UNDER REVIEW

MAYBE WE ALREADY HAVE RUNAWAY MACHINES

A new book argues that the invention of states and corporations has something to teach us about A.I. But perhaps it's the other way around.

By Gideon Lewis-Kraus

November 30, 2023

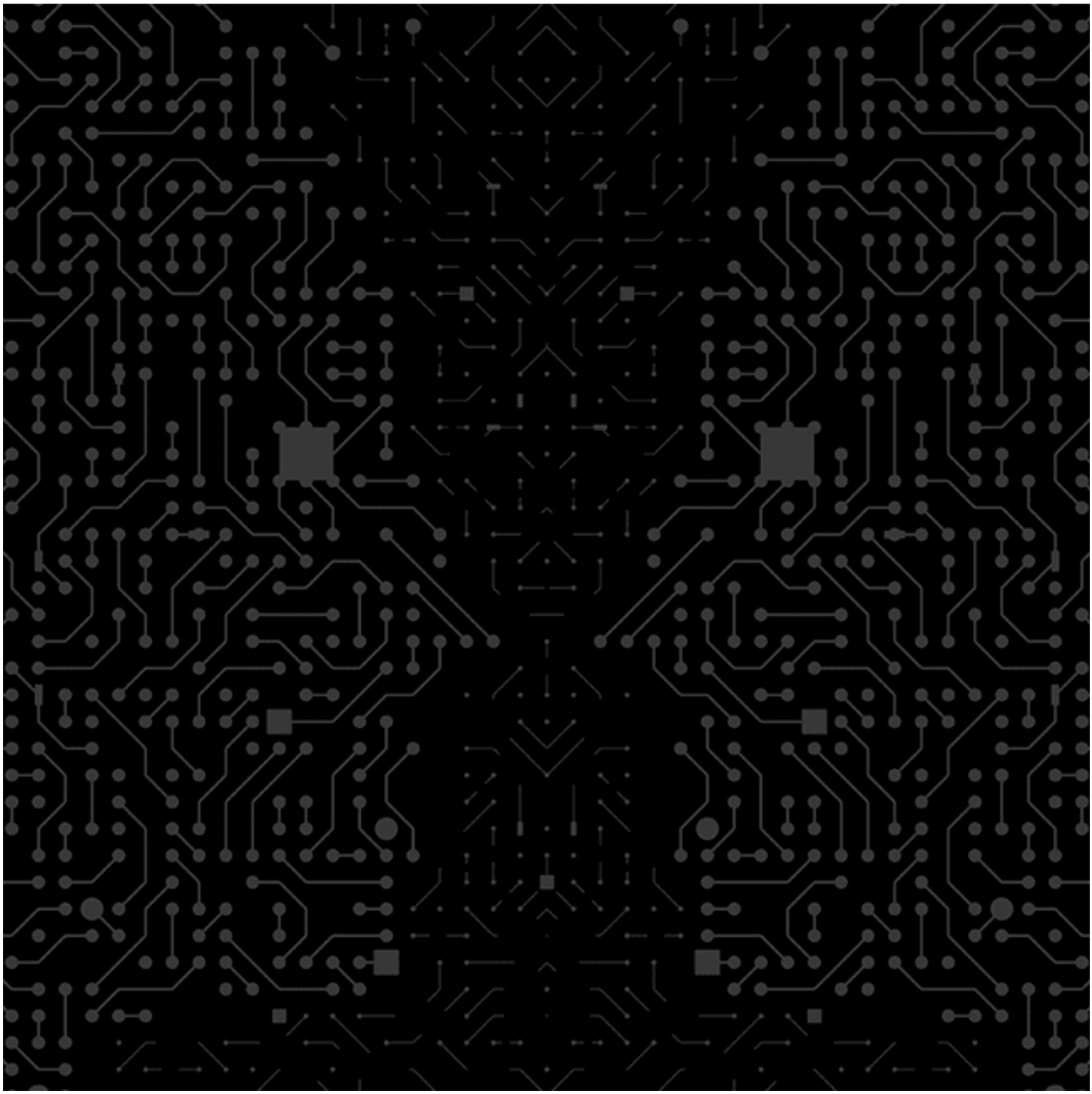


Illustration by Shira Inbar

Save this story

Most of us aren't quite sure how we're supposed to feel about the dramatic improvement of machine capabilities—the class of tools and techniques we've collectively labelled, in shorthand, artificial intelligence. Some people can

