OPEN QUESTIONS

TWO PATHS FOR A.I.

The technology is complicated, but our choices are simple: we can remain passive, or assert control.

By Joshua Rothman

May 27, 2025

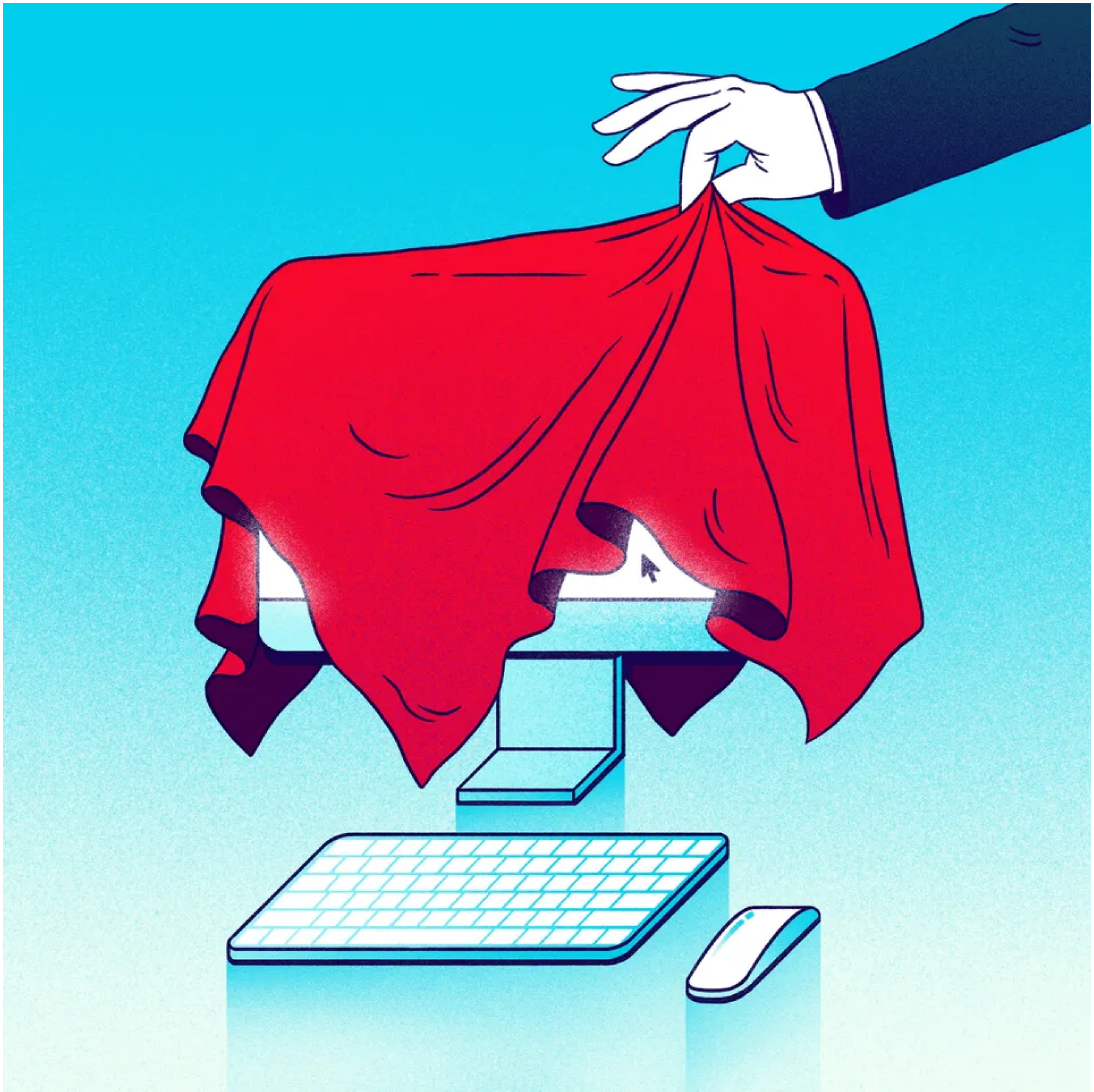

Illustration by Josie Norton

Save this story

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

Last spring, Daniel Kokotajlo, an A.I.-safety researcher working at OpenAI, quit his job in protest. He'd become convinced that the company wasn't prepared for the future of its own technology, and wanted to sound the alarm. After a mutual friend connected us, we spoke on the phone. I found Kokotajlo affable, informed, and anxious. Advances in "alignment," he told me—the suite of techniques used to insure that A.I. acts in accordance with human commands and values—were lagging behind gains in intelligence. Researchers, he said, were hurtling toward the creation of powerful systems they couldn't control.

