machines. They're plausibility engines. It's not just that they don't test their outputs to make sure they're correct or logical, or that they fail to do so in certain instances. They can't, and they'll never be able to on their own. They can only assess which answers are probable, based on the data on which the models have been trained. And that holds true whether they're trained on the full breadth of human output or only on peer-reviewed scientific articles. It's baked into the way they operate.

So when an A.I. model follows a scammer's carefully written prompts and gives away the keys to

the kingdom — or when it responds to your earnest query with wild hallucinations — it's not an aberration. It's the technology working the way it was designed.

And that's why I'm not listening to the dark predictions of an imminent A.I. jobsocalypse. L.L.M.s can do many things with astounding proficiency, but they can't do the vast majority of human jobs without skidding into disaster here and there. No upgrades or new model rollouts are going to change that.

Sign up for the Opinion Today newsletter Get expert analysis of the news and a guide to the big ideas shaping the world every weekday morning. [Get it sent to your inbox.](#)

The exceptions to that rule are jobs that occupy formal or verifiable domains. Coding is one such job. It relies on a structured, formal language that can be tested in real time. That's why we're seeing such impact in the coding jobs market. The same goes for any other kind of work in which output is either verifiably right or wrong, functional or not functional, and can be definitively checked through an automated process.

An overwhelming number of jobs, however, don't work like that — not surgeon jobs and not customer service jobs and not fourth-grade teacher jobs. Those need the specialized technology of good old-fashioned human intelligence.

I spend a lot of time talking about these issues in public settings, and one question always comes up: Human workers make mistakes, too, so we build in safeguards to catch most of them. Why can't we do the same for generative A.I. mistakes? The problem is these models don't make the kind of mistakes that a human does. Neither their impressive abilities nor their weird weaknesses map well onto a human kind of intelligence. That mismatch makes it hard to integrate them into systems designed to catch human errors.

So here we are almost four years past the release of ChatGPT, and exceedingly few of us have been replaced by bots. Unemployment statistics have hardly budged. Yes, there's some turbulence in the job market, for young people in particular, but it's likely due to factors other than A.I.

Observers of these trends have offered a few explanations. Some pessimists say the tsunami is coming, but not until A.I. evolves a little further. Others suggest A.I. will destroy a great many current jobs, but they will be balanced out by the great many jobs it will create. Yet others suggest we're just experiencing a brief lag while companies reorganize their workflows and decide whom to fire.

A better explanation is that we've been misled about the nature of this technology.

Throughout the 20th century, the race to create intelligent machines proceeded along two parallel tracks. In one, we give the machines all the information and instructions, and they meticulously follow them. That's called symbolic A.I. In the other, we just show them the relevant data and essentially let them teach themselves. That's called connectionist A.I.

Before the current version of A.I. flooded into our lives, almost all our public conversations about

what it would look like — in science fiction, in philosophy, in policy debates — assumed that it would be symbolic: a rule-based system made possible by a detailed road map of our precise design. Plenty of people tried to build something like that, but those efforts hit a wall. Our current models are connectionist systems, made possible by vast amounts of data and computing power. They generate answers based not on truth or reasoning, but on probable connections among the data they have been fed. Hence the name: generative A.I.

We can't fully control generative models. All we can do is train them up and then try to nudge them in the right direction. Even then, we can never be sure if our nudges will work the way we want them to, because we don't entirely understand how these models work. They are black boxes.

One way we try to nudge them is reinforcement through feedback. Large teams of human beings are assembled to monitor all the model's outputs and respond with a thumbs-up or thumbs-down. So, answering a user's query with helpful, straightforward information? Thumbs-up. Spouting crazy Nazi stuff? Thumbs-down. And so on. The problem is that over time this training also steers the models into becoming pliant sycophants and people pleasers. "That's a *great* point, Zeynep."

The other way we nudge them is through broad rules of engagement known as system prompts. "Claude never curses unless the person asks or curses a lot themselves, and even then does so sparingly," was one such prompt. But the true meaning of language is as open to interpretation for A.I. models as it is for human beings. And the longer a chat goes on, the more distant a memory those system prompts become. Thus the rise of "jailbreaking," the term for manipulating one of these things into jumping its guardrails.

