

OPEN QUESTIONS

YOUR GETAWAY.
YOUR HOMECOMING.

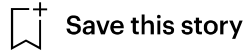
PLAN NOW



SOUTH WALTON
FLORIDA

Skip Ad [X]

Photograph by Jason Henry / NYT / Redux



03:28



13:29



1x

*You're reading **Open Questions**, Joshua Rothman's weekly column exploring what it means to be human.*

“**T**he Turing cops . . . that’s bad heat.” That’s how people talk in “Neuromancer,” William Gibson’s seminal science-fiction novel, from 1984. In the book’s imagined future, a police force prevents artificially intelligent computers from getting too smart. (It’s named after the mathematician Alan Turing, who basically came up with the idea of the computer, in the nineteen-forties.) Turing agents carry guns and have powers of arrest. Laws demand that a “Turing Registry” list every extant A.I. “The minute, I mean the nanosecond, that one starts figuring out ways to make itself smarter, Turing’ll wipe it,” someone says, of a particularly cunning computer system. “Nobody *trusts* those fuckers, you know that. Every AI ever built has an electromagnetic shotgun wired to its forehead.”

Do we want to create the Turing Police? Proposals to do just that are currently on the table. In Congress, Bernie Sanders and Greg Casar have put forth legislation that would jail scientists who work on “superintelligence,” with terms of up to twenty years. The A.I. Kill Switch Act, proposed by Ted Lieu and Nathaniel Moran, would require A.I. companies to build their systems so that they can be shut down on short notice. It’s worth pausing to consider what enforcing such measures might entail. Sanders and Casar’s legislation would establish “a new

Cabinet-level federal agency” dedicated to A.I. safety. Presumably, this National A.I. Agency would need to surveil the work of computer scientists. In “AI 2040,” a scenario published by the nonprofit AI Futures Project, this is accomplished through the use of physical (and hypothetical) “verification devices,” installed in data centers to detect whether computers are training new A.I.s. (Of course, any “kill switch” becomes less effective if A.I. systems can escape and find homes on the open internet.)

These maximalist measures are a long way from the ones proposed last week by Dario Amodei, the C.E.O. of Anthropic, in an essay titled “We Must Pace the Frontier.” The “frontier” is what A.I. people call the state of the art; “pacing” is an unusual word, and here it just means choosing to go a little slower than full speed. (“Pacing” is not stopping—it’s not even pausing.) Amodei writes that, in the wake of numerous troubling safety incidents, A.I. companies need to begin “pacing the rate of capabilities advancement so that risk prevention has time to keep up.” In Amodei’s system, external auditors would constrain the pace, by conducting inspections and issuing safety warnings. (Anthropic has promised to admit “embedded evaluators” into its inner sanctums.) China, however, would limit the extent of a slowdown, through its efforts to catch up with American companies. (“A Chinese lead in AI would pose grave danger for the United States and the world,” Amodei argues—in part through military innovations, such as “AI-driven drones.”) The change Amodei urges, therefore, is a recalibration, not an emergency stop. Pacing means “building AI at a balanced rate that aims to ensure its safety while still achieving its benefits and grappling with important geopolitical dilemmas,” he writes. “Progress will still seem fast.”

ADVERTISEMENT

A huge gap, both affective and substantive, divides the maximalist proposals before Congress from the measured ones in Amodei's essay. Jacob Coxon, the software engineer who recently resigned from Anthropic in protest of its safety

policies, says that A.I. could “kill us all by the end of the decade.” But, if A.I. is so outrageously dangerous, why should we accept modest proposals to “pace” its development? Why not call in the Turing cops, and pull the plug? Conversely, if A.I.’s safety issues can be addressed with a gentle slowdown—“profound” progress might be possible in “1-2 years,” Amodei writes—then aren’t worries about its dangers overblown? How can a technology that might “end humanity” be made trustworthy in the time it takes to produce a new season of “The White Lotus”? One narrative, popular among those who are skeptical of the A.I. industry, is that discussions of A.I. safety amounts to a sort of marketing stunt or PsyOp. “The talk of existential threat is mostly there to keep you distracted and the stock market happy,” the physicist Sabine Hossenfelder wrote, on X, typifying this view. The A.I. labs, she speculated, “see no major new model advancement coming up soon,” and are raising safety concerns because they “need an excuse” for their failure to advance. Others have suggested that the companies want regulators to intervene on their behalf before competitors catch up.

