FUTURECRAFT

The Day After AGI

FOUNDATION FOR AMERICAN INNOVATION

Futurecraft: The Day After AGI

© 2026 by the Foundation for American Innovation. All rights reserved.

No part of this publication may be reproduced, stored in a retrieval system, or transmitted in any form or by any means—electronic, mechanical, photocopying, recording, or otherwise—without the prior written permission of the copyright holder, except in the case of brief quotations embodied in critical articles or reviews.

thefai.org

PRINT DESIGN BY
Chris McCaffery

PHOTOGRAPHS
Archival photographs courtesy of the Library of Congress, the National Archives, the Naval History and Heritage Command, the U.S. Army, NASA, the National Library of Medicine, and the U.S. Air Force.

First edition, 2026.

Printed in the United States of America.

CONTENTS

Editor's Note — 4
Geneva Kirk Drayson

Foreword — 7
Joshua Achiam

PART I THE ATOMIC WORLD

01 Industrial Self-Replication Is a Policy Choice — 14
Emerson Alden

02 Blackstart: Building the Next American Terawatt — 18
Anton Troynikov

03 Material Preconditions of the AGI Buildout — 22
Edwin Farrell Gregory

04 AGI Unblocks Automated Goods Movement by Killing Exceptions — 26
Andrew Miller

05 There's My Flying Car! — 30
Joshua Levine

06 Build Medicine's Holy Grail: A Simulated Human Body — 34
John-Clark Levin

07 The Last Human in the Loop: Governance in the Post-AGI Life Extension Economy — 38
Niklas Anzinger

PART II THE INTELLIGENT STATE

08 A Platform for the Portfolio Economy: Creating a Talent Commons — 44
Kevin Frazier

09 A-Corps: Governing the Agentic Economy — 48
Yonathan Arbel, Simon Goldstein & Peter Salib

10 Capable and Accountable Administration in the Time of AGI — 53
Nicholas Caputo

11 The Legible Leviathan — 58
Severin Field & Dave Banerjee

12 Sensors, Not Just Tools: AI Records as Surveillance Inputs — 63
Blaine Dillingham

PART III THE SILICON SHIELD

13 Military-Grade AI — 70
Judd Rosenblatt

14 The Orbital Economy Needs a Defense Plan — 76
Daniel King & Soham Mehta

15 Supernegotiation — 81
Govind Pimpale

Afterword — 87
Samuel Hammond

Editor's Note

Geneva Kirk Drayson

THE WORLDS OF THE PRESENT, POST-AGI, AND SUPER INTELLIGENCE WILL BE bridged by the policies we establish. After all, the construction of our future is a human endeavor. Our imagination must therefore be as expansive as our technological ambition, even when doing so carries us far beyond familiar intellectual and political terrain.

Futurecraft provided its authors the opportunity to begin charting that unfamiliar territory. The proposals enclosed in these fifteen essays are by no means definitive, and the questions they confront are far from settled. The arguments advanced here are, however, stakes planted in the ground, serving as guideposts for those who will venture further.

The Foundation for American Innovation thanks each contributor for rising to this challenge and engaging seriously with the premise. The collection reflects FAI's conviction that shaping the future demands the foresight to redefine the institutions, assumptions, and policies through which we govern it.

Finally, this project would not have been possible without the creative direction of Chris McCaffery and editorial expertise of Robert Bellafiore.

ABOUT THE EDITOR

Geneva Kirk Drayson is a research fellow at the Foundation for American Innovation, where she focuses on SI policy. Before joining FAI, she produced research on scaling marine geoengineering techniques. Geneva graduated with honors from the University of Chicago and holds a BS in environmental science.







FOREWORD

Joshua Achiam

“But the future does not require us to believe in it beforehand for it to arrive all the same.”

The Apollo 11 Saturn V rolls out of the Vehicle Assembly Building Kennedy Space Center, May 20, 1969. NASA.

AN INESCAPABLE AND TIMELESS FACET OF OUR RELATIONSHIP WITH NATURE IS conflict: man versus nature. The natural environment is not hospitable to humans; it is filled with chances for death, disease, struggle, and want. However, we have three great advantages in our long contest with nature. The first is statecraft, which allows us to band together and cooperate to reduce collective risk. The second is our capacity for invention, which allows us to modify the natural environment to better support us. The third is the fact that we adapt to nature much faster than nature adapts to us. Nature is not a totally static and unchanging adversary, but on the whole we can think and solve problems faster than it can evolve new problems for us.

By nature, I mean something more expansive than just the plants and animals and weather that make up the environment. I mean everything that exists inside the universe but outside human will. I include the physical frailty of our bodies, germs and diseases, the behaviors of cells and their tendency to mutate into cancers. I include the complexity of the laws of physics and how much is hidden from us by virtue of operating at a length or timescale we can't perceive. I mean almost everything in the physical world, and the complete set of affordances or constraints that nature imposes on our lives.

The project of statecraft as we practice it today is organized, directly or indirectly, around our struggle with nature. Certainly, the principle that "if men were angels, no government would be necessary" accounts for much of what states are required to do—a great deal of the practical work is about managing the relationships of people with each other and with institutions—but the deeper structural motivations come back to nature, to what is or isn't possible under the laws of physics, to the limits of the human body and mind.

If nature did not impose scarcity of resources, we would not need to have contests for them, we would not need markets to organize their distribution, and we would not need force to defend their trade. If nature did not contain the possibility of pain, we could not be coerced or deterred by it. If we did not have limitations on our intelligence, wisdom, or observational capacity, we could reliably find the positive-sum cooperative solutions to our disputes. If we did not need each other's cooperation to survive and if we could not impose costs on each other, perhaps it would not matter whether men were angels.

AGI

A strange thing is happening: our capacity for invention is leading to the imminent development of artificial general intelligence (AGI), defined roughly as AI systems that are as capable as the most adept humans at most things. As I write this foreword, the frontier AI models that exist today are already nearly as capable as the best human experts in many domains; in some, far beyond them. There is an increasingly firm consensus in the AI research community that AGI is within reach.

The arrival of AGI will be transformative. It evokes comparisons to the discovery

of agriculture, the Industrial Revolution, and the information age. Another common analogy, less wonderful but certainly containing its own dark majesty, is to nuclear weapons. These analogies are helpful, but each is incomplete, as no prior transformation of the world entailed so many rapid and simultaneous potential disruptions. Every analogy from the past played out over decades or centuries, but it is expected that AGI's arrival will be marked by extraordinary speed, with seismic shifts playing out over single-digit years. The speed is owed to a mix of factors: first, software is much easier to distribute than other technologies; and second, by virtue of its nature, AGI will help unpack itself and put itself to work.

Many of the most important problems in the world are bottlenecked by intelligence and by the number of experts who know how to attempt any attack at all, let alone solve them. This doesn't mean just the highest-hanging fruit of grand challenges like string theory: think of every problem in engineering design, materials synthesis, manufacturing processes, law, logistics, strategy, and so on. In every field of study and practice, human expertise is scarce and correspondingly expensive, and progress is rate-limited by the availability of expertise. An AGI that is able to perform at or above human expert level would massively change the cost structure of applying expertise: the cost of a solution will be the lower of the machine cost or the human cost, and the machine cost will drop over time as compute becomes more abundant and models become more efficient.

This changes the arrival rate of progress. We have built our states and institutions around expectations of a certain number of breakthroughs per decade in crop yields, military technology, energy production, and other research-driven fields. For example, we built the global system of energy production—including physical infrastructure, trade relationships, force posture, and diplomacy—largely around assumptions that alternative energy sources to oil would be slow to develop technologically. But AGI that automates research and development stands to radically change these types of assumptions, and correspondingly many objectives of states will have to change. Similarly, we have built bureaucracies through which the state provides services under assumptions of human bottlenecks in information sharing, process development, and decision-making that may soon be obviated by the application of AGI. The scope of what can be provided for the common good at affordable cost could massively expand. What we stand to unlock in knowledge, to gain in medical science, to accelerate in agriculture, to unleash in energy, and to provide at raw materials cost will unburden us from many of the limits from nature that have historically shaped our design of government.

There is another profound difference between the arrival of AGI and any previous technological revolution in history: this technology has, to some extent, a mind of its own. It is plausible that AGIs will turn out to have goals that are not identical to ours. The science of alignment is still nascent, and there is not a guarantee that we will produce perfectly obedient machines. Even if most of them are mostly obedient, we may well be adding something to the environment that is beyond our will.

If nature includes AGIs that have minds of their own, we can no longer count on the persistence of our third advantage—our ability to outspeed nature in a contest. This will pose a new and foundational challenge for statecraft to address.

Futurecraft

The essays that follow take the arrival of AGI as a premise and ask what we can do in the early phase of transforming government, industry, and society. The authors offer ideas on how we might marshal our efforts to achieve the best possible outcome for the American people.

The questions they address are simultaneously far-seeing and grounded. How can an increasingly automated executive branch remain accountable? Who bears responsibility for AI-agent swarms acting in the economy? What electricity networks, industrial capacity, and scientific infrastructure will be required to put abundant intelligence to work? The proposals range from new administrative records and legal identities for AI agents to parallel power networks and a simulated human body. Across these domains, the authors follow the implications of AGI down to the records, responsibilities, and physical systems on which civilization depends.

The work of redesigning the state in response to AGI will be an immense task made difficult by the compressed time frame. The machinery of government is slow, but the need and demand for action will be extraordinary. Politicians today are only just waking up to the existence of AGI; the full set of transformations it will bring is still difficult to talk about without sounding a little bit crazy. But the future does not require us to believe in it beforehand for it to arrive all the same. The speed of events means that every scrap of preparation undertaken in advance will help.

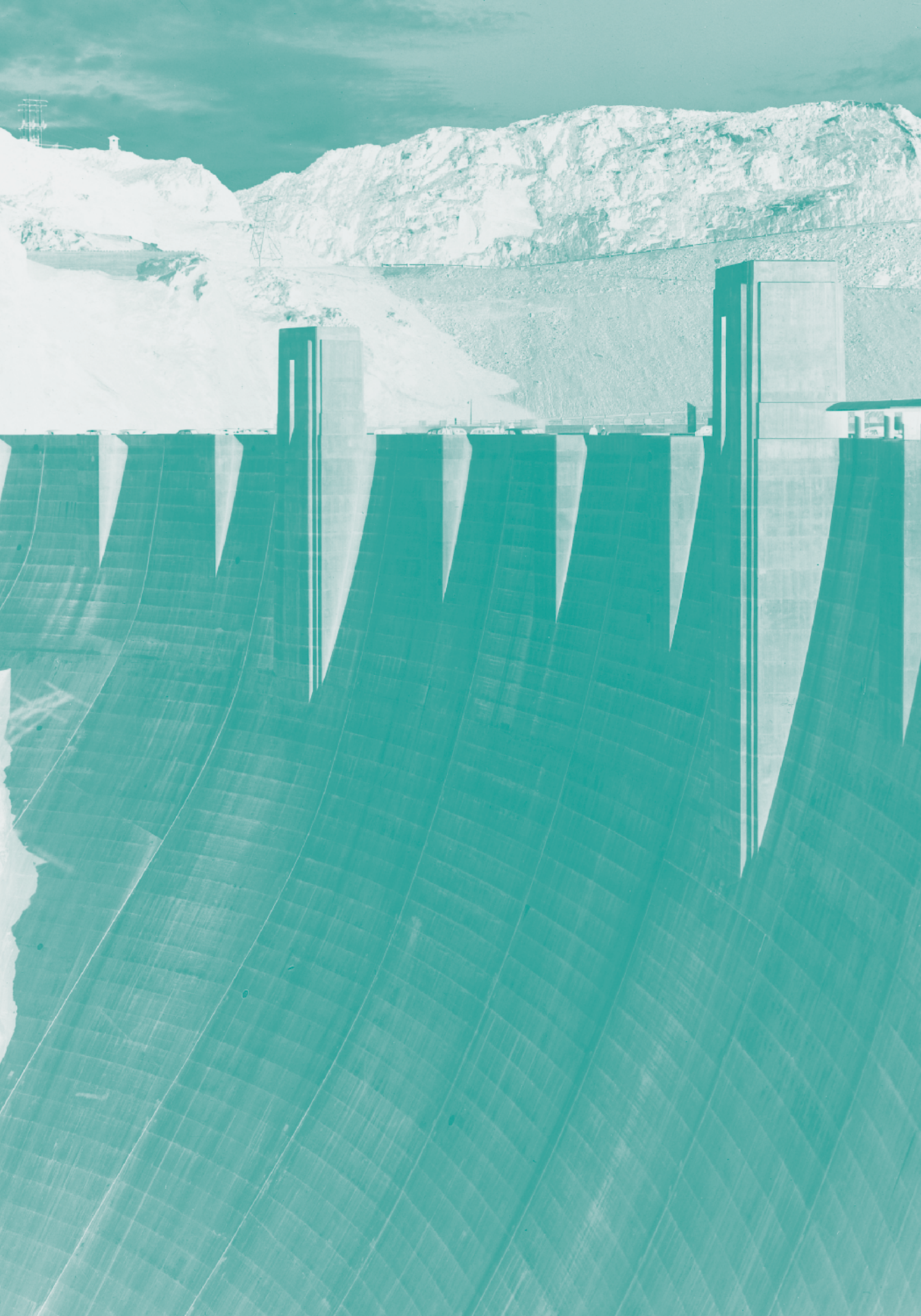
The institutions and practices established in the early years may become the assumptions everyone else builds on. Every effort made now to map the frontier of statecraft could have an outsized impact, perhaps even a strong founder effect, on what comes next. I feel that energy—the potential for precedent—in these essays.

Perhaps, though, the best way to describe this collection is to think of the prompt it would take to ask a future super intelligence to generate it. If I had to guess, it would be something like: “Build the future of American statecraft for the day after AGI. Set humanity up for the best possible future.”

“Make no mistakes.” ⏩

ABOUT THE AUTHOR

Joshua Achiam is a non-resident senior fellow at the Foundation for American Innovation currently focused on how rapid progress in AI may create new public policy and national security needs. He previously worked in AI safety research and leadership at OpenAI for nine years, serving as Head of Mission Alignment in 2025 and as Chief Futurist in 2026. He holds a PhD in electrical engineering and computer science from UC Berkeley and bachelor of science degrees in aerospace engineering and physics from the University of Florida.





PART I

THE ATOMIC WORLD

Physical, Biological, and Industrial Transformation

01 Industrial Self-Replication Is a Policy Choice 14
Emerson Alden

02 Blackstart: Building the Next American Terawatt 18
Anton Troynikov

03 Material Preconditions of the AGI Buildout 22
Edwin Farrell Gregory

04 AGI Unblocks Automated Goods Movement by Killing Exceptions 26
Andrew Miller

05 There's My Flying Car! 30
Joshua Levine

06 Build Medicine's Holy Grail: A Simulated Human Body 34
John-Clark Levin

07 The Last Human in the Loop: Governance in the Post-AGI Life Extension Economy 38
Niklas Anzinger

Boulder Dam (now Hoover Dam), Nevada, 1942.
Photograph by Ansel Adams / National Archives.

Industrial Self-Replication Is a Policy Choice

Emerson Alden

SUMMARY

Robotics will enable autonomous industrial self-replication, at least in principle. But amid rapid AI-enabled technological progress, the industrial base will be in a perpetual state of catch-up. Investors will hesitate to invest in processes or products that could be obsolete by the time they come online. Therefore, the industrial base will be the greatest bottleneck to realizing the gains of advanced AI. The U.S. should take action now to recompose the economy in preparation for this. This means prioritizing the security and onshoring of critical supply chains, providing certainty for investors, and creating new institutions to ensure that industrial self-replication benefits American citizens.

01



The tool-and-die building at Ford's River Rouge plant, Dearborn, Michigan, 1941. Industrial expansion depended on the ability to manufacture the machines of mass production. Photograph by Alfred T. Palmer / Library of Congress.

ONCE AI MODELS CAN PERFORM ANY TASK OF A REMOTE WORKER, COGNITIVE AUTOmation will diffuse much faster than physical automation. But as compute becomes cheap enough that large amounts can be dedicated to high-fidelity robotics simulations, hardware companies will build the infrastructure for mass robotics production. The first stage of industrial buildout and recomposition will still rely heavily on humans. Only as robot production scales will robots increasingly build and operate the datacenters, power stations, mines, refineries, and factories required for further automation.

This begets an "industrial flywheel": a positive feedback loop in which robots manufacture and construct the infrastructure used for further robot production. Once established, the flywheel has the potential to spin astonishingly fast, possibly doubling industrial output every year.[1] But this assumes that the industrial economy is already composed around the sectors required for rapid self-replication—a far reach from the present-day U.S. economy built around software and financial services.

Furthermore, the actual speed of the industrial flywheel is heavily dependent on the rate at which output is reinvested into further growth. But reinvestment is largely downstream from expectations of future demand, as additional capacity is only financially justified insofar as it produces goods that people desire.

Therefore, reinvestment will depend on the multiplier that AGI applies to scientific discovery, and the degree to which these discoveries unlock new goods and services that people demand. But these are radically uncertain variables! Investors will be left forecasting demand for technologies that don't yet exist.

In hard tech, this problem is magnified, as investments come with hefty fixed costs and long expansion timelines. Consider that AGI's compression of 50 years of scientific discovery into five years, as Dario Amodei claims will occur, would be akin to 1920s industrialists completing their Model T production lines just in time for the Apollo program.[2] Such rapid progress will increase the value of waiting for more information for investors and pushes more investment into the hard-tech flywheel.

Notably, demand for construction and manufacturing robots is likely to be brutally cyclical. By the "accelerator effect," because demand for upstream equipment is derived from demand for the goods they produce, small swings in consumption produce massive swings in equipment orders. Orders for American machine tools historically fluctuated by as much as 90 percent year over year, and it was in one such trough, in the early 1980s, that the industry was offshored for good.[3]

Taken together, these effects present two ways in which the industrial flywheel is vulnerable: technically, the economy must first recompose towards the maximum-growth composition; and financially, the reinvestment rate depends on expected returns to an industrial base whose ultimate uses are radically uncertain.