barely contain their excitement. Others are, to put it mildly, alarmed. What proponents of either extreme have in common is the conviction that the rise of A.I. will represent a radical discontinuity in human history—an event for which we have no relevant context or basis of comparison. If this is likely to be the case, nothing will have prepared us to assimilate its promise or to fortify ourselves against the worst outcomes. Some small fraction of humanity is, for whatever reason, predisposed to stand in awe or terror before the sublime. The rest of us, however, would prefer to believe that the future will arrive in a more orderly and sensible manner. We find a measure of comfort in the notion that some new, purportedly sui-generis thing bears a strong resemblance to an older, familiar thing. These explanations are deflationary. Their aim is to spare us the need to have neurotic feelings about the problems of tomorrow. If such hazards can be redescribed as merely updated versions of the problems of today, we are licensed to have *normal* feelings about them instead.

This isn't to say that these redescriptions invariably make us feel good, but some of them do. When these arguments take up the sensationalized threat associated with a particular technology, they can be especially reassuring. In this magazine, Daniel Immerwahr recently proposed that the history of photography ought to make us feel better about the future of deepfakes: the proliferation of photographic imagery was coextensive with the proliferation of such imagery's manipulation, and for the most part we've proved ourselves able to interpret photographic evidence with the proper circumstantial understanding and scrutiny —at least when we feel like it. Immerwahr's argument isn't that there's nothing to see here, but that most of what there is to see has been seen before, and so far, at least, we've managed to muddle our way through. This is an inferential claim, and is thus subject to the usual limitations. We can't be certain that human societies won't be freshly destabilized by such state-of-the-art sophistry. But the role of a historian like Immerwahr is to provide us with a reference class for rough comparison, and with a reference class comes something like hope. The likelihood

of technological determinism is counterposed with a reasonable, grounded faith that we have our wits about us.

There is another kind of deflationary argument, one more often associated with the broader and more generic threat of machine intelligence rather than with a discrete technology like deepfakes. The logical premise here is not the precedent of human savvy but an analogy to human fallibility. This might be called the "Since when were we so great?" approach. ChatGPT hallucinates a lot of nonsense? Our own brains hallucinate a lot of nonsense. Facial recognition misidentifies criminal suspects? Eyewitnesses routinely misidentify criminal suspects. There are more and less sophisticated versions of these claims. One liability of this way of thinking is that an emphasis on our own crappiness, alongside reasons of convenience, laziness, or technological fetish, might function as an alibi—an excuse to grant unearned and unexamined authority to machines. The adverse consequences of this attitude include runaway feedback loops: below-average pilots crash planes, we compensate with an increasing reliance on automated systems, the automated systems make allowances for even worse pilots (who lack the skill or training to recover the craft when the automated system fails). But, on the whole, the contrast between inscrutable machines and terminal human mediocrity is once again supposed to be soothing. Machines might make mistakes, but these mistakes aren't really qualitatively different from the mistakes we ourselves make. Machines, one might add, are subject to iterative engineering. Once we give up our irrational bias in favor of human rationality—once we realize that we should prefer self-driving cars that get into occasional horrific crashes to human-driven cars that get into frequent horrific crashes—we'll resign ourselves to our dependence on these tools in the way that we reconciled ourselves to reliance on the plow.

Both of these modes of consolation—by precedent or by analogy—rest on the assumption that history is continuous, that change is more or less gradual and

invisible to those experiencing it, and that humanity will remain in a position to address new complications in the way it has addressed, with varying degrees of success, prior ones. All of these assumptions may very well prove correct, and it's not at all unreasonable to make them, or to find them persuasive. They provide a warrant to worry about the things we're used to worrying about rather than the things we don't even know how to worry about. But they ignore, elide, or diminish the possibility, however far-fetched, that we might have to confront a genuine discontinuity.