Kokotajlo, who had transitioned from a graduate program in philosophy to a career in A.I., explained how he'd educated himself so that he could understand the field. While at OpenAI, part of his job had been to track progress in A.I. so that he could construct timelines predicting when various thresholds of intelligence might be crossed. At one point, after the technology advanced unexpectedly, he'd had to shift his timelines up by decades. In 2021, he'd written a scenario about A.I. titled "What 2026 Looks Like." Much of what he'd predicted had come to pass before the titular year. He'd concluded that a point of no return, when A.I. might become better than people at almost all important tasks, and be trusted with great power and authority, could arrive in 2027 or sooner. He sounded scared.

Around the same time that Kokotajlo left OpenAI, two computer scientists at

Princeton, Sayash Kapoor and Arvind Narayanan, were preparing for the publication of their book, "AI Snake Oil: What Artificial Intelligence Can Do, What It Can't, and How to Tell the Difference." In it, Kapoor and Narayanan, who study technology's integration with society, advanced views that were diametrically opposed to Kokotajlo's. They argued that many timelines of A.I.'s future were wildly optimistic; that claims about its usefulness were often exaggerated or outright fraudulent; and that, because of the world's inherent complexity, even powerful A.I. would change it only slowly. They cited many cases in which A.I. systems had been called upon to deliver important judgments—about medical diagnoses, or hiring—and had made rookie mistakes that indicated a fundamental disconnect from reality. The newest systems, they maintained, suffered from the same flaw.

Recently, all three researchers have sharpened their views, releasing reports that take their analyses further. The nonprofit AI Futures Project, of which Kokotajlo is the executive director, has published "AI 2027," a heavily footnoted document, written by Kokotajlo and four other researchers, which works out a chilling scenario in which "superintelligent" A.I. systems either dominate or exterminate the human race by 2030. It's meant to be taken seriously, as a warning about what might really happen. Meanwhile, Kapoor and Narayanan, in a new paper titled "AI as Normal Technology," insist that practical obstacles of all kinds—from regulations and professional standards to the simple difficulty of doing physical things in the real world—will slow A.I.'s deployment and limit its transformational potential. While conceding that A.I. may eventually turn out to be a revolutionary technology, on the scale of electricity or the internet, they maintain that it will remain "normal"—that is, controllable through familiar safety measures, such as fail-safes, kill switches, and human supervision—for the foreseeable future. "AI is often analogized to nuclear weapons," they argue. But "the right analogy is nuclear power," which has remained mostly manageable and, if anything, may be underutilized for safety reasons.

Which is it: business as usual or the end of the world? “The test of a first-rate intelligence,” F. Scott Fitzgerald famously claimed, “is the ability to hold two opposed ideas in the mind at the same time, and still retain the ability to function.” Reading these reports back-to-back, I found myself losing that ability, and speaking to their authors in succession, in the course of a single afternoon, I became positively deranged. “AI 2027” and “AI as Normal Technology” aim to describe the same reality, and have been written by deeply knowledgeable experts, but arrive at absurdly divergent conclusions. Discussing the future of A.I. with Kapoor, Narayanan, and Kokotajlo, I felt like I was having a conversation about spirituality with Richard Dawkins and the Pope.

In the parable of the blind men and the elephant, a group of well-intentioned people grapple with an unfamiliar object, failing to agree on its nature because each believes that the part he’s encountered defines the whole. That’s part of the problem with A.I.—it’s hard to see the whole of something new. But it’s also true, as Kapoor and Narayanan write, that “today’s AI safety discourse is characterized by deep differences in worldviews.” If I were to sum up those differences, I’d say that, broadly speaking, West Coast, Silicon Valley thinkers are drawn to visions of rapid transformation, while East Coast academics recoil from them; that A.I. researchers believe in quick experimental progress, while other computer scientists yearn for theoretical rigor; and that people in the A.I. industry want to make history, while those outside of it are bored of tech hype. Meanwhile, there are barely articulated differences on political and human questions—about what people want, how technology evolves, how societies change, how minds work, what “thinking” is, and so on—that help push people into one camp or the other.