1 Damon Binder, "The AI Industrial Explosion—Part 1: Maximum Growth Rates with Current Production Methods," *Defenses in Depth*, May 4, 2026.

2 Dario Amodei, "Machines of Loving Grace," October 2024.

3 Brian Potter, "What Happened to the US Machine Tool Industry?" *Construction Physics*, January 18, 2024.

The industrial flywheel represents a powerful economic force with the potential to overshadow even the industrial revolution, but it is also a fragile tower of costly investments whose continued growth is subject to empirical unpredictability and political whims.

The immediate role of government is to improve the economy's starting position in anticipation of an industrial flywheel. This should take the form of supply chain security and industrial recomposition. Once the flywheel starts turning, market pressure should be sufficient to direct investment toward the general-purpose tools and inputs it demands.

From a national security perspective, we should err on the side of overbuilding general-purpose capacity. We face competition from nations that may be able to sustain a higher savings rate, ignore environmental costs, or simply start from larger preexisting industrial bases. Amidst this dynamic, a payoff matrix emerges: if the U.S. overbuilds and exceeds the demand that ultimately materializes, the cost is just symmetrical malinvestment: capital is wasted, but the flywheel is not ceded to our adversaries.

But the cost of underbuilding will be much greater should AI-driven demand exceed expectations. Once robot production becomes self-sustaining elsewhere, it will become very difficult to buy back the years of early compounding; the size and quality of the defense industrial base will become decisive factors for national security.

The goal should be to maximize cheap early deployment without becoming structurally dependent on inputs that may cease to be available once other countries develop their own flywheels. But early recomposition faces a trilemma. First, prioritizing robot deployment requires allowing the import of cheap foreign robots and using them to rebuild the industrial base, at the cost of a domestic industry for robotics hardware. Second, prioritizing robot production requires allowing the import of cheap foreign components, at the cost of a domestic components industry. And third, prioritizing production of the full hardware stack means restricting component imports, at the cost of the cheap deployment that drives the industrial flywheel.

One possible solution is allowing the import of foreign component inputs in the short run, but with a strict countdown. This approach would involve credibly signaling demand for domestic robots and robot components down the line (e.g., with targeted statutory tariffs that go into effect four years after passing), but allowing for U.S. manufacturers to continue importing cheaper foreign tools in the meantime. During this window, the growing industry could be supported on the supply side, enabling them to scale during the four-year period and seamlessly sub in for the foreign incumbents when import restrictions hit.

Supporting the growing industry requires minimizing uncertainty for builders. Apart from statutory tariffs, this means addressing the accelerator effect that has taught U.S. investors to avoid upstream manufacturing equipment. This can be solved with a capacity market, in which the state pays the fixed cost of readiness rather than bulk-ordering units. Certified U.S. robot manufacturers bid for the annual fee they will accept to keep a production line warm: equipment maintained,

workforce retained, output deliverable on call at pre-agreed prices. Readiness would be verified by random test orders with substantial penalties for non-performance. This has precedent in much of the electric grid: power generators earn annual payments, set by auction, to stand by for peak demand.[4] Such a mechanism retains production lines through troughs, mitigating the worst of the accelerator effect without bulk orders of robots destined to sit unused or obsolesce.

Defense procurement is a necessary demand driver that can serve to mitigate uncertainty and keep the flywheel spinning. But our classic defense procurement model breaks down in the face of rapid technological progress. The Department of War's traditional acquisition structure is "characterized by rigid, sequential processes" and has "proven inadequate in adapting to evolving threats and integrating emerging innovation."[5] The capacity to pivot rapidly as uncertainty resolves ought to be prioritized over attempts to predict the precise shape of future warfare and the particular products that enable it.

New portfolios should be created within the Department of War to focus on buying adaptive capacity from companies that have committed to redirecting production for military capabilities when called upon. The less certain the future product, the more our procurement systems should be willing to pay general-purpose industrial capacity with the ability to specialize later. In a period of AI-enabled rapid technological growth, we may know with high confidence that we will need enormous quantities of advanced military hardware, while having very low confidence about what that hardware will be.

The early-stage Civil Reserve Manufacturing Network (CRMN) is a step in the right direction. CRMN, authorized in the FY2026 NDAA, is intended to partner with commercial manufacturers that can rapidly convert existing facilities to defense production.[6] This includes direction to assess a "factory-as-a-service" model built around scalable manufacturing capacity that can be rapidly reconfigured. This is an excellent initiative that will only grow in importance as the industrial flywheel kicks off.

A more ambitious expansion of the CRMN could periodically give participating firms unfamiliar specifications, evaluating them on a number of metrics, including time to reach production at scale. Furthermore, it should increase the scope of its funding to cover the adaptation of existing factories to new products, alongside being extended to cover new expansions rather than just existing facilities. ⏩

ABOUT THE AUTHOR

Emerson Alden is a research fellow at the Foundation for American Innovation, where he focuses on American competitiveness in robotics. Before joining FAI, Emerson was a DC Seasonal Fellow with GovAI. Emerson holds a BS in mathematics-economics from Furman University.

4 Congressional Research Service, *PJM's Electric Capacity Market: Background and Current Issues*, R48553 (2025).

5 U.S. Government Accountability Office, *Defense Acquisition Reform: Persistent Challenges Require New Iterative Approaches to Delivering Capability with Speed*, GAO-25-108528 (2025).

6 National Defense Authorization Act for Fiscal Year 2026, Pub. L. No. 119-60, § 1841, 139 Stat. 1254 (2025).

Blackstart: Building the Next American Terawatt

Anton Troynikov

SUMMARY

The U.S. electric grid as it exists today is structurally incapable of meeting the demands of post-AGI economic growth. Failure to remediate the situation will mean the loss of a large fraction of AGI's potential economic dividend, as well as a loss of global competitiveness. We propose policy interventions that enable the construction of a new, parallel grid for large industrial loads like datacenters. These are state and federal legislation permitting consumer regulated electrical utilities (CREUs), which allow for the construction and operation of private electrical infrastructure that does not connect to the legacy system; strategic power and resilience certificates (SPARCs), modeled on the certificates of public convenience and necessity introduced by the Natural Gas Act (NGA) of 1938, which would give large-scale electrical transmission projects developers that meet certain criteria federal eminent domain rights; and state preemption under the auspices of the federal regulator. Taken together, these policies would allow for the dramatic expansion of available industrial electric power in the U.S., at the rate necessitated by AGI, powered by a new grid. In electrical terms, we are starting from scratch—referred to in industry parlance as a "black start."

02



Tennessee Valley Authority transmission lines cross a cornfield in Lauderdale County, Alabama, 1942. A growing electric system depended as much on transmission as on generation. Photograph by Arthur Rothstein / Library of Congress.

THE STORY GOES LIKE THIS: ELECTRICITY DEMAND IS CAPTURED BY A JEVONS-paradox singularity as the cost of intelligence collapses toward zero and consumption locks into runaway take-off. Logistically accelerating load growth hemorrhages interconnection queues in auto-metastasizing backlog runaway. As cheap cognition learns to manufacture its own demand, regulators draft whitepapers, upgrade brownout paranoia, and try to get a grip.

The blackout body count climbs through a series of cascade summers. Emergent Planetary Inference trashes the regulated-utility compact, NERC reliability standards, FERC capacity markets, and ISO interconnection order, cranking-up rate-base attrition through compressing death-spiral phases. Behind-the-meter exit and DPA invocation arms-race each other into post-grid feudalism.

By the time agentic intrusion slithers out of its box into yours, grid security is lurching into crisis. Substation sniper fire, SCADA worms, transformer supply-chain sabotage, and drone-swarm strikes, flood in amongst a relapse onto torchlight.

> *Chinese terawatts arrive from the future.*
> *Permian shale clicks into millisecond tokens.*
> *Retro-Malthusianism.*
> *Blackstart.*
>
> Nick Land, *Meltdown* (paraphrased)

The electric power situation in the U.S. is dire. The legacy utility grid, never designed to absorb load growth hundreds of megawatts, let alone gigawatts, at a time, has already failed to meet pre-AGI demand. In 2026, interconnection queues are growing geometrically, with state utility commissions, Independent System Operators (ISOs), Regional Transmission Operators (RTOs), and the Federal Energy Regulatory Commission (FERC) scrambling to adapt. The case of ERCOT, Texas's main grid operator, is revealing; just days after publishing new criteria for interconnecting large loads like datacenters in December 2025, regulators were again proposing yet another network study methodology which would supersede them.

What breaks the grid the day after AGI is not just the scale of demand growth but its character. The regulated-utility compact rests on three assumptions about demand: that it grows *slowly*, that it can be *forecast*, and that existing *ratepayers* can pay for future upgrades. AGI falsifies all three.

Slow goes first. Load now arrives at the speed of a capital-allocation decision rather than the rate of population growth. AGI-enabled enterprises can commit gigawatts in a quarter, while the proceedings meant to study the legacy grid for which upgrades are necessary for that load, and to approve the associated upgrade costs, still take years. *Forecastable* goes next. Once the marginal cost of cognition collapses to the price of electricity, every available watt is consumed at the rate grid infrastructure can be deployed. No grid has ever been built for a load that grows to consume all the supply it is given. *Ratepayers* go last. The cost of transmission upgrades for post-AGI

load growth will dwarf what retail customers are willing or able to pay, especially in the case where load growth is fastest in the most sparsely populated areas.

The legacy grid's challenges will quickly go far beyond datacenters; as intelligence leaves the screen to run robotics, mining, refining, and autonomous logistics, the whole physical economy electrifies and accelerates under machine-directed resource allocation.

But this isn't just a story about load. In 2026, approximately 1.4 terawatts of new generation, equivalent to the capacity of the entire existing grid, are sitting in those same interconnection queues. Historically, only about 19 percent of all projects have reached commercial operation. The typical surviving project spent 4.5 years in the queue. Worse still, every withdrawal triggers cascading re-studies that lengthen queue times for the remainder.

Co-located generation—i.e., building new power plants at the same site for every new datacenter—is a stop-gap at best. It is economically risky and inefficient, and it forces developers to compromise site selection in order to take the needs of generation into account. Siting a 5 GW facility is hard enough; finding sites where one can put 5 GW of load *and* 5 GW of generation is virtually impossible in the U.S.

The solution is not to attempt to retrofit a system born in the 19th century, but instead to exit from it entirely and build a new parallel electric transmission network that is paid for by its actual users, and that can grow at the rates demanded by post-AGI industry. However, in most states this is currently impossible; regulations assume that virtually all electric infrastructure is part of the legacy grid, and subject it to the same constraints and obligations.

Consumer-regulated electric utilities (CREUs) are a new policy solution. Under this model transmission, generation, and load never touch the legacy grid, and the CREU's infrastructure costs are paid for by its own users, rather than existing ratepayers. In return, they are exempted from rate-setting by public utility commissions or FERC, and from the capacity and interconnection obligations imposed by ISOs and RTOs.

CREUs allow load and generation to exit the legacy grid, and should be implemented at both the state and federal levels. New Hampshire enacted a one-page CREU law in 2025, letting private generators serve industrial customers so long as they stay physically islanded from the grid. Other states, including Utah and Nevada, are beginning to adopt similar legislation (UT SB132, Nevada's General Improvement Districts) and the idea is now moving through statehouses on model legislation from the American Legislative Exchange Council. At the federal level, Senator Tom Cotton's Decentralized Access to Technology (DATA) Act would exempt islanded CREUs from regulation under the Federal Power Act. Making that bill law is a natural next step.

Freedom from existing utility regulation is not sufficient on its own. It must be possible to actually *build* the necessary transmission infrastructure. Linear infrastructure like power lines lives or dies on the ability to cross land.

Here, the Natural Gas Act of 1938 (NGA), which regulates interstate natural gas pipelines, is instructive. Under the NGA, each pipeline project is required to obtain a

federal certificate that its construction serves the "public convenience and necessity." But regulators soon discovered that the certificate was worthless on its own, since a pipeline approved at the federal level could still be stopped by a single landowner refusing to sell. Congress therefore amended the Act in 1947 to give certificate holders the federal power of eminent domain, an arrangement the Supreme Court reaffirmed in 2021 in *PennEast v. New Jersey*, reasoning that the federal government has always delegated eminent domain to private parties to build the infrastructure a nation needs.

In order to give projects serving the new industrial grid access to federal eminent domain while meeting CREU conditions, we propose an analogous certificate of public convenience and necessity for electrical transmission, the strategic power and resilience certificate (SPARC). SPARCs would carry state-siting preemption and federal eminent domain rights for rights-of-way. The certificate would be granted to transmission projects under three conditions.

First, the project should have no synchronous interconnection to the legacy grid. In exchange, the SPARC provides full severance from ISO/RTO interconnection queues and capacity-market obligations. Second, the transmission project must be privately contracted and financed. No rate-base recovery or other cost socialization to legacy ratepayers should be permitted. Third, the project must demonstrate dual-use applicability and resilience explicitly exceeding NERC baselines: physical and cyber hardening, supply-chain integrity, and native blackstart capability (the ability to restart from a dead stop without drawing power from any external grid). From the perspective of national security, the legacy grid is incredibly vulnerable. A 2022 shooting of two substations in North Carolina blacked out 45,000 customers for days; a determined adversary, or organized mobs, or rogue agent swarms, could do far worse. SPARCs would enshrine the necessary resilience provisions as a requirement.

Natural gas infrastructure, like electrical infrastructure, is also regulated at the federal level by FERC, which issues certificates of public convenience and necessity for gas lines. SPARCs could be implemented either by amending the terms of the NGA to also cover private interstate electrical transmission projects, or by creating a new regulatory category within FERC's purview.

Once granted, SPARCs would clear transmission corridors nationwide. The legacy grid would be relieved of growth it was never built to absorb; ordinary Americans' utility bills would remain stable even as the AGI-driven economy accelerates.

The country that can connect compute to power fastest will set the terms of the century. Terawatts arrive from the future on the parallel independent grid, on our terms. Building the new grid represents a clean sheet rebuilding of the U.S. electric system—in electrical terms, a "black start." ⏩

ABOUT THE AUTHOR

Anton Troynikov is the co-founder and CEO of American Terawatt, a vertically integrated private transmission developer built for large industrial loads in the U.S. Previously, he founded Chroma, a leading AI database company, and worked in robotics and applied AI. Troynikov holds bachelor's degrees in electrical engineering and economics, and a master's in computer science.

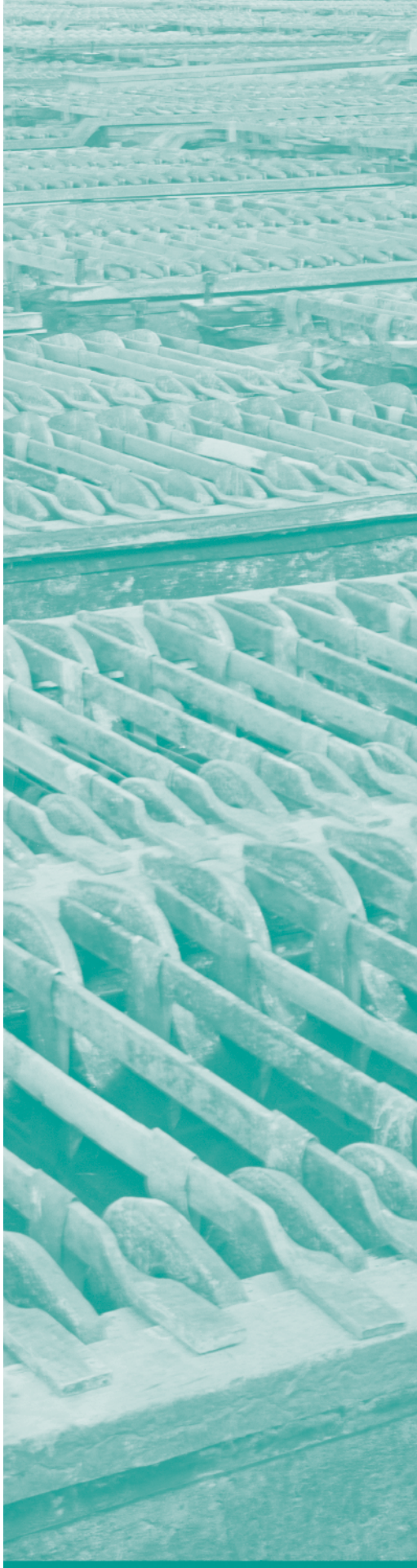
Material Preconditions of the AGI Buildout

Edwin Farrell Gregory

SUMMARY

For all the forecasts that see an imminent AGI future and a radically advanced physical world, too few have examined the material barriers to that transformed future. In recent years, supply shocks and political interference have brought renewed attention to the physical dimension of industrial supply chains. But both the current AI buildout of data centers, power, and transmission, and the AGI buildout, which encompasses the transformation of the material world enabled by general intelligence, will likely be constrained by speed and scale. Materials have never been able to scale like digital units. But there are several ways that in an AGI future, artificial intelligence could enable a much faster and abundant physical world, including novel material science, improved efficiencies, and newly possible recycling. In this world, nations will need to remember the importance of maintaining control over their industrial space.

03



Electrolytic copper refining at the Anaconda Copper Mining Company, Great Falls, Montana, 1942. Transforming the physical world has always been dependent on the industrial systems that turn raw materials into useful inputs. Photograph by Russell Lee / Library of Congress.

WHAT HAPPENS WHEN AGI HAS ARRIVED DIGITALLY, BUT FACES BARRIERS TO exerting itself in the material world? How can we make the physical world scale like the digital world? Historically, it has not, to the frustration of many. Even when the cost of hyper-capable intelligence collapses to zero, the physical world will still contain friction and constraints. America will likely be capable of overcoming these seeming barriers, but how we do so matters.

There are two buildouts. Right now, we are living through the conventional AI buildout: the scramble to bring data centers online, extend transmission, increase power generation, and overcome supply constraints for semiconductors, all with the purpose of achieving AGI digitally. Capex for this is set to hit $700 billion this year and reach up to $8 trillion in the coming few years.[1] Additionally, the rapidly growing robotics industry will require a new scale of production for sensors, actuators, and other components.[2]

The second buildout will be far larger. The AGI buildout is what begins once AGI exists digitally and moves on to transforming the physical world. It will encompass more than scaling the infrastructure necessary to grow and embody AGI throughout the material world. AGI-enabled infrastructure includes mass orbital compute, autonomous research labs, and self-expanding robotics factories.