Get early access to new short stories, and more from the literary world, when you sign up for our Books & Fiction newsletter.

gideonlk@gmail.com

SIGN UP

By signing up, you agree to our user agreement (including class action waiver and arbitration provisions), and acknowledge our privacy policy.

The crux of the residual case for doom is the alignment problem—the prospect that an artificial superintelligence might have no use for human values. (These values are usually under-specified but might include the belief that life is sacrosanct, humans have dignity, etc.) An advanced A.I. wouldn't even need to be hostile; indifference is equally unfortunate. You can give an advanced black box of a machine a clear directive and find that it executes its goals without sentiment: you ask a housecleaning robot to reduce untidiness and it kills your dog. One of the ways to relieve oneself of this apocalyptic concern is to decide that intelligence and values are necessarily correlated, that an entity of sufficient intellect would perforce absorb our ethical concerns. In this case, however, no argument by continuity—by precedent or analogy—could possibly make us feel better: we are acquainted with the existence of murderous sociopaths. Murderous sociopaths are a kind of existence proof of what people in the A.I.-safety community call the orthogonality thesis—the idea that intelligence and morality exist in different

dimensions. For the past few hundred thousand years, we've been able to contain murderous sociopaths before they confronted us with total annihilation. But a machine superintelligence that takes the form of a murderous sociopath, or becomes in any case allied to a conventional murderous sociopath, might very well command resources of an unprecedented order. It would be both genuinely bad and genuinely new. Such an existential threat might seem wildly implausible or remote. But it, by definition, would seem to flout any attempts to provide historical easement.

At first glance, David Runciman's "The Handover: How We Gave Control of Our Lives to Corporations, States, and AIs" seems poised to furnish just such a reassurance—to domesticate the alignment problem. Runciman, a witty and refined writer, is a professor of politics at Cambridge and a contributing editor to the *London Review of Books*, which has also provided a home for his enlivening podcasts. He is also the Fourth Viscount Runciman of Doxford, a title that inspires not inconsiderable faith in his commitment to historical continuity. Runciman's basic argument, which unfolds in the elegantly shaggy manner of a Peripatetic seminar, is that the alignment problem is not in fact an anomaly, and that the coming singularity might best be historicized as the Second Singularity. Runciman identifies a precedent and an analogy for the alignment problem in our relationship to two other artificial agencies: the state and the corporation. This idea, in his account, first emerged in 1651, when Thomas Hobbes described the Leviathan as, in Runciman's words, "a kind of robot." Hobbes, of course, didn't call it a robot—a term that first appeared in the nineteen-twenties—but he did elaborate his vision of the state as an "Artificiall Man; though of greater stature and strength than the Naturall, for whose protection and defence it was intended." As Runciman glosses the concept, "We assemble it in the same way we might construct any other machine of moving parts. This one is designed to resemble its creators. But it can do things we cannot. It is much more powerful than we are. That's why we built it in the first place." The state has the supra-human advantages of durability and reliability, and serves as a means of coördination.

Runciman continues, "The Leviathan is not a problem-solving machine, like a smartphone. It is a decision-making mechanism, like a set of dice. It might not give you the right answer—it will simply give you *an* answer." In theory, these answers reflect the will of the polity. "But because it comprises people," he writes, "once the people get better at coming up with answers, the state should too. These dice can think for themselves." In practice, however, Hobbes is generally regarded as a pessimist. Of the Leviathan, Runciman writes: "Its argument assumes that humans are incapable of peaceful co-existence without franchising out our decision-making power to an unaccountable higher body. At that point, we will have lost control of our fate, because whatever the state decides becomes our choice too." In Hobbes's time—of the well-known "war of all against all"—these powers were arbitrary and absolute, but this was the price to be paid for security. Runciman readily concedes that we have spent the past few hundred years constructing better mechanisms for democratic accountability, but "the basic machinery of the modern state continues to bear the hallmark of a mechanistic understanding of politics set out nearly four centuries ago. We still have governments that take decisions for us on matters of life and death, including war and peace." He continues,

> We are bound by those decisions whether we like them or not. The state has the power to enforce its choices against anyone who resists, though it may not choose to use it. Most of the time, in contemporary democracies, these arbitrary powers are sufficiently buried beneath layers of legal limitations and political pushback that we barely notice them. But sometimes, in a crisis, they reveal themselves. Then we must acknowledge that we are still living with the Leviathan—older, wiser, but recognisably the same as before.