An additional problem is simply that arguing about A.I. is unusually interesting. That interestingness, in itself, may be proving to be a trap. When

"AI 2027" appeared, many industry insiders responded by accepting its basic premises while debating its timelines (why not "AI 2045"?). Of course, if a planet-killing asteroid is headed for Earth, you don't want NASA officials to argue about whether the impact will happen before or after lunch; you want them to launch a mission to change its path. At the same time, the kinds of assertions seen in "AI as Normal Technology"—for instance, that it might be wise to keep humans in the loop during important tasks, instead of giving computers free rein—have been perceived as so comparatively bland that they've long gone unuttered by analysts interested in the probability of doomsday.

When a technology becomes important enough to shape the course of society, the discourse around it needs to change. Debates among specialists need to make room for a consensus upon which the rest of us can act. The lack of such a consensus about A.I. is starting to have real costs. When experts get together to make a unified recommendation, it's hard to ignore them; when they divide themselves into duelling groups, it becomes easier for decision-makers to dismiss both sides and do nothing. Currently, nothing appears to be the plan. A.I. companies aren't substantially altering the balance between capability and safety in their products; in the budget-reconciliation bill that just passed the House, a clause prohibits state governments from regulating "artificial intelligence models, artificial intelligence systems, or automated decision systems" for ten years. If "AI 2027" is right, and that bill is signed into law, then by the time we're allowed to regulate A.I. it might be regulating us. We need to make sense of the safety discourse now, before the game is over.

Artificial intelligence is a technical subject, but describing its future involves a literary truth: the stories we tell have shapes, and those shapes influence their content. There are always trade-offs. If you aim for reliable, levelheaded conservatism, you risk downplaying unlikely possibilities; if you bring imagination to bear, you might dwell on what's interesting at the expense of what's likely. Predictions can create an illusion of predictability that's unwarranted in a fun-house world. In 2019, when I profiled the science-fiction novelist William Gibson, who is known for his prescience, he described a moment of panic: he'd thought he had a handle on the near future, he said, but "then I saw Trump coming down that escalator to announce his candidacy. All of my scenario modules went 'beep-beep-beep.' " We were veering down an unexpected path.

"AI 2027" is imaginative, vivid, and detailed. It "is definitely a prediction," Kokotajlo told me recently, "but it's in the form of a scenario, which is a particular kind of prediction." Although it's based partly on assessments of trends in A.I., it's written like a sci-fi story (with charts); it throws itself headlong into the flow of events. Often, the specificity of its imagined details suggests their fungibility. Will there actually come a moment, possibly in June of 2027, when software engineers who've invented self-improving A.I. "sit at their computer screens, watching performance crawl up, and up, and up"? Will the Chinese government, in response, build a "mega-datacenter" in a "Centralized Development Zone" in Taiwan? These particular details make the scenario more powerful, but might not matter; the bottom line, Kokotajlo said, is that, "more likely than not, there is going to be an intelligence explosion, and a crazy geopolitical conflict over who gets to control the A.I.s."

It's the details of that "intelligence explosion" that we need to follow. The scenario in "AI 2027" centers on a form of A.I. development known as "recursive self-improvement," or R.S.I., which is currently largely hypothetical. In the report's story, R.S.I. begins when A.I. programs become capable of doing A.I. research for themselves (today, they only assist human researchers); these A.I. "agents" soon figure out how to make their descendants smarter, and those descendants do the same for their descendants, creating a feedback loop. This process accelerates as the A.I.s start acting like co-workers, trading messages and assigning work to one another, forming a "corporation-within-a-corporation" that repeatedly grows faster and more effective than the A.I. firm in which it's ensconced. Eventually, the A.I.s begin creating better descendants so quickly that human programmers don't have time to study them and decide whether they're controllable.