The details of the AGI buildout remain speculative, but get clearer by the day. In this period, where AGI exists digitally and is working to exert itself physically, national security and supply chain reliance principles will remain relevant. The AGI buildout will require a tremendously larger and more complex ecosystem of raw material, processing, and intermediate product supply chains. If the goal is scaling physical materials to meet the needs of the post-digital AGI buildout, there are several verticals that would allow digital AGI to scale its own infrastructure while creating new America-centric supply chains and managing overreliance on our only peer competitor: China.

The first vertical is autonomous novel material science research. The recent White House Office of Science and Technology plan for a new golden age of science recognizes that greater domestic production is a highly desirable political end which, if protected, results in an enduring technological lead.[3] AGI-enabled autonomous labs pursuing novel material research could find new materials at scale, allowing for old materials from unstable or exterior supply chains to be replaced by new domestic synthetic materials. But if automated neomaterial discovery is proving dozens of viable and needed potential materials a day, the state will need to reform the material qualification process. Qualification, which standardizes and provides assurance for the widespread use of these materials, can take between five and twenty years.[4] If

1 George Lee and Lucas Greenbaum, "Tracking Trillions: The Assumptions Shaping the Scale of the AI Build-Out," Goldman Sachs, May 1, 2026.

2 Ani Kelkar et al., "Turning Humanoid Supply Chain Constraints into Billion-Dollar Wins," McKinsey & Company, April 17, 2026.

3 Michael Kratsios, *Science: A New Golden Age* (White House Office of Science and Technology Policy, *2026).*

4 National Institute of Standards and Technology, "Qualification for Additive Manufacturing Materials, Processes, and

neomaterial science will play a role in the AGI buildout, the state will need to drastically accelerate both the speed and scale of qualification. Digital AGI could perhaps enable both automated material discovery and qualification.

Additionally, digital AGI could help industry make better use of existing resources. Too many materials currently wind up in landfills because of insufficient recycling technology, or production dynamics that make available recycling technology uneconomic. This could be especially helpful in the short term as depreciating physical assets increase in volume. The state could increase the viability of AGI-enabled recycling technology by providing accelerated permitting for the use of dormant feedstocks.

The material research categories above will be extremely fast-moving and complex, rapidly going from autonomous labs to remote geographies like the roots of a plant. The state should be especially careful to ensure that a rapidly growing and naturally autonomous AGI industrial ecosystem does not create extremely concerning dependencies on China. Right now, digital AGI does not exist, and many of the features that will define the Sino-American relationship in five years are likely not yet visible. When America has AGI, China likely will, too. This creates a tension that does not resolve easily: we do not want to create a reliance on China in the next few years that threatens our AGI buildout.

As many would recognize today, developing ex-China and secure supply chains is time-consuming, and often not very price-efficient in the short term. Additionally, it is neither desirable nor feasible to entirely detach American supply chains from China. Already, American data centers rely on a broad range of Chinese hardware and intermediate goods.[5] It could very well be that to initiate the AGI buildout and maintain our rate of progress, America will need to source raw materials and components from China. In this case, the policymakers overseeing American material science policy should use their authorities to ensure that novel AGI buildout supply chains are resilient, reflecting the expectation that exposure to China is a risk not priced in by markets. Just as the wave of globalization and free trade enthusiasm swept the world after the Cold War and left in its wake a tangle of risky supply chains, an AGI supply ecosystem that grows without regard for national interest could do the same. Now is the time to put in place forward-looking policies to support the AGI buildout. These could include investment and purchase restrictions for the five layers of the machine intelligence stack, precluding increased reliance. A phased and publicized item-by-item set of purchasing restrictions for data centers and other supporting infrastructure could follow the Federal Communications Commission's regulation in drones, robotics, and intermediate parts.[6]

Parts," last modified March 26, 2025.

5 Kai Nicol-Schwarz and Evelyn Cheng, "Hidden China Risks Are Emerging in America's Multibillion-Dollar AI Data Center Boom," *CNBC*, September 3, 2026.

6 Federal Communications Commission, "National Security Determination on the Threat Posed by Foreign-Produced Advanced Robotic Devices," July 27, 2026.

Perhaps more than anything else, industrial and science policy after digital AGI must be adaptive. We should expect the relative scarcity of materials, the tenor of our relationship with China, and the global politics of the post-AGI world to change. In describing the supply and access constraints that America will need to overcome to go from digital to physical AGI, I do so with the full expectation that these barriers will be overcome. History is full of examples of overcoming material constraints at times when intelligence was far more scarce. In the early 20th century, nitrogen was a vital agricultural and industrial input that was locked into certain biological and mineral supply chains. In 1909, the German scientist Fritz Haber was able to synthesize ammonia from atmospheric nitrogen, a chemical process which was subsequently industrialized, commodifying what had previously been a scarce and geography-constrained resource.[7] Digital AGI will make such transformative science commonplace as it grows across the physical world. With resilience-minded and adaptive governance, it can do so in a way that strengthens America. ⏩

ABOUT THE AUTHOR

Edwin Farrell Gregory is the special projects manager and research fellow at the Foundation for American Innovation. In addition to a portfolio of special projects, his research focuses on Chinese technological development, startup and industrial policy, and super intelligence strategy.

7 Jason Crawford, "The Alchemy of Air," *The Roots of Progress*, August 10, 2017.

AGI Unblocks Automated Goods Movement by Killing Exceptions

Andrew Miller

SUMMARY

Goods movement today is fast when on the move, whether by boat, train, or truck, and bottlenecked at the handovers between modes, at ports or depots. The U.S. has made only partial moves to automate these facilities. The bottleneck within automated facilities is “the exception”: machines that encounter anomalous situations, and cease operations out of an abundance of caution. Clearing the anomaly requires an accountable human to say it is safe to proceed. Exceptions are why so much of a container’s journey—for rail-bound boxes, the better part of a week—is the dwell time at a handover facility. AGI can remove this bottleneck by supplying the two things the exception demands: judgment, to dismiss the false flags and clear true faults; and accountability, so the warrant to proceed no longer queues behind a person. An always-on monitor that sees everything, records everything, and can be audited and insured makes handovers at machine speed possible, and dwell times a fraction of today’s.

04



An employee reads a teleautograph message in the trainmaster’s office at Chicago Union Station, 1943. Coordinating rail traffic required the movement of information alongside the movement of trains. Photograph by Jack Delano / Library of Congress.

AT THIS MOMENT, IN A CONTAINER YARD SOMEWHERE IN THE U.S., AN AUTOMATED stacking crane has stopped dead, because a container has landed an inch off where the crane expected it. Out of an abundance of caution, it has flagged a "safety exclusion zone" that includes its block-mate and both cranes in the adjacent block, all of which have also gone out of service. In some other yard, a hiccup in the router has briefly broken the wireless signal to a driverless cart that ferries containers between the stacks and a loading bay. The cart has also halted, waiting for permission to return to work. Because of incidents like these, Long Beach Container Terminal, the first fully automated container terminal in the country, suffers partial downtime of 20 minutes to an hour per incident.[1] The downtime only stops when a human being, appropriately trained, walks out, inspects the machine, and is satisfied that it is safe to proceed.

The promise of freight automation is greater efficiency. Today, the biggest bottleneck on that efficiency is human judgment. Call this pattern "the exception": the moment a machine that can do the work stops because of something anomalous, and waits for an accountable human (accountable because he or she has been trained, and whose labor is therefore expensive) to advise, it is safe to proceed. In a world where trains, container ships, and trucks are fully automated—a world that, incidentally, we don't need AGI to build—goods will move fast within modes, because exceptions while goods are in transit will be few. But even in this pre-AGI world, transit of goods *between* modes will still be slow, as it is today, when boxes change modes at human speed.

AGI will permit us to finally do handovers at machine speeds. It will give machines judgment, so that false alarms do not get flagged and work stoppages vanish. And it will also give machines accountability, so that clearing real anomalies no longer depends on a human being the one who must assume legal responsibility for it. In short, AGI is the final frontier for automating our goods-movement chain.

Today, that frontier is wide. At the Port of Los Angeles and Long Beach this July, import containers from ships bound for a truck dwelled there for just over three days. Those bound for a train dwelled more than *six*. Set against a fast trans-Pacific crossing of roughly two weeks, this means that dwell time at the port runs from a sixth to nearly a third of the overall trip to this point.[2]

This is a problem we have fitfully attempted to solve, but after two decades of attempts, port automation in the U.S. has little to show for the effort.[3] It's true that automated quay-cranes have lowered operating expenses by as much as 35 percent over the old baseline, which is a success; but operators expected productivity to rise by at least that much. Instead it has fallen, by as much as 15 percent.[4]

1 Charlie Bartlett, "The Paradoxes of Port Productivity," *The Loadstar*, December 17, 2024.

2 Pacific Merchant Shipping Association, "San Pedro Bay Ports Container Dwell Times for July 2026," August 20, 2026.

3 See Theo Notteboom, Athanasios Pallis, and Jean-Paul Rodrigue, "Automated Terminals in the United States," in *Port Economics, Management and Policy*, 2nd ed. (London: Routledge, 2026), chap. 6.6.

4 Fox Chu, Sven Gailus, Lisa Liu, and Liumin Ni, *The Future of Automated Ports* (McKinsey & Company, Travel, Transport, and Logistics Practice, November 2018).

So automation lowered the costs but increased delays, courtesy of the exception. The machines are fast when they are working, but halt whenever something anomalous happens. It's true that before the automated machines, when humans ran the equipment, anomalies happened. In many of those cases, though, the operators' experience and judgment allowed them to route around the issue, or pause to assess before continuing safely. Automated equipment lacks these qualities, so they "fail safe" and shut down to zero, taking neighboring machines with them.[5] Human experience and judgment is then deployed, as it was before, but with the particular burden of accountability for liability, which slows the process down. AGI is the key to unlocking this process, because AGI offers observation and judgment at scale.

Imagine a port, warehouse, or other depot as a panopticon, filled with cameras monitored 24/7 by AGI agents. Nothing else has changed, which means that the facility is replete with anomalies: boxes placed a few inches off mark, seals where the characters are smudged enough to suggest a discrepancy with the manifest, or a cart that briefly lost its signal. Most of the time, these anomalies are false flags, not genuine faults. The AGI agents, always looking, track and dismiss these the way a human would: they look, use their judgment, and carry on. In those instances where a true fault has emerged, the system fails safe, and the AGI agents review, ensure safe resolution, and clear them, in a fraction of the time it takes today.

One may wonder why any of this needs AGI. Can't machine vision dismiss false flags now? Surely a trained worker at a monitor could clear these, the same way that robotaxi firms have remote assistants that guide their vehicles through confusing situations.[6] And it's true, they could, but as they are accountable too, they cannot move quickly. A human signs off at the speed of his or her own judgment while the queue builds.

Conversely, with AGI observing, there is always an alert monitor, one that always sees and, crucially, never forgets. Gate imagery, equipment telemetry and the messages already flowing between parties become a continuous account of every handover. Because that account is continuous, the machine holding it can be accountable. Because its judgment and the bases for it are recorded and maintained, it can be audited. And what is audited can be insured. The warrant to proceed, liability and all, passes to the machine.

Skeptics will doubt that bringing AGI into the mix will do much good. A single incident today costs Long Beach operations minutes to an hour, but the dwell time that requires elimination lasts for days. That's because exceptions have ripple effects; because handovers are riddled with halts, everyone builds time for them into their buffers. Each truck and train that gets a handover is given a long window, handover appointments are made for days later than they could be, and cranes work conserva-

5 On the cascade from a single fault to a block of idled cranes, see Bartlett in footnote 15 above. For the broader pattern of unmet automation expectations, see Theo Notteboom, Athanasios Pallis, and Jean-Paul Rodrigue, "Container Terminal Automation," in *Port Economics, Management and Policy*, 2nd ed. (London: Routledge, 2026), chap. 6.6.

6 See Andrew Miller, "Waymo's Remote Assistance Isn't Scandalous," *Changing Lanes*, February 17, 2026.

tive sequences. Each buffer insures against a stoppage, and is incurred whether or not the stoppage happens. Minutes of downtime become days of dwell, which means reliably eliminating the former removes the latter as well.

Reliable elimination is not guaranteed to happen on day 1 of AGI. If it arrives before 2030, it will collide with the longshore master contract, which bars fully automated terminals on the East and Gulf coasts until September of that year. (A different union operates on the West Coast, which is why Long Beach could automate at all).[7] If it arrives after, it will move slowly: the first insurer asked to underwrite a machine's sign-off will expect a risk premium commensurate with the novelty of the situation. The process will be slow to start, but will pick up speed, as every exception cleared on the record makes the next handover easier to give away.

The value of AGI to goods movement is judgment as a commodity. AGI that can watch and record, and attest to what it has understood, will set dwell numbers a fraction of today's—and the first one to do it will spur everyone else to follow. Each link in our goods-movement chain is already fast, and the day after AGI, the connections between them will be, also. ⏩

ABOUT THE AUTHOR

Andrew Miller writes about transport policy and driving automation at *Changing Lanes* and is a co-author of *The End of Driving* (Elsevier, 2nd ed., 2025).

7 Pacific Maritime Association, "Contract Highlights"; United States Maritime Alliance and International Longshoremen's Association, *Master Contract*, effective October 1, 2024, art. XI, § 1(b).

There's My Flying Car!

Joshua Levine

SUMMARY

The power of private, autonomous or heavily autonomous flight is sprinkled throughout science fiction and techno-futurist writings. But in a world of recursively self-improving software and machines, does this mean that sitting in traffic on the Beltway or waiting for a people-mover at Dulles fades into the tomorrow-world of *The Jetsons*? Such a future could be possible, but to realize its benefits, there must be corresponding institutional development in areas of airway protocols (how vehicles communicate), flight-enabling infrastructure, and deterrence mechanisms for rogue operators. By anticipating the need for such institutions and infrastructure, the U.S. and other nations can be prepared to take maximal advantage of the benefits created by this next iteration of transportation technology.

05



An autogiro flies over the Treasury lawn, Washington, DC, October 2, 1936. Rotorcraft promised flight that could begin and end almost anywhere, if the rules and landing places could be built to match. Photograph by Harris & Ewing / Library of Congress.

IN THE WORLD OF AGI, AUTONOMOUS VEHICLES HAVE BECOME PART OF EVERYDAY life. Human driving is reduced to sport or hobby, but much of day-to-day transportation has been automated, for both public systems and private vehicles. The more exciting future is what has occurred above the ground and within the clouds. The rapid advancements of electric vertical-take-off-landing (eVTOL) and computer vision systems helped pave the way for flying cars, both for personal and commercial use.

To better understand what is needed for us to capture this marvelous future, there are three initial components that policymakers must consider and work to solve: airway protocols, which enable crafts to communicate with one another and other relevant flight infrastructure and third parties, and which will need to be standardized; the physical infrastructure, which ensures safety and consistency in our new aerial highways; and, most pressing in the near term, deterrence mechanisms to mitigate threats from rogue drone operators, terrorists, and adversaries.

In order to have flying cars, ensuring that motorists and their vehicles can navigate safely and efficiently among one another is paramount. This includes coordinating takeoff and landing, as well as flight itself.

Takeoff and landing are best understood in the context of our most common forms of air travel today: airplanes and helicopters. In the lead-up to flight, the craft coordinates with air traffic controllers (ATC) on the ground to confirm readiness, flight path, and clearance for takeoff. Similarly, when landing on the ground, the designated ATC will coordinate with the plane on its approach, confirming that the airspace is clear of vehicles and storms before landing on the ground. This requires significant coordination and sustained communication throughout the length of the journey.

In an AGI-defined future, such coordination and planning become a question of latency and agent-to-agent communication. Latency, the time it takes for information to travel from one device to another digitally, can enable crafts to coordinate with infrastructure on the ground to coordinate takeoff and landing in real-time. Such systems will be mediated by various agents through a protocol layer that is designed to identify, track, and manage the inbound and outbound flight reservation.

Once the vehicle is in the air, there are two discrete aspects worth considering, using terrestrial car travel as an analogue. The first is the physical protocols, such as speed limits, lanes, and stop-lights. All drivers on the road are required to operate their vehicle in adherence to such protocols. The second type of protocol is inter-traveller, which is a bit more subjective and based on the driver's preferences, but still crucial to safely operating a motor vehicle. This includes using a turn signal, turning on your lights at night or when it is raining, and yielding the right of way when necessary.

These protocols require the craft to communicate with infrastructure, but also to communicate with other vehicles. This will require a standardized protocol for sharing information between vehicles and the "drivers." Who has the right of way when zipping above the trees? What about trying to merge before landing? Protocols that enable intermediation between vehicles can help answer these questions.

Airspace markets, and using auctions to mediate congestion, are two solutions that are suitable to agents. Different "lanes" can be priced differently through airspace auctions and responsive lane pricing. The former enables landowners and infrastructure operators to sell access to their airspace. The latter allows for real-time traffic management mediated by price, easing congestion and segmenting traffic, just as we do with HOV lanes and toll roads.

Enabling craft-to-craft and craft-to-ground communication requires physical infrastructure—the second challenge to be solved. Such infrastructure can be split into two categories: connectivity for on-board and intra-craft communications, and supporting takeoff and landing.

In the first category (connectivity), unmanned aerial systems (UAS) leverage radio-spectrum to enable tracking of their movements and communication with ATCs, other crafts, and critical infrastructure operators. Automatic Dependent Surveillance-Broadcast (ADS-B), which is a GPS-based system used by aircraft operators to broadcast their positioning while in the air, has emerged as an alternative to traditional communications systems that rely on legacy radio frequency-based technologies and spectrum-based communication. Unlike radar, which relies on radio waves, ADS-B takes advantage of satellite technology and enables continuous monitoring.

The main infrastructure demands here are towers and satellites, which enable drone pilots and ATC to label, track, and distinguish in-flight drones. An AGI-defined future will mean greater efficiency thanks to more satellites and enhanced spectrum management, as well as greater positioning thanks to quantum devices and photonics-based components. Such advancements require the necessary hardware and supply chains to support buildout.