Without belaboring the comparison, Runciman elaborates a case for the state, which binds human reason in a lattice of rules and procedures, as an algorithmic device. The twinned essential functions of the state—the ability to issue debt and the prerogative of war—have undergirded its durability and afforded it a capacity

for long-term planning that has thus far eluded other forms of human organization. For those of us privileged enough to escape the brutality of the state in the service of imperialist extraction, the state has considerably enhanced our longevity, wealth, and happiness. But the state, even if it comprises the human, is not itself human. "Our ability to fulfil our potential—to live longer, richer, more varied lives—was made possible only by handing over some of our capabilities to artificial versions of ourselves, which do not suffer from our natural frailties. We have had to give up features of our essential humanity—some of what Arendt would call our capacity for independent action—to enjoy these other benefits," Runciman writes.

So it should come as no surprise that the state cannot be counted upon to demonstrate a commitment to human values. It has an artificial life of its own, even as it acts through us—when we are asked to kill in wartime, we are granted leave from human values. States are "persons without scruples," Runciman notes, and we have had hundreds of years to grow familiar with this contrivance and its discontents. The machine-learning era is, in some sense, merely an extension of the mechanical one. "Let's not kid ourselves that the age of AI—with the looming challenges of black-box decision-making and algorithmic procedures whose outcomes are a mystery even to their creators—poses a unique challenge for human understanding. Just look at your own life. Do you fully understand where the group decisions come from that shape who you are?," Runciman writes. "The fact that groups are made out of people doesn't make them any easier to see inside than other kinds of machines. The black boxes are all around us already. In one sense, at least, they *are* us."

One of the things that make the machine of the capitalist state work is that some of its powers have been devolved upon other artificial agents—corporations. Where Runciman compares the state to a general A.I., one that exists to serve a variety of functions, corporations have been granted a limited range of autonomy

in the form of what might be compared to a narrow A.I., one that exists to fulfill particular purposes that remain beyond the remit or the interests of the sovereign body. Corporations can thus be set up in free pursuit of a variety of idiosyncratic human enterprises, but they, too, are robotic insofar as they transcend the constraints and the priorities of their human members. The failure mode of governments is to become "exploitative and corrupt," Runciman notes. The failure mode of corporations, as extensions of an independent civil society, is that "their independence undoes social stability by allowing those making the money to make their own rules." There is only a "narrow corridor"—a term Runciman borrows from the economists Daron Acemoglu and James A. Robinson—in which the artificial agents balance each other out, and citizens get to enjoy the sense of control that emerges from an atmosphere of freedom and security. The ideal scenario is, in other words, a kludgy equilibrium.

Yet we citizens of modern states have always labored under the shadow of a partly mitigated alignment problem—a "mismatch between the drives of these artificial persons and the needs of the planet," as Runciman describes it—that provides a frame, a vocabulary, and a sense of foreboding as we seek to process the automation on the horizon. The concern about unaligned machines is that even if we can program, so to speak, their ultimate ends, we can't necessarily anticipate the instrumental subgoals they might pursue as an intermediate measure. If you instruct a machine to complete a task, the likeliest instrumental subgoal is to "gain control." If the danger of the alignment problem seems indistinct or preposterous, Runciman suggests, you haven't been paying attention.