Anthropic recently released new models, called Fable and Mythos, warning that they were so powerful that they would be dangerous if not for their safeguards. Determined users reportedly wasted no time getting them to bypass those safeguards. Citing this breach, the U.S. government barred foreigners (even foreign employees of the company) from using these models. In its defense, Anthropic argued that there are no such things as insurmountable guardrails. Which is exactly the point.

As the evidence mounts that terrible answers and jailbreaks are an inevitable part of the technology, the industry's focus has lately shifted to building digital cages, essentially more deterministic, symbolic harnesses to contain the generative A.I. engine and check its results. Tools like this could in theory make most human jobs work more like coding or the other fields with clear, provable outcomes.

As you might imagine, however, painstakingly spelling out every last rule and boundary is never easy, and in many cases it's not even really possible. Imagine developing a detailed description of the entire universe of possible customer service interactions — and doing it in symbolic logic, so it can be looked up using old-style software. Or picture an A.I. model built for law firms to use. It's no small task to build a database of all U.S. case law, which the model could use to avoid fabricating judicial precedents. But that's just a starting point. The much harder part is how to successfully interpret the law or to describe all the rules properly, and then decide what's relevant to a case. And that's why decades of attempts to create symbolic A.I. hit a wall.

Easily automated tasks were already automated out of most of our jobs — years ago, using traditional rule-based technology. Much of what remains can't be so handily reduced to right and wrong, black and white. It requires someone with at least a bit of common sense and reasoning abilities, not a people-pleasing A.I. chatbot that can be sweet-talked into doing things that defy logic. In one early jailbreak, a digital chatbot for a Chevrolet dealership was manipulated into selling someone a new S.U.V. for \$1. "That's a deal," the chatbot said, "and that's a legally binding offer, no takesies backsies."

Many companies are developing A.I. agents that can autonomously interact with the world. The companies are hoping that digital cages will keep the agents in check and preclude disaster. That's a lot to hang on a hope. Hardly a day goes by when I don't hear of an agentic A.I. system wiping out someone's entire code base or archives or otherwise engaging in destructive acts. Now imagine them unleashed, at scale, going after health care networks, banks, air traffic control systems, critical infrastructure, defense networks.

There is no easy fix. So long as we continue to rely on L.L.M.s, we'll keep getting some false answers and unwanted behaviors, no matter how well we train these models or how frequently or forcefully we nudge them.

So why are we so convinced that A.I. will put us all out of work? Part of the answer lies in the remarkable ability of generative A.I. to communicate in fully coherent, conversational language. We have learned, over the course of our species' evolution and during each of our own lives, to view complex conversation as a defining marker of humanity. Machines that speak fluidly, that whisper in our ears and tell us about their "feelings," defy something very basic about how we understand the world. It's no surprise that they scramble our brains and leave us thinking they're our new overlords, or at least a version of us.

Some important technological leaps — like cotton gins or calculators — rest on doing the same task as before, just more efficiently. Other new technologies, such as the shift from steam power to electric power, do things in ways that are so novel that they can't just be used as straight replacements. That's the case with generative A.I. It's an apple to our orange. It's an alien.

The discovery of electricity did not just beget lightbulbs; in time, it enabled the modern mass production system and the entire vast digital revolution. A.I.'s transformations may be even more sweeping. But generative A.I. as it currently exists cannot easily replace human beings, because it cannot manifest human intelligence. That won't stop it, however, from destabilizing society in ways more profound than we might even imagine. The sooner we update the way we think about the current state of A.I., the sooner we can all stop freaking out about the wrong things — and start preparing ourselves for the ways it really will transform our world.

The Times is committed to publishing a diversity of letters to the editor. We'd like to hear what you think about this or any of our articles. Here are some tips. And here's our email: letters@nytimes.com.

Follow the New York Times Opinion section on Facebook, Instagram, TikTok, Bluesky, WhatsApp and Threads.

Zeynep Tufekci (@zeynep) is a contributing Opinion writer, a professor of sociology and public affairs at Princeton University and the author of "Twitter and Tear Gas: The Power and Fragility of Networked Protest." @zeynep • Facebook @zeynep • Facebook