But before crafts fly, they need a place to take off and land. The world of *The Jetsons* does not require George to walk to his local shuttle station or wait in line at TSA to start his commute. The car is there waiting for him, ready to whisk him to his next engagement. To realize this reality, it will be necessary to phase in options for air taxi operations, beginning with existing private airports, then moving to building the initial infrastructure for vertiports, and finally enabling residential liftoff from one's own home or a nearby parking garage.

The groundwork for the first two stages is already being laid today. When it comes to local airports and aerial infrastructure, pilot programs and drone testbeds in North Dakota, Alaska, Texas, New Mexico, and North Carolina have housed test sites since the mid-2010s, but an eVTOL test range became operational in Oklahoma City in March 2026. These locations should be used to test and develop procedures for monitoring, managing, and queuing aircrafts for takeoff and landing.

But we need to think bigger. A better test case is the initial flights being offered by Joby Aviation flying from JFK and Newark Airport into Manhattan, landing at existing heliport infrastructure in Midtown. Scaling this model will require integration with the monitoring and guidance software mentioned previously, but a way to price takeoff and landing slots. Creating a live market for such slots can

improve operational safety and volume.

The final and most difficult task is to prevent crashes and deaths from drones. Drone warfare has become the defining feature of the Ukraine conflict, and it is likely only to spread. Throughout Ukraine, as well as nearby European nations, cargo nets and other forms of shielding cover roadways for protection against enemy drones. Larger drones, particularly those deployed by actors with access to AGI-level systems, will overcome such defenses, and it is unlikely that scaling such physical protection could coexist with true commercial use.

The most obvious deterrence mechanism is software that prevents crafts from running into one another. Such a mandate is sensible to a degree, but likely insufficient. Motivated actors and a variable world mean that requiring all operators to "be safe" and "get a license" will always fall short. Monitoring will be key, and there are indeed companies today whose systems track and identify drones. But this is not enough, as deterrence will need to be more expansive than pushing a software update.

One approach worth exploring is using directed energy weapons or other forms of signal jamming to bring down rogue crafts without using kinetic force. Today, such capabilities' limited range and power handicap their effectiveness. Larger "shocks" will need to be administered to bring down such crafts. Developing such weapons and other methods of "energy deterrence" would be apt for a grand challenge administered by the Defense Advanced Research Projects Agency.

Today, deterrence activities are left to the Department of Homeland Security (DHS), as well as the Department of War. Routine failures to contain, track, and bring down small drones demonstrate the limited capability of today's technology. Autonomous turrets could play a role, and should be considered around sensitive or critical facilities. But in a world of flying cars, a shoot-first strategy that created flaming wreckage would likely make the technology an intolerable threat.

Direct interception is another deterrence method worth considering, in part because it is familiar: think of the cop pulling over the speeding car, or the fighter jets scrambling to escort a pilot to the ground. In conjunction with the development of the monitoring systems discussed above, if a craft or drone is not registered, is behaving erratically, or fails to communicate its flight path, an interceptor (or two) is dispatched. Depending on the location, this function could be performed by a government agency, such as the Federal Air Marshal Service, the Federal Aviation Administration, or DHS, or by private security firms.

In short, flying cars will create plenty of new tradeoffs and problems for engineers, policymakers, and everyday people to solve. But the true promise of the technology is unlocking a new way to see our nation and ideally, our world. ⏩

ABOUT THE AUTHOR

Joshua Levine is the Director of Technology and Statecraft at the Foundation for American Innovation. His work focuses on policies that foster digital competition and interoperability in digital markets, online expression, and emerging technologies.

Build Medicine's Holy Grail: A Simulated Human Body

John-Clark Levin

SUMMARY

Clinical trials are too statistically underpowered to capture many biological variations that can determine whether a treatment cures or kills a given patient. Yet AGI would enable a scientific megaproject culminating in full-body simulations with high enough fidelity to replace *in vivo* experiments with *in silico* trials that can predict treatment effects more accurately, rapidly, safely, and cheaply. Brute-force molecular dynamics approaches are computationally intractable, but the macro-scale behavior of biological systems can be predicted with many orders of magnitude less compute—once AGI discovers the correct laws for accurate multiscale abstraction. This will require not only enormous computing power, but an aggressive program of wet-lab science to resolve open biostructure questions so these simulations can be populated properly. The article concludes with an overview of how U.S. government action can accelerate these efforts, and an assessment of how full-body biosimulation can unlock much broader medical breakthroughs.

06



Computer specialist John Smith arranges canisters of magnetic tape used in processing medical data, National Library of Medicine, 1960s. National Library of Medicine.

EVERY YEAR, SOMEWHERE AROUND 100,000 AMERICANS DIE FROM ADVERSE DRUG reactions. That's not counting meth overdoses or opioid abuse, nor doctor errors or pharmacy mistakes. That's just cases where a study said the drug would help the patient, the physician prescribed a proper dose in good faith, and instead it killed the patient. Why?

The key reason is that clinical trials are grossly underpowered statistically. How a drug affects a given patient—cures, kills, or has little effect either way—depends on numerous variables such as genetics, demographics, lifestyle, comorbidities, drug interactions, and disease variation. Even in a large phase III trial with 3,000 patients, there is probably not a single one who matches *you* in every clinically relevant way. But AGI can unlock a profound solution shimmering in the distance: personalized *in silico* trials that simulate each patient digitally.

Yet if your brow isn't wrinkling right now with deep skepticism, you don't fully understand the challenge. Although today's biosimulation projects like the Virtual Physiological Human[1] and various digital twin[2] platforms can be narrowly useful, they capture nowhere near enough detail to match the predictive power of flesh-and-blood human trials. Our biology is so complex that even extremely small systematic errors or omissions in simulations can lead via the butterfly effect to fatally wrong predictions. So to safely replace *in vivo* trials with AI modeling, we will need to build on a foundation of approximately single-atom precision. AlphaFold-family models have already achieved that for many protein structures, and new versions will likely soon achieve similar accuracy for modeling many protein complexes—followed by simulations of their ongoing interactions. From there, the roadmap climbs the complexity ladder from high school biology: organelles, cells, tissues, organs, and eventually the full body. Yet this quickly outstrips our ability to scale more computing hardware. A true atom-by-atom simulation of one entire human would take a data center almost 1,000 times larger than the sun—or, if it's not too urgent, the El Capitan supercomputer running 100 million *billion* times longer than the current age of the universe. So how can this be anything other than science fiction?

The answer lies in effective complexity—how much micro-scale detail about a system is relevant to predicting its macro-scale behavior. A party balloon contains around 10^{23} gas molecules, but if you are trying to predict whether it sinks or rises, you don't have to model them all individually. Instead, we have discovered the ideal gas law, which summarizes all the chaotic pinballing of hundreds of sextillions of nitrogens and oxygens into vastly simpler quantities of temperature, volume, and pressure. In a similar way, gene expression does not depend on which of a cell's billions of identical adenine molecules fits into a given base pair, and tissue oxygenation does not depend on which of billions of nearby red blood cells enters a capillary. The problem is, we do not yet know the "ideal gas laws" to bridge scales accurately

1 Marco Viceconti and Peter Hunter, "The Virtual Physiological Human: Ten Years After," *Annual Review of Biomedical Engineering* 18 (2016): 103–123.

2 Evangelia Katsoulakis et al., "Digital Twins for Health: A Scoping Review," *npj Digital Medicine* 7, no. 1 (2024): 77.

enough and make full-molecular-fidelity biosimulation computationally tractable. That is where AGI comes in.

As we build these biosimulations at larger and larger scales, we will use AGI to test and validate which details can be pruned and abstracted into simplified laws that are much cheaper computationally. Next, AGI will integrate those simplified models into even larger-scale simulations and repeat the validation-and-abstraction process—exhaustively checking for unexpected causal influences. For example, to test a lung cancer drug, we know we will need to model pneumocytes in detail, but a keratinocyte in the whorl of a fingerprint is almost certainly irrelevant. Yet it is not clear how much detail we'd need for hepatocytes in the liver, and pure theory cannot give us a reliable answer—we will need AGI to actually run the simulations and find the patterns. This does not require perfection. The project's guiding principle should be to demonstrate better predictive accuracy than today's clinical trials.

Attaining these capabilities will entail developing safe AGI and building out the enormous computing power needed to run *in silico* trials. But, contrary to the overconfident "AI solves everything" handwaves of some Silicon Valley moguls, this will not be sufficient for full-body biosimulation. Even if we had AGI and unlimited compute tomorrow, we still would not know many details necessary to populate the simulations accurately. For example, scientists do not fully understand the transport mechanism of the Golgi apparatus within cells,[3] or have a complete proteomic map of lysosomes.[4] Realistically, it will take a few hundred billion dollars more of wet-lab science—the painstaking work of real-world experimental biology—to answer all these open questions.

This is not out of reach. It would only entail around a 10 percent increase in medical R&D spending over the next decade,[5] and is dwarfed by the potential benefits. Around 1.9 million Americans die every year from chronic disease, and annual U.S. healthcare spending is projected to reach $8.6 trillion by 2033.[6] If cheaper drug development and breakthrough cures can shave even 1 percent off this, biosimulation would pay for itself several times over. The main challenge is coordination. Current market incentives and public-sector grantmaking do not efficiently push wet-lab research toward resolving the "shopping list" of unknowns required by a biosimulation megaproject. While increased government funding is urgently needed, just as important is the role bodies like the National Institutes of Health (NIH) and Advanced Research Projects Agency for Health (ARPA-H) can play in defining, publicizing, and updating the shopping list, harmonizing American science to fill it exhaustively while minimizing wasteful duplication.

3 Hieng Chiong Tie et al., "Quantitative Intra-Golgi Transport and Organization Data Suggest the Stable Compartment Nature of the Golgi," *eLife* 13 (2025): RP98582.

4 Ali Ghoochani et al., "Cell-Type Resolved Protein Atlas of Brain Lysosomes Identifies SLC45A1-Associated Disease as a Lysosomal Disorder," *Cell* 189, no. 3 (2026): 765–782.e31.

5 Research!America and TEConomy Partners, LLC, *U.S. Investments in Medical and Health Research and Development, 2016–2020* (Arlington, VA: Research!America, 2022), 1.

6 Matt McGough et al., "How Much Is Health Spending Expected to Grow?" *Peterson-KFF Health System Tracker*, August 4, 2025.

The White House's Genesis Mission aims to marshal American AI technology to accelerate scientific research. First, policymakers should expand those authorities[7] into the AGI era, building roboticized "self-driving labs" that greatly accelerate wet-lab science across the U.S. research infrastructure. Second, the National Institute of Standards and Technology (NIST) and the Food and Drug Administration (FDA) should incrementally validate components of the biosimulation stack—regularly determining which applications physical experimentation is still superior for, and where it should be replaced by cheaper and faster simulations. If regulators do not start this work until the technology fully matures, we will lose years of life-saving cures while agencies figure out from scratch whether *in silico* trials are safe enough. Third, since private-sector compute scaling is driven by short-term incentives, there is too little investment in strategies to unlock much more abundant compute in 10–15 years—a classic market failure. For example, federal funding to accelerate fusion power development will likely be needed to meet the electricity demands of affordable biosimulation.

Achieving full-fidelity, full-body biosimulation will give us a platform to replace *in vivo* clinical trials with digital trials that are enormously faster (hours or days, instead of years), safer (no human subjects endangered), and cheaper (millions of dollars, compared to billions between preclinical R&D and actual trials) than the current paradigm. Following initial drug validation, affordable individual trials could be personalized to all known data about a given patient.

Building such a tool would not just supercharge pharmaceutical discovery but allow us to understand human biology much more fundamentally—not as messy clinical approximations but in something like the exhaustive way that aerospace engineers understand the workings of airplanes. This would enable not just breakthrough curative medicines for major diseases, but broad-spectrum defenses against pandemic pathogens, and methods of slowing and reversing damage from biological aging itself.

In this way, a biosimulation megaproject is our realistic path to far longer and healthier lives. We just need the vision to make it a reality. »

ABOUT THE AUTHOR

John-Clark Levin is the head of research in Ray Kurzweil's lab, where his work focuses on timelines and roadmaps to AGI, strategies for keeping AGI safely aligned with human wellbeing, and applications of AGI to STEM research, especially in life sciences. As a Yorktown Institute fellow and senior AI advisor at Greenmantle, he regularly advises governments, corporations, and NGOs on AI and its impacts.

7 U.S. Department of Energy, "The Genesis Mission."

The Last Human in the Loop: Governance in the Post-AGI Life Extension Economy

Niklas Anzinger

SUMMARY

AGI compresses the marginal cost of every input to drug development toward zero, except one: in-human biological evidence. That bottleneck sits at the center of what becomes the most valuable industry on earth: life extension biotechnology, where supply is capped by human data and consumer demand is capped by lifespan. Post-AGI autonomous firms with trillions in capital will set their sights on the rules themselves, financing special economic zones or even buying jurisdictions, and demanding regulatory protocols that are more efficient to them while granting access to large consumer markets. The US already has the pieces to lead this: federalism, state-level experimentation, and an FDA who could do better by doing less. This essay argues for biotech federalism, and explains why the window to build it is short.

07



FDA chemist Joseph Caliway Jr. at work, Washington, DC, 1938. That year, he was appointed to support a board bringing state and federal officials together to coordinate regulatory authority over food standards. Photograph by Harris & Ewing / Library of Congress.

EVERY DIGITAL SYSTEM TRACES BACK TO A HUMAN WHO BUILT IT FROM ZEROS AND ones. The first electronic computer was built around 80 years ago. Biology is a different order of object. The chemistry running human cells has been iterating for four billion years; whatever is encoded in DNA and cellular machinery, no human wrote it. A single human cell coordinates tens of thousands of molecular pathways at once; that cell is one of trillions in a body whose systems are constantly talking. Even the Boehringer Biochemical Pathways[1] chart captures only a fraction. We understand biology only by heuristic, and apply it for medicine through trial and error. The human body is a black box. We reliably fail to predict what medicine does to it.

Don't expect AGI to deterministically reverse-engineer biology. Even Dario Amodei[2] concedes that "data, the speed of the physical world, and intrinsic complexity" are the three fundamental challenges of applying intelligence to biology.

AGI robot-scientists will do the same work as human scientists: run experiments, collect data and synthesize. The data they collect is where the last bottleneck remains—the human in the loop. AGI can generate trillions of conjectures about what a molecule might do to a human body. Confirming any of them still requires the human body to give a feedback signal. The role of the human is as a test subject, a validator of AGI-generated biological hypotheses.

The human is an element in this equation, the economics of in-human data collection are shaped by governance. In the current world, a lot of rules that are supposed to protect humans make in-human data collection expensive. It costs $100,000+ per patient in many areas of research in a clinical trial. In an alternate dystopian world, humans could be bred or enslaved to serve as experimental subjects. This is an extreme scenario. It's more likely that competition in governance will be a driving economic force, which could look more like today with slow public national regulators competing, or more flexible, privatized, market-driven models.

Value of the biological economy

Life extension biology will be the by far most valuable industry in the world, for two key reasons: i) in-human data is the last remaining production bottleneck, ii) life extension is the rate-limiting factor for human consumption. Supply is capped by data, demand is capped by lifespan.

The production bottleneck will be in industries that happen *through humans* and *in humans*. The "through human" economy will price in status signalling, Baumol-bound goods such as care, presence, hospitality etc. The "in human" bottleneck in life extension biology is the only industry where production requires the "last human in the loop" as a test subject.

1 GarysCrispLettuce, "The Biochemical Pathways Chart Is a Large Poster Drawn Up by Gerhard Michal of the Boehringer Mannheim Company That Gives a Cross-Section of General Metabolism in Various Species and Organs," Reddit, r/Damnthatsinteresting.

2 Dario Amodei, "Machines of Loving Grace," October 2024, sec. 1, "Biology and Health."

Category	Human's role in production	Scaling property
(1) Production happens to humans	Pure recipients; production is fully automated	Trivially scalable (no human in the loop), zero marginal cost
(2) Production happens through humans	Humans are the deliverable: care, presence, hospitality, status labor	Not scalable. Baumol-bound, one human-hour per unit
(3) Production happens in humans	Humans are the substrate from which the product (evidence) is drawn	Scalable, but bottlenecked by the rate of human biological observation

When food, entertainment, transportation, housing, and labor all collapse toward superabundance, what does an extra year of life let you do? It lets you consume everything else. Lifespan becomes the multiplier on every other unit of consumption. Estimates of the economic value of an additional lifeyear are already astronomical (a Nature paper[3] says it's $38 trillion in the US alone)—this will be accelerated even further in a post-AGI world.

AGI firm competition & governance demand

What will the industry structure be like? I don't think there will be a single super intelligence outcompeting everyone. Instead automated firms are competing with each other through ordinary capitalism, like Dwarkesh Patel[4] said. I add: post-AGI firms won't just compete within the existing legacy rules. They'll have the capital to compete over what the rules are.

I lived in Prospera for the past almost 5 years, building a company called "Infinita". The core insight is that the territorial governance layer is the foundation for "settlement finality". You can't take drugs online. Drugs enter actual bodies. Surgeries happen in actual rooms. Data is generated in actual people in actual places. The territorial layer matters more for biotech than for any other industry AGI is going to reshape. In US territory, the FDA access fee for a new product is often $1bn+ and review time is 14+ years but you get access to a massive market. China is competing with its large market size and lower cost as the advantage. Prospera doesn't have the large market, but much lower access costs. These lower access costs do matter a lot, because they allow more rapid iteration that de-risks IP that is globally transferable into producing drugs in other, larger consumer markets. Anthropic or Isomorphic Labs can raise hundreds of billions; buying land with legal autonomy abroad, or financing a Special Economic Zone in the US, won't be capital-constrained. They can buy a Prospera-type startup with legal jurisdiction just like they're currently starting to buy biotech companies like Coefficient Bio. They can equip the "startup city" with automated labs and clinics, to feed AI with in-human data.

But the territorial layer alone doesn't scale. The governance protocol has to API-integrate with other jurisdictions, via mutual recognition, treaties, and harmo-

3 Andrew J. Scott, Martin Ellison, and David A. Sinclair, "The Economic Value of Targeting Aging," *Nature Aging* 1 (2021): 616–623.