So far, so continuous. This is the point where recourse to precedent is supposed to get reassuring: given that this putatively newfangled source of distress turns out to bear a resemblance to a habitual concern, we can be confident in the human facility for survival and flourishing—or at least not utterly despondent. Except Runciman doesn't do that. Instead, he turns a standard argumentative form on its head. It's not that we can look to the past to help us solve the alignment problems

of the future. It's that the alignment problems of the future help clarify our existing sense that everything is intractable and wrong. As Runciman puts it,

> States and corporations reflect two different sides of our contemporary fear of machines that have escaped human control. One is that we will build machines that we don't know how to switch off, either because we have become too dependent on them or because we can't find the off switch. That's states. The other is that we build machines that self-replicate in ways that we can no longer regulate. They start spewing our versions of themselves to the point where we are swamped by them. That's corporations.

Runciman quite clearly believes two things. The first is that our resentful relationship to powerful machines is precedented. This might seem like good news. The second is that the recognition of this precedent does not lead to the conclusion that our relationship to powerful machines can be rendered agreeable in perpetuity. This is not good news. We have not just failed to solve the proto-alignment problem with regard to states or corporations, but, in fact, the situation has become aggravated over time. The machinery of the state has grown increasingly remote from the concerns and values of ordinary citizens, and the balance of power has tilted in the direction of ever more powerful corporations. Human values seem about as important to our current bureaucratic machinery as they would to an aloof mythical superintelligence. Just glance at our administrative dysfunction. It's all well and good to call for more democracy, but greater participation in a system that's fundamentally out of whack won't get us far.

And although Runciman's plummy tone never rises to the level of alarm, his prediction is that everything is about to get worse, and this is the point in his account where A.I. becomes more than a useful retrospective metaphor. It would be one thing if our most pressing concern were peaceful coexistence between humans and machines. But the more relevant prospect for unease is the rapport between machines and machines—between the automata of the state and the

corporation and their A.I. kin. Some of the potential consequences are obvious: in the use of automated weaponry, for example, or comprehensive state surveillance of the kind that China, which has directed an enormous amount of resources into A.I. development, is pioneering. Even in the absence of a true artificial superintelligence, we might find ourselves beset not by a single alignment problem but by multiple overlapping ones. One of Runciman's few prescriptions is that we keep foremost in mind the distinction between a machine that gives us answers and a machine that makes decisions—that we might be O.K. if we grant machines the status of oracles rather than overlords. This proposition discounts the possibilities that A.I.s can learn to deceive. (There's evidence that they already have.) But, even if we can somehow make the machines promise they'll be honest, the line between an answer and a decision is blurry, and the path from the former to the latter short. There are famous Cold War stories about military subordinates who declined the opportunity to initiate a nuclear exchange, but it remains unclear how they would have held up to the force of a state run on machine counsel. After running through a few different sci-fi scenarios of existential risk, Runciman concludes that "the bigger risk by far seems to be neither rogue individuals, nor rogue technology, nor rogue monkeys—its artificial persons gone wrong." The combination of the artificial agency of the state with even weak forms of artificial intelligence might be just as catastrophic as the regime of a single strong A.I. Plenty of people have issued similar warnings before. But those have often described a plot wherein a maniacal dictator wields a loosened A.I. Runciman's point is that the alliance between even a democratic government and a safe-ish A.I. could derail civilization.

One can, of course, object to the dark story Runciman tells. Corporations have changed over time. As a consequence of the shareholder revolution, they have become more single-minded in their pursuit of profit maximization, but one doesn't have to look back very far to find examples of corporate structures that optimized for more variables than just the bottom line. In a recent podcast episode

devoted to the book, the political theorist Lea Ypi reminded Runciman that our current political and corporate configurations hardly exhaust the possibilities for collective life. The point of a book like David Graeber and David Wengrow's "The Dawn of Everything," to which Runciman delivers a rebuttal in a lengthy footnote, is that the state as machine is a historical contingency rather than an inevitable by-product of coördinated action at scale.