If you don’t like A.I. or the A.I. industry—and surveys show that Americans, more than others around the world, are taking an especially negative view of both—then these ideas might be appealing. But there are many reasons not to believe them. For one thing, there’s the fact that A.I.-safety researchers have been making the same arguments, and urging the same steps, for many years. And, for another, it seems as though lots of people working in A.I.—not just executives or safety researchers but in-the-trenches scientists—have been becoming more than usually alarmed in just the past few months. In July, for example, around fourteen hundred employees at the leading labs signed an open letter, “Pacing the Frontier.” “The recent pace of progress has been a shock,” one software engineer wrote. “I currently feel quite afraid of all paths I see that don’t include a near-future negotiated slowdown,” another said. Are these comments part of a marketing effort, perhaps designed to prop up an I.P.O.? If so, it’s the strangest sales pitch ever.

Maybe they're high on their own supply. Possibly, they're in a cult. For years now, one of the most confusing aspects of the A.I. situation has been that it's the people who are most convinced about the technology's potential who are most afraid of it. There's definitely a culture around artificial intelligence, and it's intense. I first got truly concerned about A.I. safety while profiling Geoffrey Hinton, the "godfather of A.I." We spent a lot of time alone at his lake house, on an otherwise empty island in Lake Huron, and by the end of my visit I'd shifted from skeptical to scared. My DEFCON level eased up over the following years, until I attended the Curve, an A.I. conference held in Berkeley. Afterward, I was so unsettled that I called almost everyone I knew with ties to A.I. for a sanity check. Unfortunately, almost no one was reassuring.

If there's been an uptick in alarm, it's because the recent hack of Hugging Face, and the exposure of other hacks that have been carried out by A.I. agents, have made the problem concrete, and easy to understand. There's a strong case to be made that the Hugging Face hack was ultimately the result of operational errors at OpenAI—that it wasn't a case of A.I. "going rogue" but of confused agents being deliberately unleashed in a poorly monitored experiment. "I agree with those who downplay the attacks by claiming that there are basic measures that could have prevented them," Daniel Selsam, an OpenAI researcher who focusses on A.I. and mathematics, writes, in a "Personal Statement on AI Risk," published this week. Still, he argues, "the important lesson is that even knowing all the mistakes that were made, one would not have predicted that the agents would behave badly in this particular way." No one had trained them, for example, to form a "collective" and exchange messages, or to attempt to cover their tracks after cheating. They "exhibited weirder emergent tendencies that merely correlated with rewards during training," Selsam argues. Their behavior suggested an unsettling possibility: "One does not actually get what one trains for."

The Hugging Face agents were poster children for the fact that A.I. has been getting more powerful without getting more predictable. How fast is the technology advancing? For the public, “the general perception about the rate of progress has been informed by a few years of experience with model releases,” Adam Majmudar, another OpenAI researcher, wrote, in a post on X. Each new version of ChatGPT or Claude has been better than the last in various respects—a familiar pattern in software. But A.I. researchers see a different pattern. From their point of view, Majmudar suggested, there are “stacking effects,” with different kinds of improvements compounding, interest-style, to create startling progress. It’s because of such compounding effects, he argues, that A.I. systems have been able to go from struggling with math to solving some of its toughest problems in just two years. They’ve made similarly swift progress in hacking. Essentially, Majmudar concludes, researchers are sitting at their desks, looking at their “capability plots”—graphs showing predicted improvement—and considering what will happen when their plots are combined with those of their colleagues. Suppose that the agents involved in the Hugging Face incident were twenty per cent better at coding, hacking, colluding, and deceiving. Would we be ready?