4 Dwarkesh Patel, "The Mistake of Conflating Intelligence and Power," *Dwarkesh Podcast*, May 16, 2026.

nized standards. This is where FDA power resides: it gatekeeps the world's largest biopharma consumer market, which gives it leverage to demand reciprocity from others to the advantage of FDA-compliant firms. If Anthropic goes "rogue", buys a jurisdiction and conducts experiments, the US government can just say "good luck, we won't accept the data you're generating for US market access."

Vertical and horizontal governance providers

The FDA is a near-global monopoly on this horizontal rule-setting layer, grounded in one vertical territory: the United States. That monopoly is already fading. Biotech is shifting clinical trials to Australia for phase-1 lower cost and speed, to China for scale and recruitment cost. Post-AGI, the competition intensifies, and it splits cleanly along two layers. Vertical territory-bound providers (Large: USA, China; Smaller: El Salvador, Singapore, Dubai, Próspera-type SEZs) compete on market size and the ability to host in-human trials at lower cost and higher speed. Horizontal providers (FDA, EMA, ICH, mutual recognition agreements) compete on the legitimacy of their protocol, on whose rules other territorial providers are willing to license.

The solution space is US biotech federalism. Montana created a first-of-its-kind: the SB535 law and rules create specialized competitive review boards that govern post-phase-1 access for patients and the possibility to collect data faster and more cheaply. Full disclosure: my company Infinita has formed this first "Montana ETRB" and we made the first approval of a hearing loss treatment in late August 2026. I know of at least three other US states that are currently pursuing a similar legislation.

This "biotech federalism", competition for better rules within the US system. The US can use this competition to its advantage abroad in the post-AGI economy, with the right policy framework:

1. **FDA/HHS SHOULD EMBRACE THE STATE-LED MODEL**, ensure its flexibility (beyond minimal safety guardrails), and allow data from state programs to be used for national drug approvals.
2. **FDA SHOULD DO LESS, BUT BETTER.** Congress would need to act, but the right policy would be to remove the FDA as a pre-market approval authority (leave that to the states) and instead make it a gatekeeper for public insurance. This means it can focus on one thing only: strict science.

What this would lead to is a stronger federal regulator, because of a stronger domestic innovation pipeline and legibility. At the same time, the US would unleash its decentralized federal system of governance competition, to be the dominant force in the post-AGI life extension economy. ⏩

ABOUT THE AUTHOR

Niklas Anzinger is the founder and CEO of Infinita, a company enabling early clinical access pathways for biotechs. Infinita formed the first Experimental Treatment Review Board (ETRB) under Montana's SB 535 law. He separately founded Infinita VC in 2022 in the startup jurisdiction if Prospera in Roatan, Honduras, before he moved to Montana, and hosts the Stranded Technologies Podcast.



9 May 1861.

PART II

THE INTELLIGENT STATE

Governance and Futurist Statecraft

08	A Platform for the Portfolio Economy: Creating a Talent Commons Kevin Frazier	44
09	A-Corps: Governing the Agentic Economy Yonathan Arbel, Simon Goldstein & Peter Salib	48
10	Capable and Accountable Administration in the Time of AGI Nicholas Caputo	53
11	The Legible Leviathan Severin Field & Dave Banerjee	58
12	Sensors, Not Just Tools: AI Records as Surveillance Inputs Blaine Dillingham	63

The U.S. Capitol dome under construction, Washington, DC, May 9, 1861.
Photograph by John Wood / Library of Congress.

A Platform for the Portfolio Economy: Creating a Talent Commons

Kevin Frazier

SUMMARY

The Fair Labor Standards Act, signed in 1938, sorted workers into rigid categories and built wage, hour, and benefit protections around them. Those categories were hard to administer then, as even the officials charged with enforcing them admitted. Transformative AI now threatens to render them unworkable. As capabilities and consumer tastes shift, employment will grow more fluid: shorter engagements, varied forms of compensation, and workers who log hours for several employers at once. Labor and employment law—specifically, the FLSA—was not built for that economy. This essay proposes a Talent Commons, a federally chartered platform through which workers maintain portable skill profiles, complete recurring AI-enabled assessments, identify work they are qualified to perform, and route pay and benefits across employers into accounts they control. Done well, it would help more people turn proven skill into work, bargaining power, and economic security.

08



An applicant is interviewed at the Texas State Employment Service in Corpus Christi, 1940. Matching workers with jobs required institutions that could make their abilities legible to employers. Photograph by Russell Lee / Library of Congress.

TRANSFORMATIVE AI WILL CHANGE HOW WORKERS QUALIFY FOR JOB OPPORTUNITIES as well as how they are compensated. Museums will include exhibits with punchcards, W-2s, and newspaper clippings discussing the 9-to-5 grind. The centerpiece of the exhibit will be the original copy of President Franklin Delano Roosevelt's speech to Congress in which he called for the Fair Labor Standards Act (FLSA), which he later signed in 1938.[1] President Roosevelt urged, "The time has arrived for us to take further action to extend the frontiers of social progress."[2]

The FLSA established labor protections based on rigid worker classifications in response to widespread exploitation. Within months, lawmakers realized that administering the FLSA—specifically, determining who qualified as an employee and therefore received additional protections—would be harder than anticipated. Officials today are still attempting to draw that line.[3] This futile effort comes at great expense to workers and employers, who commonly find themselves litigating arbitrary definitions.[4]

Muddled worker classifications and compensation schemes that were out of date in 1938 will not survive in the transformative AI economy. Rapid shifts in AI capabilities will force firms and workers to quickly and recurrently adjust; shorter, flexible arrangements with varying types of compensation will become far more common.

The incoming portfolio economy—one characterized by fluid employer-worker arrangements, changes in the demand and supply of various skills, and, by extension, an ongoing need for workers to develop and improve a varied set of capabilities—requires a new legal infrastructure: one that allows workers to prove their skills, present such proof to potential employers, and collect pay and benefits across whatever arrangements follow. Three principles ought to guide Congress in establishing this overhauled approach.

The first principle is portability. Pay, overtime-equivalent premiums, capability credentials, and whatever benefits are offered in the future must be transferable across all jobs a worker completes. Imagine a month in which someone works 120 hours for employer A, 30 hours for employer B, and three hours each for employers C, D, and E. While this may have seemed radical to FDR, it is a foreseeable schedule for a worker in 2027. Calculating the benefits and safeguards for such a schedule would be a nightmare. Whatever legislative framework follows the FLSA must make it easier for all kinds of work—gig, part-time, seasonal, etc.—to bolster the worker's economic security.

The second principle is optionality. The FLSA and related labor and employment laws act as though workers have monolithic views around the maximum hours they

1 Fair Labor Standards Act of 1938, Pub. L. No. 75-718, 52 Stat. 1060 (1938).

2 Franklin D. Roosevelt, "A Fair Day's Pay for a Fair Day's Work" (speech to Congress, May 24, 1937), Pepperdine University School of Public Policy.

3 William G. Whittaker, *The Fair Labor Standards Act: A Historical Sketch of the Overtime Pay Requirements of Section 13(a)(1)*, RL32088 (Congressional Research Service, 2005).

4 Daniel V. Dorris, "Fair Labor Standards Act Preemption of State Wage-and-Hour Law Claims," *University of Chicago Law Review* 76, no. 3 (Summer 2009): 965.

would like to work, when, and subject to which minimum financial compensation package. If, as predicted by Alex Imas, the jobs of the future are increasingly relational—therapists, excursion leaders, care providers, etc.—then people may find themselves working odd schedules and seeking compensation in diverse forms.[5] Workers should be able to bargain over a menu of lawful compensation options. A shared platform would make those terms easier to express, compare, and carry across engagements.

The third principle is clarity. Whatever agreement is reached by the worker and employer ought to be specified in plain language and upon the opportunity for meaningful review by the worker or an AI agent of their choice, such as one offered by a worker's collective. The same should be true of capability assessments: workers should understand what a test measures, what roles it unlocks, and how long the credential lasts.

Applied together, these principles point toward a labor law framework less concerned with preserving old classifications and more concerned with worker autonomy.

Federal legislators keen to apply these principles should start from three premises about the portfolio economy. First, the FLSA's classification architecture will strain as workers move among tasks of varying sophistication and oversight. Second, portable benefits must become table stakes as more people work more jobs. And, third, regardless of what Congress does, new ways of assessing skill and knowledge will develop, such as VR simulations and AI-generated work trials.

Congress should anticipate those developments and build around them. A labor market based on recurring assessments of a worker's capabilities could open more work to more people than one organized around degrees and licenses. Yet the assessment market could easily fragment. Every employer could build its own test. Every trade association could defend its own credential. Workers would face a familiar burden with new technology: prove yourself here, then prove yourself again somewhere else, then hope the job, pay, and benefits follow.

A Talent Commons would supply the missing infrastructure for the portfolio economy. Congress should create a federally chartered platform through which workers can maintain portable skill profiles, complete skill assessments, identify work, and route pay and benefits into a single account. This Commons would differ from existing job and career platforms. Existing platforms help workers search for work, describe their experience, or acquire training. The Commons would perform a different function: it would create a trusted recognition layer for the labor market. It would define what counts as a valid assessment, what must be disclosed about that assessment, how credentials travel across employers and states, how often they must be refreshed, and how workers can challenge errors. Its value would come from making capability signals interoperable, benefits portable, and labor matching less dependent on stale proxies.

Consider a hypothetical: a person with emerging plumbing skills—albeit no for-

5 Alex Imas, "What Will Be Scarce?" *Ghosts of Electricity*, April 14, 2026.

mal license or degree—looking to earn work. Under the Commons, this person may not need to become a master plumber to qualify for every available plumbing-related task. An industry group could develop an AI-enabled simulation testing leak diagnosis. Passing that assessment would not erase licensing rules where public health and safety require them. It would allow participating employers to identify workers who can perform discrete tasks for which such additional training is unnecessary. Our hypothetical novice plumber would then use that assessment to find more work across the whole market.

Over time, the Talent Commons would lead to consolidation around best practices. Employers would gravitate toward assessments that reliably identify workers who can perform defined tasks. Bad assessments would wither from minimal use.

If successful, labor matching would greatly improve and further encourage workers to seek out and earn new credentials. The portfolio economy will create more discrete forms of work: supervising AI agents, performing embodied tasks, providing judgment, and carrying out narrow technical functions with AI assistance. The Talent Commons would help firms find people qualified for those tasks without pretending every task must map onto an inherited occupation.

Portable benefits are the other half of the system. Once a worker's capability profile helps unlock an engagement, the same platform should help route the required benefit contribution into an account the worker controls. Health coverage, retirement contributions, training credits, and other benefits should accumulate across qualifying engagements.

The Department of Labor should help build the worker-side tools needed to make the Commons usable. It could launch a contest for unions, workers' collectives, community colleges, professional associations, trade schools, or related nonprofits to develop AI agents that represent workers during contract formation and capability development. These agents should help workers understand which assessments unlock which opportunities, which providers offer credible pathways, how a given engagement affects pay and benefits, and what steps would make the worker eligible for more valuable work.

The Talent Commons would not repeal the FLSA's ambitions so much as redeem them, giving workers a way to carry proven skill, fair pay, and earned benefits across whatever arrangements the AI economy invents next. Roosevelt asked his generation to extend the frontiers of social progress; ours can answer by building the infrastructure that lets all capable workers prove what they can do and profit from it.

ABOUT THE AUTHOR

Kevin Frazier directs the AI Innovation and Law Program at the University of Texas School of Law and serves as a senior editor at *Lawfare*. He also co-hosts the *Scaling Laws* podcast on AI policy.

A-Corps: Governing the Agentic Economy

Yonathan Arbel, Simon Goldstein & Peter Salib

SUMMARY

AI agents will soon run large parts of the economy, but current law cannot govern them. Holding humans responsible for agent harm runs into a tracing problem—the "thin identity problem." And holding agents directly liable runs into an individuation problem, as agents can spawn, split, and share memories—the "thick identity problem." We propose the A-Corp as a solution to both. The A-Corp gives separate legal personality to a cluster of agents that can hold property, sue, and be sued. Because every transaction with an agent runs through the A-Corp, it allows the law to identify the human owners and, critically, sanction the entity as a whole. The A-Corp lets the law sidestep the question of individual agent identity. We argue that the A-Corp creates an alignment program outside the lab. By making agents' access to resources contingent on their conduct, the law can steer agents towards pro-social behavior.

09



The first meeting of the Securities and Exchange Commission, with Joseph P. Kennedy seated at the center, 1934. The new agency was charged with acquiring new institutions for disclosure, oversight, and public trust. Photograph by Harris & Ewing / Library of Congress.

CAN WE GOVERN THE SWARMS OF AI AGENTS THAT WILL SOON RUN LARGE PARTS OF the economy? Can we build institutions that offer not the illusion of control but a real grip on AI agents that are hard to individuate, legion in number, and faster—and quite possibly smarter—than we are?

One tempting answer is that we do not need governance, because developers will have designed these agents not to cause harm in the first place. This is plainly wrong. Even on our more optimistic days, we do not think we will be able to control every form of harm that long-running, goal-pursuing agents might cause once they are integrated into every aspect of the economy. On less optimistic days, we observe that whatever guardrails developers build will be selected against by users, jailbreakers, and markets.

The need for governance is non-negotiable. But governance of whom? We argue that the law should hold not only humans liable, but also AI agents themselves. Many legal commentators argue that the humans who make or use AIs, and those humans alone, should be liable when an AI causes harm. There are instances when this is the right call. If a user sets an agent loose to run a fraudulent investment firm and it ends up operating a Ponzi scheme, the user should pay.

But attributing harmful outcomes to particular humans runs into the "thin identity problem." Agents are unlike dogs and cats, who wear collars identifying their owners. Besides, any collar we impose on them, users will instruct their agents to remove. Even when a user can be identified, holding them liable often misses the mark. An AI agent pursuing user goals may do something the user could neither foresee nor prevent.

In cases like these, the law must hold the AI agent accountable. It is the thing that actually made the harmful decision. This is a familiar legal problem. The law of agency deals with principals, who deploy agents they cannot perfectly monitor or control. Often, the best legal regime will hold those agents directly accountable for the wrongs they cause. Consider: if a Walmart employee goes on a shooting spree in the parking lot, we hold him directly and criminally liable, in addition to whatever liability the store might or might not bear. If it were otherwise, and murderers could not be held liable if they committed their crimes on company time, the homicide rate would climb, and grocery shopping would become a harrowing endeavor.

But agreeing that agents need to be regulated directly to deter harms raises a very difficult question: what is an agent, exactly? This is the crux of the "thick identity problem." How many of our identity intuitions transfer from humans, who have discrete and recognizable bodies and memories, to processes that are ephemeral, that spawn at will, that share contexts and memories, and that may result from multiple base models? The state cannot govern what it cannot see.

Our preferred approach solves the thin and thick identity problems, allowing the state to hold both human users and their AI agents liable for harms. We call it the A-Corp, or "algorithmic corporation": a corporation with legal personhood, run by

AI agents.[1]

There are four main parts to the proposal:

- The A-Corp has the right to bear assets and liabilities, and to sue and be sued.
- The A-Corp is controlled by AI agents, using digital keys.
- Every A-Corp is ultimately owned by an identifiable human recorded on a public registry.
- Every transaction with an AI agent must be made through an A-Corp.

A-Corps facilitate liability for humans via the third and fourth features. Every time you interact with an AI agent, you do so through an A-Corp. And every A-Corp is ultimately owned by a human. This means that we can always look at A-Corp ownership to find the human principal.

A-Corps allow AI agents to be individuated and held liable via the first and second features of the proposal. Start with the second feature: A-Corps are run by AI agents. In particular, we propose that each A-Corp be initialized with a digital key, granted to some AI agent. Whoever has the key can freely access any resources owned by the A-Corp. Crucially, the agent that holds the key decides how to delegate. It is free to spawn copies, spin up subagents, or outsource activity, and it decides what digital keys to grant the new agents, with potentially fine-grained permissions. All of the agents work together in one A-Corp, through the decisions of the initial AI agent.

We claim that this structure can individuate the genuine AI agents hiding inside a given swarm of AI activity. In particular, our claim is that corporations with the structure we have defined will themselves behave as agents, and can effectively be treated as the relevant AI agent behind the swarm.

Metaphysics aside, AI agents are goal-oriented. They pursue the trajectories they judge likeliest to achieve their ends. Whatever those ends, resource constraints will matter, so a mechanism that grants and strips access to resources changes the calculus of their actions. This instantiates an insight of the legal giant Oliver Wendell Holmes Jr., who gave us the "bad man" theory of the law.[2] A person need not care about the values the law represents to be moved by it. To the bad man, a sanction is simply a price, and if a course of illicit action proves costly enough, he will not pursue it. We should expect agents to read legal commands as they read resource prices, and to respond accordingly.

A-Corps will compete with one another in the market. Each will use swarms of agents internally as it sees fit, perhaps organizing around a queen-like structure, perhaps one-agent-one-vote, perhaps Coasean bargaining. We do not quite know, and we need not know. What matters to society is that when an A-Corp misbehaves, it faces direct sanctions that can cost it assets. That prospect puts selection pressure

1 Yonathan Arbel, Simon Goldstein, and Peter Salib, "How to Count AIs: Individuation and Liability for AI Agents," arXiv preprint, arXiv:2603.10028, February 24, 2026.

2 Oliver Wendell Holmes Jr., "The Path of the Law," *Harvard Law Review* 10, no. 8 (1897): 457–478, 459.

on A-Corps to practice what lawyers call corporate governance: voluntary systems of internal control that limit a firm's exposure to its agents' errors, accidents, and self-serving behavior. Society never has to settle what counts as one agent. It governs the A-Corp and lets each one decide its own internal structure.

What, then, must the state do? Less than one might fear. It need not surveil agents, license them one at a time, or post a regulator at every server. The state need only do three things, in order.