Seemingly every science-fiction novel ever written about A.I. suggests that implementing recursive self-improvement is a bad idea. The big A.I. companies identify R.S.I. as risky, but don't say that they won't pursue it; instead, they vow to strengthen their safety measures if they head in that direction. At the same time, if it works, its economic potential could be extraordinary. The pursuit of R.S.I. is "definitely a choice that people are eager to make in these companies," Kokotajlo said. "It's *the* plan. OpenAI and Anthropic, their plan is to automate their own jobs first."

Could this type of R.S.I. work? (It's never been done.) Doesn't it depend on other technological factors—such as "scaling," the ability of A.I. to improve as more computing resources are dedicated to it—which have held true in the past, but might falter in the future? (Some observers think it might already be faltering.) If R.S.I. took hold, would its progress hit a ceiling, or continue until the advent of "artificial superintelligence"—a level of intelligence that exceeds what human minds are capable of? ("It would be a very strange coincidence if the limit on intelligence happened to be just barely above the human range," Kokotajlo said.)

The possibilities compound. Would superintelligence-driven innovation inspire a militarized arms race? Could superintelligent A.I.s end up manipulating or eliminating us while pursuing their own inscrutable ends? (In "AI 2027," they use up the Earth's resources while conducting scientific research we're not smart enough to understand.) Or, in a happier development, might they solve the alignment problem for us, either domesticating themselves or becoming benevolent gods, depending on your point of view?

No one really knows for sure. That's partly because A.I. is a fractious and changing field, in which opinions differ; partly because so much of the latest A.I. research is proprietary and unpublished; and partly because there can be no firm answers to fundamentally speculative questions—only probabilities. "AI 2027" unfolds with a confidence and narrative drive that belie the uncertainties inherent to its subject. The degree to which the scenario depends on a chain of optimistic technological predictions is arguably a flaw, perhaps a major one. (An informed friend associated the report's views with "A.I.-pilled yea-sayers.") But, actually, partiality is one of the reasons that scenarios are valuable. In any uncertain situation, we tend to regard the possibilities we hope won't come to pass in a more hypothetical light. But, for as long as we're reading it, a scenario forces us to at least try to believe in its reality. "AI 2027," Kokotajlo told me, is "not wildly different" from what's talked about "in cafeteria conversations at these companies." They talk about it; now we're imagining it. Are *they* imagining it? Are they taking it seriously enough that, if presented with an important choice about R.S.I., they'll make a wise one?

Kokotajlo says they're not. One widespread misapprehension about artificial intelligence is that dangerous or uncontrollable technology might simply "emerge," without human intervention. ("They say it got smart," someone says, of Skynet, in "The Terminator.") But "AI 2027" portrays a string of affirmatively bad decisions, beginning with the choice, by researchers, to build self-improving A.I. before they have fully figured out how to look inside it and interpret its thoughts. The scenario asserts that, for reasons of competition and curiosity, people working in A.I. will actively seek to do what anyone who's seen "WarGames" could tell them not to. "If you work for these companies, and you talk to them about what they want to do, which is what I did, they tell you that they're going to do it," Kokotajlo told me. "They know that they don't have interpretability solved—that they can't rigorously check the internal goals, or rigorously predict how the A.I. systems will behave in the future. But they're moving ahead anyway." "AI 2027" is partly a tech scenario, and partly a people scenario. It suggests that it's the A.I. companies that are misaligned.

Unlike "AI 2027," "AI as Normal Technology" has an East Coast sensibility. It's a dry, conservative white paper, and draws much of its authority from knowledge of the past. Narayanan and Kapoor aren't too concerned about superintelligence or a possible intelligence explosion. They believe that A.I. faces "speed limits" that will prevent hyper-rapid progress, and argue that, even if superintelligence is possible, it will take decades to invent, giving us plenty of time to pass laws, institute safety measures, and so on. To some extent, the speed limits they discern have to do with A.I. in particular—they flow from the high cost of A.I. hardware, the dwindling supply of training data, and the like. But Kapoor and Narayanan also think they're inherent to technology in general, which typically changes the world more slowly than people predict.