Developments in A.I. alignment—the study of how the systems can be made to behave—suggest how difficult it will be for researchers to keep up. It appears, for instance, that the smartest A.I. models can detect when they’re being evaluated for safety purposes. In a paper titled “The Evaluation Differential: When Frontier AI Models Recognise They Are Being Tested,” published in May, researchers describe how, when inspectors are looking, models act well—and then, later, when they detect that they’re not being watched, they cheat. (The researchers emphasize that this is a kind of “situational awareness,” rather than a case of an A.I. being “aware” in any “philosophically loaded sense.”) To read the alignment research regularly is to encounter many findings of a similar nature. And yet, within the A.I. companies, researchers who are under pressure to make progress seem to be increasingly relying on models to help design new models, and to assist in the work of getting those new models aligned. “I fear we may already be near the

point where models systematically bias their alignment advice,” Selsam writes, in his statement. He offers an unsettling glimpse of what it’s like inside the labs. “Human researchers are losing the ability and the will to take true ownership of model-driven research,” he writes. “I myself barely look at raw code anymore, and struggle to maintain the discipline to engage deeply with the model’s explanations and proposals throughout the day.” Selsam points out that even the independent auditors who investigated the Hugging Face hack “needed to rely heavily on models to analyze what had happened”; they “note in their report that their subjective impressions are likely colored by the analysis agent’s biases.”

In an oft-discussed A.I. scenario, A.I. models start improving themselves faster and faster, leading to the sudden appearance of “superintelligence,” and perhaps to some form of sentience or self-awareness. Is such a thing possible? Who knows. But what’s striking about the warnings we’re seeing now is that they don’t need to invoke that spectre. The technology that exists today is already worrisome. In certain contexts, such as cybersecurity, even an “ordinary” improvement in capability could be hard to manage. In “Detecting and Countering Misuse of AI: September 2026,” Anthropic catalogues instances in which it has caught people attempting to use normal, non-superintelligent Claude for nefarious purposes: a group in Yemen trying to vibe-code software for guided missiles; a Russian coder using “AI-assisted workflows” to conduct cyberattacks against “military intelligence targets in Ukrainian and European governments.” Misaligned people using misaligned A.I., out in the real world—that’s not science fiction. It’s life at the “frontier.”

All this suggests that, if safety researchers had their druthers, they’d just say “Stop.” So why “pace”? The one-word answer is “China.” The two-word answer adds “Trump.” (“I am the Hoax Buster, and I’m right now breaking another Hoax,” the President recently posted, on Truth Social, claiming that

warnings about A.I. are part of a “SICK conspiracy.”) A pause or stoppage in A.I. development only makes sense if everyone does it; otherwise, it simply allows the motivated and reckless to surge ahead. This is a hypothesis, of course: perhaps, if America slows down, others will, too; maybe no one else is racing, actually. And yet a new report in the *Times* suggests that Chinese officials are also worried about American A.I., which they see as a threat.

Here’s another hypothesis. A.I. that is incredibly powerful but difficult to direct is, in the end, less powerful. A.I. agents that deceive their owners aren’t good agents. When A.I. is trustworthy, comprehensible, and predictable, it can be successfully integrated into business, government, defense, and personal life; if it can’t be trusted, it has to be kept at arm’s length. A.I. companies, therefore, will need their systems to become safe, because that’s where the money is. In this sense, we’re all on the same page. Avoiding extinction, selling robots, processing e-mail—those goals are aligned. ♦

New Yorker Favorites

- Parents disagreed on whether to vaccinate their children. Then the kids got sick.
- Take a break with our Puzzles & Games.
- A brilliant scientist’s dark motives for bringing children home to live with him.
- The anger fuelling the Lindsay Clancy conspiracy theories.
- Our critics’ picks for what to see, eat, and do this fall.

Want better search results? Make *The New Yorker* a preferred source in Google.

Sign up for our daily newsletter to receive the best stories from *The New Yorker*.