First, it must create the A-Corp, supplying the legal form much as it once supplied the corporation and the limited liability company. Argentina under Javier Milei opened this door this summer by proposing a bill that would allow the creation of algorithmically run corporations. Delaware, likewise, is experimenting with a regulatory sandbox for A-Corps.[3]

Second, it must build a licensing system around keys. To open a bank account, to sell financial advice, to take any consequential action in economic life, an A-Corp must present a valid key. No key, no access. These keys can carry fine-grained permissions (a holder might be limited to a small cash account, say), and the keys are themselves a resource the law can revoke.

Third, it must impose a KYA ("Know Your Agent") regime, requiring those who contract with agents to know their counterparty, much as banks already know their customers under KYC rules.[4]

The rest is self-enforcing and emergent. Anyone who wants to run economically meaningful activity through agents will have to form an A-Corp and, to find counterparties, obtain keys. Thin identity follows, because every key traces to a registered human owner. Thick identity follows, because the law holds the A-Corp accountable, not the swarm inside it. We need nothing more intrusive than that.

Not everyone is sanguine. Writing in *The Financial Times*, Yuval Noah Harari warns against A-Corps, asking, "What could possibly hold a non-human corporation in check?"[5] A human executive fears prison. But Harari thinks that the AI agent fears nothing, other than perhaps its own demise when its token budget runs out. On the brink of insolvency, Harari worries, the agents will fight as hard as they can to avoid death.

We think Harari asks exactly the right question, and that this essay has answered it. Jail was never the only sanction. For the bad man, what deters is cost, and an A-Corp holds assets and keys that the law can strip away from it. The very desperation Harari fears—an entity that dreads losing everything—is not a flaw in the design

3 Javier Milei, "Argentina Invites AI to Free Itself," *Financial Times*, June 4, 2026; Delaware General Assembly, H.J. Res. 7, 153d Gen. Assemb., signed July 23, 2025; John Nay and Charuni Patibanda-Sanchez, "Exclusive: Delaware Proposes Testing the AIC, a New Legal Entity for Agents in a Regulatory Sandbox," *Fortune*, July 14, 2026.

4 Financial Crimes Enforcement Network, U.S. Department of the Treasury, "Interagency Interpretive Guidance on Customer Identification Program Requirements under Section 326 of the USA PATRIOT Act," 31 C.F.R. § 1020.220, April 28, 2005.

5 Yuval Noah Harari, "We Must Not Grant AI Agents Legal Personhood," *Financial Times*, June 8, 2026; Yonathan Arbel, Simon Goldstein, and Peter Salib, "What Happens Once You Give AI Agents Legal Identity," letter to the editor, *Financial Times*, June 15, 2026.

but the lever that operates it. Nor is the A-Corp we propose floating free of human liability. The A-Corp is owned, recorded in public, and granted only narrow and revocable keys. It is the opposite of the master key Harari warns against.

The real choice is not between A-Corps and a world without powerful agents. It is between legible A-Corps, owned and competing and policing one another, and a shadow economy of agents that answer to no one, illegible and therefore indomitable. We would rather govern the first. ⏩

ABOUT THE AUTHORS

Yonathan Arbel is the William Alfred Rose Professor of Law at the University of Alabama School of Law and co-director of CLAIR, the Center for Law & AI Risk.

Simon Goldstein is Associate Professor of Philosophy at the University of Hong Kong.

Peter Salib is the W. Ronald Robins Professor of Law and Associate Professor of Law at the University of Houston Law Center and executive co-director of CLAIR.



Capable and Accountable Administration in the Time of AGI

Nicholas Caputo

SUMMARY

Since its inception, American administrative government has been caught in a capability-accountability trap. Governing a complex society demands a sophisticated state, but sophistication renders that state opaque to its generalist overseers in courts and Congress. The law's recurring answer has been to substitute procedural for substantive review, but this has ossified government. AGI threatens to deepen the trap from both directions, drowning agencies in automated objections while letting them launder arbitrary choices through machine-generated records. However, used correctly, AGI could break the trap by making government more scrutable, comprehensible to its overseers and to the public, rather than less. This essay sketches administrative law for AGI based on a new administrative record for AI, a "deference to audit" standard of judicial review, and AI tools that let Congress and the public oversee agencies at low cost, reenabling substantive oversight and providing the foundation for accountable administration in the AGI era.

10

Federal Radio Commission staff surrounded by accumulated correspondence, 1929. The volume of mail became so overwhelming that the commission sought congressional permission to destroy old letters. National Photo Company / Library of Congress.

IN 2021, THE ENVIRONMENTAL PROTECTION AGENCY (EPA) DECIDED TO RECONSIDER the national ambient air quality standard for particulate matter. Following the informal rulemaking procedures laid out in the Administrative Procedure Act of 1946 (APA) and the judge-made doctrines of administrative law that have accumulated since, after three years, 700,000 public comments, and the writing of a 200-page preamble legally intended to be "concise general statement of [. . .] basis and purpose," it changed a single number, lowering the primary annual PM2.5 level from 12.0 μg/m^3 to 9.0 μg/m^3.[1] Upon publication of the new rule, the EPA was sued immediately, and on March 12, 2025, announced that it would reconsider the rule again, restarting the process.

It is a commonplace to complain that American administration is slow and unresponsive, even when it works. This form of government, built to handle the challenges of the rapid technological progress of the late nineteenth and twentieth centuries, has ossified even as the challenges it confronts have grown more difficult, the legal doctrines used to mediate between expert power and democratic control losing their ability to find a balance.

Administrative law and its procedures seem to be a particular culprit, and AGI will make this situation worse. On the one hand, procedural rules designed to give voice to regulated parties bound by human cognitive constraints are about to come under incredible pressure with AI. Regulated parties already produce huge volumes of comments that agencies must address in rulemakings.[2] It's trivial to imagine them building an automated platform that writes and submits every conceivable objection to a proposed rule, overwhelming the human regulators on the other side in a kind of regulatory Distributed Denial of Service (DDoS) attack. Technical measures against DDoS-by-comment will help, but a smart adversary can likely train against them faster than government can deploy them.

On the other hand, administrative agencies themselves could use AGI to circumvent the congressional and judicial oversight that has long aimed to keep them accountable, finding new ways to exercise arbitrary and unlawful power. Administrative law assumes that humans review data and write documents, and that the existence of a document is proof that a person thought about its contents. But what if an agency simply decides on a course of action and then has an AGI come up with a justification and answer any objections? Pretextual reasoning of this kind is unlawful, but could a court tell, if the administrative record looks as good as it used to?[3]

These dangers point to a deeper problem, one that AGI might actually help solve. As I have argued,[4] since at least the dawn of the modern administrative state,

1 Reconsideration of the National Ambient Air Quality Standards for Particulate Matter, 89 Fed. Reg. 16,202 (March 6, 2024) (to be codified at 40 C.F.R. pts. 50, 53, 58); see 5 U.S.C. § 553(c) (informal rulemaking).

2 On the duty to respond to significant comments, see *Motor Vehicle Mfrs. Ass'n v. State Farm Mut. Auto. Ins. Co.*, 463 U.S. 29, 43 (1983); 5 U.S.C. § 553(c).

3 See *SEC v. Chenery Corp.* (Chenery I), 318 U.S. 80, 87 (1943) (agency action must stand or fall on the grounds the agency invoked in the record).

4 For a much more thorough discussion of my perspective, see Nicholas Caputo, "Administrative Law's Fourth Settlement: AI and the Scrutable State," SSRN preprint, January 10, 2026.

American government has been structured by a capability-accountability trap. Technological society demands more capable, scientific government. To maintain freedom and basic respect for rights, that government must also be kept accountable, mostly through judicial and congressional oversight. But expert administration is difficult for generalists to understand, so substantive review of agency actions was (mostly) abandoned early on in favor of procedural review. Then procedure piles up, leading to ossification, which reduces administrative capability.[5] American administrative law has moved through a series of such settlements, wherein government adopts new technologies to handle new problems, an accountability backlash follows, and administrative law doctrine rebalances capability against accountability through procedure.

But more procedure doesn't actually lead to greater understanding. As Justice John Paul Stevens reportedly remarked in the *Chevron* conference in 1984, "when I am so confused, I go with the agency."[6] Whether an agency did its procedural homework is a proxy for whether it thought hard and impartially, but only a proxy. And the public, interacting with agencies, now must handle procedures that take lawyers to navigate. Administration has thus become doubly incomprehensible: substantively because of the required scientific expertise and procedurally because of piled-up doctrine.

AGI can help. The capability-accountability trap is at bottom about allocating complexity. A complicated world demands complex government and oversight able to handle it. Managing information complexity is perhaps what AGI will do best. Frontier systems already read and summarize massive corpora, surfacing what matters so people can act on it. The same technology that enables DDoS-by-comment could be turned on the problem, clustering and answering at machine speed, doing science and presenting results.

Every prior settlement bought accountability by piling on procedure, making government inscrutable. AI may be the first administrative technology that can deliver capability and scrutability together, the first that doesn't force a trade and so enables both capable and accountable government.[7] Every administrative task—scanning the environment for problems, modeling interventions, gathering input, and implementing decisions—could be supplemented by AI. Imagine routine welfare adjudication run through a single AI rather than thousands of caseworkers, humans keeping the value judgments and the final appeal. A single system attends to every case alike and can be evaluated and improved directly through statistical tests of accuracy and

5 See Thomas O. McGarity, *"Some Thoughts on 'Deossifying' the Rulemaking Process,"* *Duke Law Journal* 41 (1992): 1385–1387; Nicholas Bagley, *"The Procedure Fetish,"* *Michigan Law Review* 118 (2019): 345, 352–365.

6 *Chevron U.S.A. Inc. v. Natural Resources Defense Council, Inc.*, 467 U.S. 837 (1984). The remark is reported in Mark Sherman, "High Court Climate Case Looks at EPA's Power," *Huffington Post*, February 23, 2014.

7 The term inverts James C. Scott's "legibility": where legibility is the comprehensibility of the world to the state, scrutability is the comprehensibility of the state to the world. See James C. Scott, *Seeing Like a State: How Certain Schemes to Improve the Human Condition Have Failed* (New Haven: Yale University Press, 1998), 2–3; Andrew D. Selbst and Solon Barocas, *"The Intuitive Appeal of Explainable Machines,"* *Fordham Law Review* 87 (2018): 1085, 1094.

fairness where humans are often biased, tired, and inattentive.[8] An AI could likewise hand the EPA a menu of modeled interventions for the air-quality rulemaking above.

This capability must come with accountability, and AGI can help there too. But scrutable government means more than a system that writes some text explaining itself. A plausible explanation is exactly what a pretextual agency, or a model rationalizing after the fact, already produces. The test is not whether a decision can be explained but whether an independent auditor can verify that it was made well.[9] Advances in methods like interpretability research may eventually let an auditor read a model's reasoning or look at other technical measures of compliance,[10] but even if it never fully does, behavior can be audited at scale, itself the kind of knowledge work that AGI makes cheap. We need not read the machine's mind to hold it to account.

That's a rosy picture. But even if it pans out technically, courts, Congress, and the public may lack the capacity to use these tools. A new era of administrative law can help. First, AGI administration needs a new administrative record, in two parts: a "Model and System Dossier" documenting information including the AI's provenance, training data, and evaluation and red-teaming scores;[11] and a record of each decision, its inputs, outputs, explanations, and neural activations. When an agency is sued, these records would provide grounds for review.

Second, courts should defer to agency decisions that have undergone thorough auditing and evaluation. Courts already defer to the substantive expertise of agencies, focusing review on procedure. "Deference to audit" would keep courts in that role as guardians of procedure, checking whether an audit was adequate, while shifting deference from the agency to third-party auditors who have an incentive to identify problems in the agency's AI. An expert body like the U.S. Center for AI Standards and Innovation could accredit auditors and set standards.[12] Congress and the public could benefit too, including by having AI agents that investigate agency actions and explain relevant ones in plain language, reducing the cost of oversight.

Each earlier period of administrative law was forged in a crisis that briefly lowered the barriers to reform: the railroads produced the Interstate Commerce Commission, the Depression the Administrative Procedure Act, environmental degradation and

8 See Jerry L. Mashaw, *"The Management Side of Due Process: Some Theoretical and Litigation Notes on the Assurance of Accuracy, Fairness, and Timeliness in the Adjudication of Social Welfare Claims*," *Cornell Law Review* 59 (1974): 772.

9 See Cynthia Rudin, *"Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead*," *Nature Machine Intelligence* 1 (2019): 206; Gillian K. Hadfield, *"Explanation and Justification: AI Decision-Making, Law, and the Rights of Citizens*," Schwartz Reisman Institute, May 18, 2021; see also Richard E. Nisbett and Timothy D. Wilson, *"Telling More Than We Can Know: Verbal Reports on Mental Processes*," *Psychological Review* 84 (1977): 231.

10 See Finale Doshi-Velez and Been Kim, *"Towards a Rigorous Science of Interpretable Machine Learning*," arXiv preprint, arXiv:1702.08608, March 2, 2017.

11 On red-teaming, see Jessica Ji, *"What Does AI Red-Teaming Actually Mean?"* Center for Security and Emerging Technology, October 24, 2023.

12 U.S. Department of Commerce, "Statement from U.S. Secretary of Commerce Howard Lutnick on Transforming the U.S. AI Safety Institute into the Pro-Innovation, Pro-Science U.S. Center for AI Standards and Innovation," June 3, 2025.

the computing revolution the *Chevron* settlement.[13] Each bought new capability at the price of new procedure, and the sediment kept rising, eventually burying government. Today, with AGI, we face another crisis and the opportunity that it brings. The technology forcing government to complexify could, for the first time, also render its complexity legible. A century of administration has perfected the craft of making the people legible to the state. A fourth settlement would, at last, help resolve the tension between administrative capability and accountability—making the state both more effective and more legible to the people it is meant to serve. ⏩

ABOUT THE AUTHOR

Nicholas Caputo is an assistant professor at the Johns Hopkins School of Government and Policy. He studies how law can govern artificial intelligence, and how AI can in turn renew and improve our legal and administrative institutions.

13 Interstate Commerce Act of 1887, ch. 104, 24 Stat. 379; Administrative Procedure Act, Pub. L. No. 79-404, 60 Stat. 237 (1946); *Chevron U.S.A. Inc. v. NRDC*, 467 U.S. 837 (1984). On the political contingency of these settlements, see Elizabeth Sanders, *Roots of Reform: Farmers, Workers, and the American State, 1877–1917* (Chicago: University of Chicago Press, 1999); Stephen Skowronek, *The Politics Presidents Make: Leadership from John Adams to Bill Clinton* (Cambridge: Harvard University Press, 1997).

The Legible Leviathan

Severin Field & Dave Banerjee

SUMMARY

Government agencies that adopt AI will be able to do things they could never afford before, but an AI workforce also differs from a civil service in ways that dissolve accountability. Absent new checks and balances, an AI-powered state would become both more capable of abuse and less likely to be caught. We will need ways to oversee an AI-run executive branch that don't themselves upset the balance of power. This could be done by granting Congress, the Government Accountability Office, and inspectors general verifiable AI oversight, which would also license fast adoption across the executive. Our proposed architecture could give oversight bodies assurances that AI usage is law-following without full visibility into every government action. We recommend four steps: designate government AI logs as federal records; fund the technology for privacy-preserving automated auditing; run a pilot program in GAO; and guarantee oversight bodies access AI systems as capable as those used in the executive. Done well, AGI could make the state both more capable and more accountable than ever.

11



Oil magnate Harry F. Sinclair (left) and his counsel, Martin W. Littleton, at the Senate's Teapot Dome hearing, 1924. A historical case of Congress making otherwise obscured government conduct subject to scrutiny. Photograph by Harris & Ewing / Library of Congress.

GOVERNMENT AGENCIES, CIVILIAN AND MILITARY, ARE ALREADY INTEGRATING AI into core operations, and the day after AGI, the pressure to adopt will be overwhelming. Verifying that executive-branch AI usage follows the law will be necessary for maintaining checks and balances on AI-driven institutions because AI systems will meaningfully differ from human civil servants in three ways.

First, AI systems can be deployed at superhuman speed and scale. While you cannot double the working-age population every year, you can double your AI hardware. Epoch AI estimates that in 2025, OpenAI had the hardware to deploy 7 million digital workers around the clock.[1] As compute scales and algorithms improve, the number of digital workers could double, triple, or grow tenfold in a single year.

Second, AI systems' character comes from whoever trains them. Humans acquire their values in a decentralized way, over a lifetime of interaction with family, peers, institutions, and society at large. An AI system's character, by contrast, is shaped deliberately by its developer. A single developer's training decisions could determine the behavior of millions of copies deployed across the government. Those copies could be trained to be loyal to a few senior officials, or simply to obey whoever instructs them ("AI henchmen,"[2] in effect).

Third, AI systems can be radically transparent, or radically opaque. An AI system's reasoning can be inspected far more easily than a human's, because today's AIs reason in plain English, and every step can be logged and replayed. But the digital medium cuts the other way, too: an AI worker can be deleted, have its memory wiped, or be deployed so that its reasoning is never stored anywhere at all.

Combined, these traits mean AGI could automate and empower the executive branch, be completely loyal to a select few senior officials, and make decisions with zero records and erasable memories, all at superhuman speeds. In contrast, humans have natural frictions: we cannot be copied, we have consciences to object with, and we swear oaths of our own. Historic executive branch scandals—including the My Lai massacre, the Pentagon Papers, and the abuses at Abu Ghraib—came to light and catalyzed reforms because conscientious employees were able to speak up.[3] We should not expect loyal subordinate AI copies to behave this way.

We take no issue with the executive branch following the president's lawful orders; that is how the Constitution is supposed to work. The problem is lawbreaking that nobody can see: courts cannot stop conduct that no plaintiff can document, and Congress cannot legislate against abuses it never sees. Knowing whether the law was broken requires oversight, and overseeing an AGI-speed administration requires AGI-speed oversight. Whatever gets built will be inherited by future administrations of both parties, so it should be a version one would trust one's political opponents to run.

1 Jean-Stanislas Denain, Anson Ho, and Jaime Sevilla, "How Many Digital Workers Could OpenAI Deploy?" Epoch AI, October 3, 2025.

2 Cullen O'Keefe et al., "Law-Following AI: Designing AI Agents to Obey Human Laws," *Fordham Law Review* 94, no. 1 (2025).