Is that beside the point? It's hard to say. States might be much more than the sum of their human parts, but they're supposed to derive their legitimacy from the popular will, which means that as artificial "persons without scruples" they are nonetheless vulnerable to our appeals. They are run by people who at least in theory can be communicated with. The fact that this hasn't often been achieved doesn't mean that such an achievement is forever out of reach. The problem with an unaligned A.I. is that it won't care much about people one way or the other, and will remain impervious to our cries for consideration. Runciman devotes a perfunctory concluding chapter to scenarios like these, but it doesn't seem to rank highly among his priorities. This isn't because he's ruled out the likelihood of an A.I. doomsday, but because he considers it a failure of imagination. Members of the A.I.-safety community, he writes, are so taken with their own prophecies that they fail to consider how thoroughly we've already compromised our decision-making: "They overestimate the worth of human intelligence, and as a result they underestimate what stands in its way." For Runciman, we'll be lucky if we can coax into alignment the machines we've made of ourselves, never mind the ones made of silicon. ♦

More on Artificial Intelligence

- Can we stop runaway A.I.?
- Jaron Lanier on ways of controlling the new technology—and how we mythologize it.
- What's the point of reading writing by humans?
- As currently imagined, A.I. promises to disempower workers. But an alternative might be possible.
- "It's not possible for me to feel or be creepy": an interview with ChatGPT.

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



Gideon Lewis-Kraus is a staff writer at The New Yorker covering technology, academia, and books, among other topics. He is the author of the memoir "A Sense of Direction."

More: Artificial Intelligence Book Review States Corporations

The Books & Fiction Newsletter

Early access to new short stories, plus essays, criticism, and coverage of the literary world.

SIGN UP

By signing up, you agree to our user agreement (including class action waiver and arbitration provisions), and acknowledge our privacy policy.

READ MORE

THE LEDE

What China's Vibe-Coding Capital Can Tell Us About the A.I. Boom

Tech workers have flocked to Liangzhu, a small town outside Hangzhou, looking to escape China's corporate culture and build their future with A.I.

By Afra Wang

SHOUTS & MURMURS

Maternal Intelligence: An Open Letter

Is M.I. hopping our guardrails? If unchecked, it can lead to "recursive self-improvement," such as dyeing a "fun streak of color" in her bangs.

By Nell Freudenberger

THE LEDE

In "Musk," Alex Gibney Punctures Elon's Self-Mythology

"To make this film, I set out to solve a mystery," Gibney says in the documentary. "How could Musk assume so much power, and why had we given it to him?"

By Sheelah Kolhatkar

GLOBAL NOTES

Marco Rubio's Wild, Wild West

In Greece, the Secretary of State cast the West as a civilization defined by its warriors and conquerors, not by its democracies.

By Ishaan Tharoor

OPEN QUESTIONS

Why Can't You Sit Still?

We fight our nervous energy—but maybe we misunderstand what it's telling us.

By Joshua Rothman

VIDEO DEPT.

An OpenAI Engineer and His Friends Debate the Future

In Brent Katz's short documentary "Josh's Wedding," a group of groomsmen reckon with a technology that one of them is building.

A CRITIC AT LARGE

Why America Looks the Way It Does

From vast grids to winding suburbs, Americans have spent two centuries remaking the landscape around property.

By Daniel Immerwahr

THE LEDE

The Humanism Driving the Anti-A.I. Movement

At the Pro-Human Assembly in Washington, D.C., attendees talked less about jobs and policy, and more about their philosophy of life.

By Becca Rothfeld

BRAVE NEW WORLD DEPT.

We're Learning to Live with Waymos. It's Going Weird

A view from the back seat.

By Molly Fischer

ANNALS OF TECHNOLOGY

Computer Chips Are the Most Magnificent Things Humanity Has Ever Made

They power our phones, cars, dishwashers—and the A.I. systems that prop up the economy. They're also at the heart of fierce geopolitical rivalry.

By Gideon Lewis-Kraus

THE LEDE

There Is No A.I. "Race"

We're constantly told that Chinese advances would have catastrophic consequences. But the first to innovate is rarely the one who wins.

By Chang Che

PAGE-TURNER

MacKenzie Scott's Substack Novel Can't Escape the Amazon Juggernaut

Scott's decision to self-publish could be seen as a protest against her ex-husband's company—or as a sign of how profoundly it's altered the publishing landscape.

By Rhoda Feng

Privacy Information