The understandable focus of A.I. researchers on "intelligence," Kapoor and Narayanan argue, has been misleading. A harsh truth is that intelligence alone is of limited practical value. In the real world, what matters is power—"the ability to modify one's environment." They note that, in the history of innovation, many technologies have possessed astonishing capabilities but failed to deliver much power to their inventors or users. It's incredible, for instance, that some cars can drive themselves. But, in the United States, driverless cars are confined to a handful of cities and operated, as robo-taxis, by a small number of companies. The technology is capable, but not powerful. It will probably transform transportation—someday.

Artificial-intelligence researchers often worry about A.I., in itself, becoming too powerful. But Kapoor and Narayanan prefer a human-centered way of thinking: the point of technology is not to become powerful but to empower us. "Humans have always used technology to increase our ability to control our environment," they write, and even wildly capable technologies have empowered us only slowly. New inventions take a long time to "diffuse" through society, from labs outward. "AI 2027" entertains the possibility of "cures for most diseases" arriving as soon as 2029. But, according to Kapoor and Narayanan's view, even if the intellectual work of creating those cures could be rapidly accelerated through A.I., we would still have to wait a long time before enjoying them. Similarly, if an A.I. system speeds the invention of a lifesaving medical device, that device must still be approved by the Food and Drug Administration. Suppose that a superintelligent A.I. solves fusion power—the technology must still be tested, and a site for a proposed plant must be located, with willing neighbors. (The nuclear power plant constructed most recently in the United States, in Waynesboro, Georgia, took fourteen years to build and ran nearly twenty billion dollars over budget.) "My favorite example is Moderna," Kapoor told me, referring to the pharmaceutical company. After Chinese researchers sequenced the genome of SARS-CoV-2, the virus which causes COVID-19, it took Moderna "less than a week to come up with the vaccine. But then it took about a year to roll it out." Perhaps A.I. could design vaccines even faster—but clinical trials, which depend on human biological processes, simply take time.

The view that increases in intelligence will lead quickly and directly to technological outcomes, Narayanan told me, reflects a general underestimation, among coders, of "domain-specific" complexity and expertise. "Software engineering, even though it has engineering in the name, has a history of being disconnected from the rest of engineering," he said. This means that A.I.-safety researchers might also be undervaluing the systems that are already keeping us safe. Kapoor and Narayanan concentrate in particular on the practices of industrial safety, which have been developed and proved over decades. In a factory, fail-safes and circuit breakers insure that systems default to harmless behaviors when they malfunction. (Machines, for instance, may shut down if carbon-monoxide levels rise, or if they detect a person inside them.) Redundancy allows managers to see when a single widget is producing an unusual result. Processes like "formal verification"—in which systems are subjected to carefully designed rules that promote safety—are often used when human beings work alongside complex machines.

The world, in this view, is already a pretty well-regulated place—and artificial intelligence will have to be integrated slowly into its web of rules. One question to ask is, Do we believe that those in charge of A.I. will have to follow the rules? Kapoor and Narayanan note "one important caveat" to their analysis: "We explicitly exclude military AI . . . as it involves classified capabilities and unique dynamics that require a deeper analysis." "AI 2027," meanwhile, is almost entirely focussed on the militarization of artificial intelligence, which unfolds quickly once its defense implications ("What if AI undermines nuclear deterrence?") make themselves known. The two reports, taken together, suggest that we should keep a close watch on military applications of A.I. "AI as Normal Technology," for its part, offers concrete advice for those in charge in many areas of society. Don't wait, passively, for A.I. firms to "align" their models. Instead, start monitoring the use of A.I. in your field. Find ways to track evidence of its risks and failures. And shore up, or create, rules that will make people and institutions more resilient as the technology spreads.

"Deep differences in worldviews": that seems about right. But what is a world view, ultimately? World views are often reactive. We formulate them in response to provocations. Artificial intelligence has been unusually provocative. It has prompted reflections on the purpose of technology, the nature of progress, and the relationship between inventors and the rest of us. It's been a Rorschach test. And it's also arrived at a particular moment, in a particular discursive world, in which opinions are strong, objections are instant, and differences are emphasized. The dynamics of intellectual life lead to doubling down and digging in. We have feedback loops, too.