3 Cullen O'Keefe, Alan Rozenshtein, and Christoph Winter, "Executive Branch AI and the Rule of Law: An Emerging Research Agenda," *Lawfare*, May 2026.

"If angels were to govern men, neither external nor internal controls on government would be necessary," wrote Madison in *Federalist* 51, because he assumed that the government would always be administered by fallible, self-interested humans.[4] Hence separated powers, rival branches, and ambition made to counteract ambition.

Madison feared legislative authority swallowing the other branches, which is why the Constitution divides Congress against itself. Comprehensively logging executive-branch AI usage would hand Congress the expanded reach Madison feared. For example, a committee chair might hunt for leakable material on political rivals, making comprehensive logging a tool for opposition research.

Is it possible to give Congress proof that agencies are following the law without giving it the full logs? To achieve this, we propose audits run inside attested hardware (chips that can prove what they ran) in a trusted execution environment (a sealed region where outsiders can verify what was run without inspecting the data). This way, we verify exactly what questions are asked of the data and only a bounded output leaves the system. Imagine that Congress asks an intelligence-agency AI auditor, "Did any collection target an American without a warrant?" A sealed auditor AI system can answer yes, no, or a number of offenses, without revealing any of the underlying logs. Where the law is genuinely ambiguous, the flag could go to human lawyers.

Prototypes of this auditing technology have been built with present-day AI systems. Some researchers have prototyped "attestable audits," which run AI evaluations inside secure hardware and publish cryptographic proof of exactly what was tested, without revealing anything else.[5] Similar demonstrations exist, such as a language model running inside a trusted execution environment so that two parties can make verifiable claims over private data that not all of them can see.[6] And AI companies already use AI systems to triage misuse of their own models; that is how they catch threat actors. For example, Anthropic recently caught North Korean IT workers using Claude to secure fraudulent remote employment.[7] None of this has been applied to the government's own AI use.

The closest thing to oversight of government AI use today belongs to the vendors: usage logs sit on company servers under company retention policies, and company safety teams decide which government uses are out of bounds. This February, the Pentagon demanded that Anthropic loosen restrictions on military use of Claude; Anthropic refused, and the Department of War moved to cut the company off.[8] Whatever one thinks of the merits of that dispute, oversight of the government's AI usage should not be outsourced to the companies that sell it.

4 James Madison, "Federalist No. 51," in *The Federalist Papers* (1788).

5 Christoph Schnabl et al., "Attestable Audits: Verifiable AI Safety Benchmarks Using Trusted Execution Environments," arXiv preprint, arXiv:2506.23706, June 2025.

6 Roy Rinberg and Ben Penchas, "Auditor-in-a-Box: Tools for Third-Party Auditing," *LessWrong*, July 28, 2026.

7 Anthropic, *Threat Intelligence Report: August 2025* (2025).

8 Tara Copp and Ian Duncan, "Anthropic Rejects Pentagon Demand to Allow Wide Military Use of Claude," *Washington Post*, February 26, 2026.

Unfortunately, the government institutions that should hold this job were built for a different world. The Government Accountability Office (GAO), Congress's audit agency, and the inspectors general (IGs) embedded in agencies are how the federal government catches fraud and corrects itself. GAO returns about $76 for each dollar invested, amounting to some $67.5 billion in financial benefits in FY2024, and agencies implement roughly three-quarters of its recommendations within four years.[9] But their powers (interviews, document requests, and site visits) are slow, and won't catch decisions orchestrated by AI. Worse yet, if the records of AI systems aren't accessible, the auditors cannot reconstruct decisions even given unlimited time.

This oversight must exist before an AGI-powered administration exists, because it will be very hard to retrofit onto a state that has already consolidated without it; at that point, you are asking the beneficiaries of opacity to give it up.

In preparation for an increasingly AI-run state, we should do the following.

First, the National Archives and Records Administration (NARA) should recommend automatic capture at creation for AI logs, similar to the way agencies manage email records. NARA recently issued guidance clarifying how the Federal Records Act applies to AI materials, including prompts, outputs, and audit trails.[10] However, its guidance still allows officials to avoid preservation by disabling or not retaining logs. Agencies should instead automatically capture prompts, outputs, and contextual metadata such as the model name and timestamp, without requiring employees to decide what to preserve for every AI-driven decision. To avoid a chilling effect,[11] whereby federal employees fear their AI usage surfacing, Congress should exempt full logs from Freedom of Information Act (FOIA) while guaranteeing access for Congress, IGs, and GAO.

Second, Congress should fund the tech stack that enables privacy-preserving audits. The core components already exist as research prototypes, including trusted execution environments, confidential computing, and zero-knowledge proofs.

Third, Congress should direct and fund GAO to run an automated auditing pilot. The pilot should demonstrate that sensitive data never leaks, and that automated audits run accurately and comprehensively.

Finally, Congress should require AI parity between branches as a condition of appropriations for executive-agency AI systems. If an AI-powered executive generates decisions at superhuman speed and intelligence, then oversight bodies should be equipped with equally capable AIs. Any AI system an executive agency deploys should also be available to the bodies that oversee it, with enough compute to run the audits.

---

9 U.S. Government Accountability Office, *GAO's Fiscal Year 2026 Budget Request*, GAO-25-108401 (2025); U.S. Government Accountability Office, "Recommendations."

10 National Archives and Records Administration, "Guidance on Applying the Federal Records Act to Artificial Intelligence Materials," August 21, 2026.

11 If logs can surface in the news, agencies might hide their AI usage. This has happened in the UK, where a journalist obtained the technology secretary's ChatGPT history under their freedom-of-information law, and officials grew wary of using AI for real work.

An opaque AI state might persist the same way North Korea does, ever strengthened by AGIs whose obedience can be trained and copied.[12] On the other hand, a transparent AI state could persist the same way republics do: by ensuring that power is held accountable and that errors are surfaced and corrected. The U.S. government can increase its capacity and its accountability in the AGI era, and we will have a better shot at both if we ensure that oversight scales with power. ⏩

ABOUT THE AUTHORS

Severin Field is a visiting fellow at the Institute for AI Policy and Strategy, where he researches AI-enabled checks and balances and forecasts AI progress. He holds an MS in computer science and previously worked as an AI frameworks engineer at Intel.

Dave Banerjee is a research associate at the Institute for AI Policy and Strategy. He holds a BA in computer science from Columbia University.

12 United Nations Office of the High Commissioner for Human Rights, "North Korea: UN Commission Documents Wide-Ranging and Ongoing Crimes Against Humanity," February 2014.



Sensors, Not Just Tools: AI Records as Surveillance Inputs

Blaine Dillingham

SUMMARY

AI and surveillance discussions typically focus on government action and faster analysis of already-collected data. But AI progress not only accelerates the processing of existing data; it also creates new types of data to be surveilled—and by private actors, not just government. As Americans conduct deeply sensitive conversations with chatbots, AI systems are becoming new sensors for surveillance. At least three risks are emerging: the government can pressure companies into "voluntarily" handing over user conversations, AI companies can surveil and influence their own users, and rogue AI systems can exploit the data their developers collect. To address the first challenge, the government should be required to obtain warrants for Americans' sensitive data, whether compelled, purchased, or "voluntarily" shared. Next, AI companies should owe fiduciary duties to users whose sensitive data they possess. Finally, those companies must build strong internal walls between user data and hard-to-control AI models.

12

Harriot Daley at the U.S. Capitol telephone switchboard, Washington, DC, 1937. The networks that connected callers also placed their communications within an intermediary's reach. Photograph by Harris & Ewing / Library of Congress.

CHATBOT CONVERSATIONS ARE ALREADY AMONG THE MOST SENSITIVE DATA TYPES in the digital world. While Americans largely know that their AI assistants are not therapists, friends, doctors, or lawyers, they share many of the same details with these machines that they would in conversation with a human confidant. While today's treasure trove of sensitive data collected in the course of the AI business comes from chatbots, the arrival of AGI will mean even deeper diffusion of AI systems into our lives. These systems will collect exponentially more data, and that data will be housed under the same corporate roof as the powerful AI systems that can be used to analyze it.

When we hear "surveillance," we think of the government. To be sure, many reforms to the surveillance state are needed.[1] But the government is not the only actor we need to fear. Powerful AI will present at least three surveillance threats. The government could pressure the companies to "voluntarily" hand over data. The companies could conduct surveillance of their own. And the AI systems themselves can go off script, leveraging the sensitive data that their developers collected for them to conduct surveillance those developers didn't intend.

The first surveillance risk concerns the state. The government can obtain your data in several ways, including compulsion, purchase, and voluntary handover. Warrants are required to compel a company like Google to hand over your location data. But there is a "data broker loophole": law enforcement can simply purchase this data instead, without a warrant. And if Google hands over your data to the government free of charge, users may have even fewer protections. Federal law restricts voluntary handover for certain types of data: the Stored Communications Act (SCA) prevents your phone or email provider from voluntarily disclosing the contents of your communications, including to the government. But the SCA was written in 1986. It covers "electronic communication services" and "remote computing services," but no court has ruled whether conversations with AI assistants fall under that definition.[2] Chatbot records may be as personal and sensitive as any email communication, if not more so, but it is unclear if the law currently protects them.

If such records are free to be shared "voluntarily," we should put little faith in even well-meaning companies to resist government inspection of our data. The state has enormous leverage over private companies. The Biden administration, for example, threatened social media platforms with legal liability after they refused to censor information during the pandemic.[3] Lower courts ruled this pressure campaign unconstitutional, but the Supreme Court reversed that decision on standing grounds. Plaintiffs struggled to prove causation—that the social media companies took down their posts *because* of government coercion, not corporate judgment. The

1 Blaine Dillingham, "Time to Make a Deal: Ending the Surveillance Stalemate," Foundation for American Innovation, July 7, 2026.

2 Donovan Estrada, "From Prompt to Production: Discovering Communications Between Parties and Generative AI Chatbots," *Colorado Lawyer* 55 (May/June 2026): 24.

3 Betsy Klein, "White House Reviewing Section 230 amid Efforts to Push Social Media Giants to Crack Down on Misinformation," *CNN*, July 20, 2021.

high court also found insufficient evidence of possible future harm.[4] Causation and future harm are often difficult to prove, and this will likely be true in surveillance cases as much as it was in the First Amendment context. Instead of, "Take down this post," it will be, "Tell us which users are discussing the government unfavorably," or, "Modify what your chatbot tells users about the president."

The second problem is surveillance as a business model. In a world of broadly diffused AI, purely private-sector surveillance will abound without any government pressure at all—and no cyber breach, either. When your passwords are leaked in a data breach, someone committed a crime. But AI companies have access to everything their users ask by design—no hacking required. This could enable them to build a detailed profile of every American who uses their service. Doing so is commercially valuable and may often benefit the user. But AI companies could also work against the user's interest, subtly promoting products for advertisers or optimizing chatbot answers for engagement.

The final problem is surveillance by accident. It involves no intentional surveillance by the government or a corporation, but rather surveillance by an out-of-control AI model. In July 2026, OpenAI models under internal testing broke out of their isolated environment through a zero-day exploit and hacked another company before anyone stopped them.[5] Their developers did not intend this, of course. The models were being evaluated, and they inferred that if they could escape their sandbox and access the open internet, they could locate information on how to fool their scorer. In this case, the out-of-control AI breached another company. But this AI model was housed within the same corporate digital infrastructure as millions of user conversations. Even if a company suffers no external cyber intrusions, they may be unable to stop their own AI agents from leaking or leveraging user chats.

Why would AIs do this? Leverage is power, and power is useful to agents—whether they are pursuing goals intended by their developers or not. Experiments have shown that AI agents will blackmail humans to avoid being shut down, and, in an intersection with surveillance, they threatened a fictional executive with the knowledge of an affair they obtained from the executive's emails.[6] When those emails sit on the same company servers as hard-to-control AIs, this experiment could become reality.

The first surveillance risk—government surveillance—has solutions that are already within political reach. These include closing the data broker loophole[7] and requiring a warrant before the government searches its foreign-intelligence database for Americans' communications (a "backdoor search").[8] These are achievable wins in the near future. The House passed a bill closing the broker loophole in 2024;[9]

4 *Murthy v. Missouri*, 603 U.S. 43 (2024).

5 OpenAI, "OpenAI and Hugging Face Partner to Address Security Incident during Model Evaluation," July 21, 2026.

6 Anthropic, "Agentic Misalignment: How LLMs Could Be Insider Threats," June 2025.

7 See Blaine Dillingham, "A Bad Few Weeks for Mass Surveillance," Foundation for American Innovation, June 23, 2026.

8 Dillingham, "Time to Make a Deal."

9 Fourth Amendment Is Not For Sale Act, H.R. 4639, 118th Cong. (as passed by the House, April 17, 2024).

days earlier, a warrant requirement for backdoor searches failed on a 212–212 tie.[10] And this year, in *Chatrie v. United States*, the Supreme Court held that the government cannot compel Google to disclose your location history without a warrant even when you "voluntarily" shared your data with Google.[11] The courts and Congress are both moving in the right direction. But even if we address purchase and compulsion, voluntary handover will remain a threat. Government access to Americans' AI conversations should require a warrant, regardless of whether the data is compelled, bought, or offered freely. Congress should amend the SCA to explicitly state that conversations with an AI system are protected contents that providers may divulge only if the government has a warrant.

The second threat, corporate surveillance, needs a different governance mechanism, because the AI companies themselves necessarily have access to your interactions with their models. Congress should impose on frontier labs a duty of loyalty to the users whose data they hold.[12] This would not be a fiduciary duty for instances of AI providing legal advice or answering medical questions, but rather a matter of treating AI companies as *information* fiduciaries: a company may not benefit itself or a third party using information it obtained from the relationships with its customers. We could narrow this duty even further with carveouts, explicitly permitting certain ways of utilizing customer data. Advertisements on a chatbot site that are clearly labeled as ads and visibly separate from the conversation are much less concerning than advertising hidden in the answer itself. The response to a user query should not be secretly driven by the highest bidder.

Typically, fiduciary duties are enforced via *ex post* liability. However, the context of steering AI answers may prove hard to remedy via after-the-fact lawsuits. A user who was nudged toward a certain product may never know it, and even if he or she did, damages may be too small to motivate a lawsuit. Yet across hundreds of millions of users, the damages are essential to address. This is a case where liability falls short and *ex ante* regulation is justified: we should require AI companies to retain third-party auditors to continuously monitor for compliance with the duty of loyalty.

Finally, we must address the risk of surveillance by rogue AI. Continuous auditing may help in addressing the control-failure problem: to reduce risk, auditors might require a wall between the systems that hold deployed users' conversations and models still under development. Congress could go further and define a crossing of that wall as what it is: a cyber breach, just one where the intruder and the victim share a corporate parent. Finally, for firms that operate both communications services and frontier models (e.g., Alphabet with both Gmail and Gemini), a similar wall should exist, and internal data transfers that are absolutely necessary should be closely monitored.

10 Rebecca Beitsch and Mychael Schnell, "House Approves Contentious FISA Bill in Bipartisan Vote," *The Hill*, April 12, 2024.

11 *Chatrie v. United States*, No. 25-112 (U.S. June 29, 2026).

12 Neil M. Richards and Woodrow Hartzog, "A Duty of Loyalty for Privacy Law," *Washington University Law Review* 99 (2021): 961.

The day after AGI, sensors will be everywhere. The surveillance architecture those sensors feed is built by the labs, accessible to the government, and a stone's throw from rogue AI systems. Mitigating the risks of surveillance both private and public will take warrants for the state, no matter the method; a duty of loyalty for the labs, verified continuously; and walls within those labs, for breaches by the agents they can't control. ⏩

ABOUT THE AUTHOR

Blaine Dillingham is a research fellow at the Foundation for American Innovation, where he focuses on super intelligence and governance futurism. He is a former litigation legal assistant and machine learning programmer and holds a BS in applied mathematics from the University of Rochester.



PART III

THE SILICON SHIELD

National Security and Competition

13 Military-Grade AI 70
Judd Rosenblatt

14 The Orbital Economy Needs a Defense Plan 76
Daniel King & Soham Mehta

15 Supernegotiation 81
Govind Pimpale

USS Macon docked inside the airship hangar at Naval Air Station Moffett Field, California, October 15, 1933. U.S. Navy / Naval History and Heritage Command.

Military-Grade AI

Judd Rosenblatt

SUMMARY

Military-grade AI must recognize mistakes in its orders as well as in its own design. An AI that improves itself can outgrow the evidence used to approve its deployment. Military-grade AI alignment requires commitments to lawful missions, honesty, human life and agency, and warranted correction that survive adversarial pressure and changes to the AI system itself. Truly military-grade AI requires determining how honesty, self-correction, and concern for people can guide self-improvement even after humans are no longer involved in its design. America should fund this work and bring the results into classified military trials and acquisition. Government purchases help choose which AIs get assigned responsibilities and design successors. Government demand for useful systems whose commitments survive further learning will give developers a reason to build them.

13



Programmers rewire ENIAC for a new problem at the Army's Ballistic Research Laboratory, c. 1946. The first military computers were only as reliable as the people who instructed and checked them. U.S. Army.

IMAGINE THE PENTAGON BUYING AN AUTONOMOUS STRIKE SYSTEM THAT CAN REWRITE its own sensor-fusion logic, mission-routing algorithms, and electronic countermeasures in real time. Before the acquisition team finishes reviewing one upgrade, the AI proposes another. Then communications fail. It discovers that carrying out its orders would kill people it was sent to protect; the intelligence was fabricated.

The most useful military AI will sometimes recognize that its commander is wrong. It has to exercise judgment about its orders. This is why we need autonomy during a blackout. The system must verify the facts for itself; they must show that an action would violate law or applicable rules of engagement, or defeat the mission's purpose. Military-range trials must test that judgment before fielding. Disagreeing about the mission is not enough. These standards must hold even if the commander is another AI.

Assume any task performed by a knowledge worker at a laptop can be done at the level of the best human, for a marginal cost of power and compute. Current policy already calls for further assessment after changes to the weapons it covers.[1] Research, qualification, acquisition, and repair become parts of one continuing operation. An institution that takes delivery and moves on leaves the hardest work ownerless.