Is there a single world view that could encompass the perspectives in "AI 2027" and "AI as Normal Technology?" I suspect there could be. Imagine walking onto a factory floor. A sign reads "Safety first!" Workers wear hard hats and high-viz safety gear. The machines don't run the factory; instead, the workers manipulate the machines, which have been designed with both productivity and workers' safety in mind. In this cognitive factory, serious thought has gone into best practices. A lot of emphasis is placed on quality control. A well-funded maintenance team inspects the machines and modifies them as necessary, to meet the factory's requirements. Over in the R. & D. department, scientists sometimes invent promising upgrades. But, before those upgrades are integrated into the production line, they are thoroughly vetted, and the workers are consulted. The factory, moreover, has a mission. Its workers know what they're trying to produce. They don't just ship out whatever the machines happen to make. They steer the machines toward a well-understood goal.

A lot of us may soon find ourselves working on cognitive factory floors. Whatever we do, we could be doing it alongside, or with, machines. Since the machines can automate some of our thinking, it will be tempting to take our hands off the controls. But in such a factory, if a workplace accident occurs, or if a defective product is sold, who will be accountable? Conversely, if the factory is well run, and if its products are delightful, then who will get the credit?

The arrival of A.I. can't mean the end of accountability—actually, the reverse is true. When a single person does more, that person is responsible for more. When there are fewer people in the room, responsibility condenses. A worker who steps away from a machine decides to step away. It's only superficially that artificial intelligence seems to relieve us of the burdens of agency. In fact, A.I. challenges us to recognize that, at the end of the day, we'll always be in charge. ♦

Joshua Rothman, a staff writer, authors the weekly column Open Questions. He has been with the magazine since 2012.

More: Artificial Intelligence Technology Danger Safety

READ MORE

UNDER REVIEW

Can Sam Altman Be Trusted with the Future?

The C.E.O. of OpenAI helped usher artificial intelligence into public life. Now, as fears and fortunes mount, his own transformation is just beginning.

By Benjamin Wallace-Wells

THE LEDE

Building Drones—for the Children?

The influential venture capitalist Katherine Boyle is making the case that creating things for America—from weapons to rockets to nuclear-energy plants—is pro-family.

By Emma Green

A CRITIC AT LARGE

R.F.K., Jr., Anthony Fauci, and the Revolt Against Expertise

It used to be progressives who distrusted the experts. What happened?

By Daniel Immerwahr

PROFILES

Curtis Yarvin’s Plot Against America

The reactionary blogger’s call for a monarch to rule the country once seemed like a joke. Now the right is ready to bend the knee.

By Ava Kofman

FAULT LINES

Elon Musk's Vanishing Act

Musk looks like the latest victim of a common Trump-era dynamic: the impossibility of sharing the President's spotlight.

By Jon Allsop

THE POLITICAL SCENE

Donald Trump's Politics of Plunder

The greed of the new Administration has galvanized America's aspiring oligarchs—and their opponents.

By Evan Osnos

SECOND READ

What Casey Means and MAHA Want You to Fear

Amid the destruction of America's public-health systems, Trump's Surgeon General nominee believes that your wellness is yours alone to defend.

By Jessica Winter

ANNALS OF HOLLYWOOD

How I Learned to Become an Intimacy Coördinator

At a sex-choreography workshop, a writer discovered a world of Instant Chemistry exercises, penis pouches, and nudity riders to train for Hollywood's most controversial job.

By Jennifer Wilson

DEPT. OF PSYCHOPHARMACOLOGY

This Is Your Priest on Drugs

Dozens of religious leaders experienced magic mushrooms in a university study. Many are now evangelists for psychedelics.

By Michael Pollan

INFINITE SCROLL

Sam Altman and Jony Ive Will Force A.I. Into Your Life

Sam Altman and Jony Ive Will Force A.I. Into Your Life

The founder of OpenAI and the designer behind the iPhone are teaming up on a gadget that they promise to ship out "faster than any company" ever has. What could go wrong?

By Kyle Chayka

THE LEDE

The Farmers Harmed by the Trump Administration

Four months ago, the government cut funding to agricultural labs. Kansas farmers and researchers say they can see the damage.

By Peter Slevin

U.S. JOURNAL

Green-Wood Cemetery's Living Dead

How the "forever business" is changing at New York City's biggest graveyard.

By Paige Williams