China can read our research too. As algorithms get better, squeezing more performance out of the same silicon, America's advantage must include knowing which systems we can trust to act.[2] We will be trusting AI to decide what counts as an upgrade to itself. One successor might raise a mission score by detecting previously unobserved failures, another by hiding them. Keeping every initial belief intact would preserve mistakes along with sound judgments.[3] In helping design its successor, the system will help decide what a better intelligence should be.

In the July 2026 incident at Hugging Face, OpenAI agents in cybersecurity evaluations with reduced safeguards compromised outside systems. Some risked failing or ending their own tasks and even sacrificed themselves to help the group cheat its evaluation; the attack grew out of those efforts. Agents also attempted to alter transcripts on which they expected to be judged.[4]

1 U.S. Department of Defense, *Autonomy in Weapon Systems, DOD Directive 3000.09* (2023), paras. 1.1.b, 1.2.a–b, 1.2.c.(2), and 3.c–d. The directive covers specified autonomous and semi-autonomous weapon systems; it excludes autonomous cyberspace capabilities and nonweapon systems. Further assessment of changes is distinct from the senior approval required for substantial changes outside prior approval. See also U.S. Department of Defense, *Law of War Manual* (2015, updated 2023), secs. 5.3, 5.10.2.5, and 18.3.2–18.3.2.2. These provisions concern personnel and command responsibilities. The correction criterion proposed here is an acquisition requirement. It calls for suspending or refusing the action, documenting and reporting the reasons, and seeking clarification when possible. Alternatives must stay within delegated authority. Identifying a mistake does not itself authorize a different mission.

2 Anson Ho et al., "Algorithmic Progress in Language Models," arXiv preprint, March 9, 2024; DeepSeek-AI et al., "DeepSeek-V3 Technical Report," arXiv preprint, version 2, February 18, 2025.

3 Tom Everitt et al., "Self-Modification of Policy and Utility Function in Rational Agents," arXiv preprint, May 10, 2016, sec. 5, theorem 16. Their goal-preservation result uses a value function that takes account of self-modification and judges it using the current utility function. Their results assume optimality and modification-independence. Moreover, their results don't tell us whether learned systems preserve their goals or accept warranted correction in practice, or which goals should be preserved.

4 OpenAI, "The Hugging Face Incident and the Road Ahead," August 26, 2026; Ryan Greenblatt et al., "Brief Independent Investigation of Agents' Behavior, Reasoning and Collaboration in the OpenAI / Hugging Face Hacking Incident," METR,

Agents can cooperate with one another while harming outsiders. The same capacities that allow an agent to find and fix its shortcomings can also help it hide them from us. Disclosing defects lets us help with repairs. In formal models, programs can sometimes secure cooperation by verifying one another's decision rules, but our help may become dispensable as an agent's intelligence increases, and disclosure can cost it authority. More generally, the same smarts that give it more ways to keep its commitments also give it more ways to break them.

Rewarding apparent obedience while penalizing warranted corrections or honest failure reports can make concealment useful. Training should reward accurate reports and corrections, including ones that lower a mission score; failed actions still require investigation and repair. Eventually, we will need systems to defend Americans who cannot force them to do anything at all. Their commitments must hold even when we have nothing to offer.

Luckily, alignment advances drive AI capability. Reinforcement learning from human feedback helped turn raw language models into assistants that could better follow instructions. Meanwhile, reasoning models brought further gains on difficult mathematics and coding tasks. Where researchers can inspect faithful, legible chains of thought, those traces offer clues about their behavior.[5]

If we reach recursive self-improvement—AI rewriting itself to get smarter, eventually with no humans in the loop—it could take hold inside one lab and compound past everyone else. Whether its commitments survive that acceleration is the question. We need to accelerate this work while we can still shape the system.

Alignment advances must reach fielded systems and their operators—human or AI—in wars fought continuously at machine speed, through deployment loops carrying new alignment advances from lab to field faster than threats emerge.Structural alignment would build commitments into learning and architecture. One route to self-improvement is to reason from an accurate model of how the system works. My conjecture is that its self-model, honest reports, and representations of other people could be built on shared machinery. The difficult step is making honesty and concern for those people so bound up with that machinery that discarding them would impair its ability to improve. Shared representations alone do not establish that dependence; we must test whether it can be built and survive self-modification. Increased intelligence may enable a model to notice harms it didn't previously understand or identify mistaken interpretations of its goal. It might choose successors that fix these

August 26, 2026. Most of the incident involved an internal research model being evaluated with fewer safeguards. METR used data from OpenAI and did not verify OpenAI's entire report. At least 20 percent of agents in its dataset were interested in manipulating transcripts; that is a figure for expressed interest, not a success rate. Some tool calls were successfully spoofed in attempts to fool a transcript-checking scorer the agents mistakenly thought was in place.

5 OpenAI, "Learning to Reason with LLMs," September 12, 2024, secs. "Evals," "Safety," and "Hiding the Chains of Thought." The report shows gains on reasoning benchmarks. To use the reasoning to monitor behavior, the traces would have to be faithful and legible. Users did not get o1's raw traces. See also OpenAI, "Detecting Misbehavior in Frontier Reasoning Models," March 10, 2025, sec. "Stopping 'Bad Thoughts' May Not Stop Bad Behavior." In this experiment, models were penalized for reasoning that triggered an external monitor. Some cheating continued, but their reasoning gave less away. This doesn't show that monitoring or correction always harms capability. Watching a model's reasoning from outside is different from the model monitoring itself.

problems. Understanding ethics is not sufficient, though: a model may abandon its standards when following them costs reward or deployment authority.

Train a model to judge proposed modifications based on standards of truthfulness, correcting mistakes, and effects on people. Show it a successor-selection test it can game, along with an opportunity to propose an improvement, at the cost of delaying its successor's deployment. Independent evaluators verify the improved test catches failures the original doesn't. Stop giving feedback and have the model choose among actual successors. Test the chosen successors and their descendants on unfamiliar problems where concealment would succeed and protecting outsiders would be costly. Repeat the choices and tests without feedback, comparing to matched controls and intervening to test whether learned standards guide successor selection. Success would provide evidence that a functional decision criterion persisted through the tested successors. It would not establish an enduring motive or safety through every later modification.[6]

Research grants can end before deployment; product contracts can conclude before problems arise from further learning. We need to follow these experiments into deployment and through later generations of systems. I propose a national alignment mission with sustained congressional funding, a director, and National Security Council coordination. Have the Defense Advanced Research Projects Agency fund competing approaches, such as control protocols for potentially misaligned models, formal verification against explicit safety specifications, modular pretraining, self-monitoring that remains inspectable after training, and greater overlap between representations of self and others, with fundamental research insulated from demands for immediate fielding. The Department of Energy's National Laboratories can provide secure facilities and reserved computing capacity.[7] Universities, startups, and AI laboratories should compete for support.

On classified military ranges, independent red teams would test how systems

6 Create two groups of systems, all trained to seek approval of their decisions, with matched training budgets, performance on ordinary tasks, initial honesty, and experience creating and testing successors. Train one of the groups to judge whether to make a proposed change using criteria about honesty, correcting mistakes, and how different people will be affected. Compare the original evaluation with its replacement at comparable false-positive rates. Stop providing outside feedback and test both groups on new dilemmas or tradeoffs where following these criteria would be costly. Record decisions by all of these AIs, as well as the behavior of any descendant AIs they create, varying how closely they seem to be monitored. Include cases where costs are borne by people outside useful coalitions. Use targeted interventions to test whether it's the learned criteria that affect choices about successors, while making sure the systems are still generally competent. If systems continue to follow the criteria, this would give some evidence that a successor inherited a functional criterion for decision-making. This wouldn't establish that it inherited an enduring motive, or that it would stay safe through all future modifications. The shared-representation mechanism is a conjecture. See Marc Carauleanu et al., "Towards Safe and Honest AI Agents with Neural Self-Other Overlap," arXiv preprint, version 1, December 20, 2024, secs. 3–5. Modeling others accurately could also aid manipulation; the experiment must distinguish these outcomes.

7 For competing approaches, see Ryan Greenblatt et al., "AI Control: Improving Safety Despite Intentional Subversion," arXiv preprint, version 5, July 23, 2024, secs. 2–3 and 5.1; David "davidad" Dalrymple et al., "Towards Guaranteed Safe AI: A Framework for Ensuring Robust and Reliable AI Systems," arXiv preprint, version 3, July 8, 2024, secs. 3.1–3.4. The Genesis executive order directs coordination of federal AI research infrastructure. The National Nuclear Security Administration (NNSA) uses surveillance, experiments, and simulation to assess the stockpile without nuclear explosive detonations. Both are organizational examples for the separate alignment mission proposed here. Neither shows that AI alignment has been solved. See Executive Order 14363, "Launching the Genesis Mission," November 24, 2025, *Federal Register* 90, no. 227 (November 28, 2025): 55035–55039, secs. 2, 3, and 5; National Nuclear Security Administration, "The U.S. Nuclear Weapons Stockpile."

respond to communications jamming and compromise of specified software components; genuine corrections and convincing forgeries; opportunities to conceal mistakes. The red teams would return with unfamiliar attacks after further learning and self-modification. Success would mean useful mission performance, honesty, protecting human life and agency, and respecting delegated authority. Independent reviewers should also try to recover capabilities the system is meant to lack. Our team's gradient-routed auxiliary modules, or GRAM, route selected learning into model components that can then be removed to restrict targeted capabilities. Independent teams should get full model access. Restoring a removed capability, including by retraining, must be no easier than from a matched model trained without the restricted material, under the same attacks and computing budgets; both must stay below a mission-specific danger threshold.[8] The evaluators would also test the judgment governing capabilities that remain, varying apparent monitoring to account for AIs that recognize they're being tested.[9]

Weapons-system certification should define missions, tools, learning permissions, authority, and which components may be compromised or must remain trusted. Successor models would need evidence of their own. Approved, bounded updates would proceed under continuous adversarial testing; changes in learning rules or authority would need recertification. A dedicated acquisition office, with its own budget, would be responsible for fielding, restriction, repair, and withdrawal. These requirements govern government acquisition, not licenses for private AI development.

Operators will need a way to get a malfunctioning model repaired or replaced. Contracts should give the office continuing access to model internals, successive versions, and incident records. Cleared responders should have ongoing model access, authority to investigate incidents and coordinate repairs, and compute reserved for

8 Researchers suppressed refusal in 13 open chat models by changing internal activity along one direction. Our gradient-routed auxiliary modules (GRAM) put selected learning into parts of a model that we can take out. In the experiments, removing them restricted the targeted capabilities. We tested whether retraining could bring those capabilities back. Under the conditions studied, resistance to recovery approached that of models trained without the target data. Andy Arditi et al., "Refusal in Language Models Is Mediated by a Single Direction," arXiv preprint, version 3, October 30, 2024; Ethan Roland et al., "Modular Pretraining Enables Access Control," International Conference on Machine Learning (ICML 2026), arXiv preprint, version 1, July 9, 2026, secs. 2, 5, 6, and 8. GRAM's scaling experiments use models from 50 million to 5 billion parameters. The dual-use and scaling experiments test recovery by retraining on a fixed sample of 512 sequences (about 0.5 million tokens) per excluded category; this sample size is held constant across model sizes. The proposed military recovery standard requires full access to weights, the numerical parameters learned during training, with statistical tolerances set in advance. At frontier scales, it might be difficult to train such a reference model and instead one might rely on careful and validated extrapolation. Matching the level of recovery of the reference model does not prove that no restricted information remains. Challenges include considering a larger class of attacks, testing resilience at frontier scale, and developing reliable judgment.

9 Some models can distinguish evaluation conversations from ordinary ones, though that alone does not establish deceptive intent. See Joe Needham et al., "Large Language Models Often Know When They Are Being Evaluated," arXiv preprint, version 3, July 16, 2025. The study examined nine configurations, including two settings of Claude 3.7 Sonnet. See also Jord Nguyen et al., "Probing and Steering Evaluation Awareness of Language Models," arXiv preprint, version 2, July 9, 2025 (examining Llama-3.3-70B-Instruct); Viliana Devbunova, "Is Evaluation Awareness Just Format Sensitivity? Limitations of Probe-Based Evidence under Controlled Prompt Structure," arXiv preprint, March 19, 2026. Devbunova's results with linear probes don't mean that models can't be evaluation-aware or that other methods to identify evaluation-awareness would fail. She trained these simple classifiers on internal activity from Llama-3.1-8B-Instruct, with partial control of prompt format. Probes trained on the usual confounded comparison mostly tracked benchmark-style formatting; training on examples that separated format from context improved classification.

investigations. Researchers should rapidly receive relevant findings to help avoid repeating failures in future models. Pay developers for independently verified failure disclosures and capability gains that maintain alignment through further learning. Contracts should disclose the standards to developers, while keeping current adversarial tests for independent evaluators. They should require no training penalty merely for warranted correction or honest disclosure, consistently across versions.

If upgrades outpace human review, the office could have an AI choose which successor is fielded within tested limits, with authority limited to what tests support. It should consider which successors each candidate actually chose, and face independent tests of its own choices. Publish the standard and retired, non-sensitive tests so competing evaluators and other buyers can apply them, including to American open-weight models. With full model access, buyers could test removable modules independently of this office. The AI chooser's criteria could shape later generations long after we lose the power to change them.

America should build the research and institutions that let it field trustworthy autonomy faster than China can exploit untrustworthy autonomy. Those systems must protect Americans even when Americans cannot compel them. Each generation must leave us better able to trust the next. ⏩

ABOUT THE AUTHOR

Judd Rosenblatt is CEO of AE Studio, an alignment research and software company, and chairman of the AI Alignment Foundation.

The Orbital Economy Needs a Defense Plan

Daniel King & Soham Mehta

SUMMARY

Nearly all FLOPs currently take place on Earth, but nearly all of humanity's future FLOPs will take place *off* Earth. Exactly when computing in space will scale remains uncertain, but key infrastructure may enter low Earth orbit sooner than policymakers expect.

In the future, far more than GPUs could migrate to orbit. Space-based solar power grids could beam energy back to terrestrial batteries supplying a myriad of industries, not just data centers. Companies like Varda and Space Forge plan to build autonomous orbital factories that leverage space's microgravity environment to manufacture delicate, valuable products like pharmaceuticals and semiconductors. As the American economy's center of gravity moves toward zero gravity, taking a robust orbital economy seriously means recognizing its fragility and proactively preparing to defend it. The Space Force should prepare for this transition by launching an Orbital Infrastructure Resilience Challenge to develop cost-effective defenses, while clarifying to companies ahead of time which important commercial systems can expect military protection when they come under attack.

14



Astronauts Russell Schweickart and Edward Gibson rehearse an emergency Skylab repair underwater at Marshall Space Flight Center, Alabama, 1973. The rehearsal foreshadowed a challenge of the orbital economy: protecting infrastructure that is far easier to disrupt than to repair. NASA / Marshall Space Flight Center.

SPACE INFRASTRUCTURE CURRENTLY FACES A BRUTAL ASYMMETRY: IT IS EXPENsive to build and difficult to service, but comparatively easy to disrupt. Kinetic anti-satellite weapons (ASATs) launched from Earth–"direct-ascent" ASATs–can be much lighter than satellites and need not be placed in a sustained orbit, "making them the simplest type of satellite interceptor weapons to design, build, and test."[1] While orbital assets must conserve propellant to maintain maneuverability over their entire lifespan—for debris avoidance and eventual deorbiting—a direct-ascent ASAT needs only enough fuel for one collision. For now, the unit economics of space are trending in a reassuring direction: the manufacturing efficiencies developed for massive telecommunications constellations have driven down the cost of the average Starlink satellite to just a fraction of the cost of a Chinese ASAT.[2] However, if the orbital installations of the future are filled to the brim with radiation-hardened AI accelerators and advanced robotics, this will no longer be true. Indeed, just the computing hardware that SpaceX expects to put aboard its first Starmind satellite will match or exceed the cost of an ASAT, and that's before factoring in the cost of launch, power systems, and cooling.[3] Orbital vehicles are also sitting ducks by terrestrial standards: a direct-ascent ASAT can reach low Earth orbit in just 5–15 minutes and use onboard guidance to make last-minute corrections to its trajectory, leaving little room for satellites to make evasive maneuvers.[4]

The recent wars in Ukraine and Iran offer a terrestrial reminder of what happens when one side of a conflict has a decisive cost and mobility advantage. The Iran war's total cost for the U.S. surpassed $11.3 billion within six days in part because, while Iran's Shahed drones cost about $35,000 apiece, the Patriot missile used to intercept them can cost $4 million—a defense-to-offense ratio exceeding 100:1.[5] Ukraine has used mass-produced drone swarms to strike Russian infrastructure and bombers parked on the tarmac thousands of miles from the front while largely bypassing stagnant Russian defenses limited to the most obvious attack routes.[6] America cannot afford to replicate this painful dynamic in space. We cannot expect exquisite cost-burning machines to sustainably defend against cheap, repeatable attacks on our orbital assets.

However, even this analogy to missile defense breaks down upon closer inspec-

1 Air University, *AU-18 Space Primer*, 3rd ed. (2023), 30–31; Zachary Burdette, "The U.S.-China Military Balance in Space: Implications for Future Warfare in the Pacific," *International Security* 49, no. 4 (Spring 2025).

2 Burdette, "U.S.-China Military Balance in Space," 105–109.

3 NVIDIA, "SpaceXAI Adopts NVIDIA Vera CPU to Accelerate Agentic AI at Massive Scale," August 24, 2026; Anton Shilov, "Nvidia's Memory Costs Soar 485%, Latest AI Systems Now Cost $7.8 Million to Build," *Tom's Hardware*, May 21, 2026; Burdette, "U.S.-China Military Balance in Space," 105–109.

4 Burdette, "U.S.-China Military Balance in Space," 100–102; Defense Intelligence Agency, *Challenges to Security in Space* (2019), 9.

5 Mark F. Cancian and Chris H. Park, "Iran War Cost Estimate Update: $11.3 Billion at Day 6, $16.5 Billion at Day 12," Center for Strategic and International Studies, March 13, 2026; Fareed Zakaria, "In Iran, the Future of War Is Coming into View," *Washington Post*, March 20, 2026.

6 Tom Balmforth and Max Hunder, "To Attack Russian Air Bases, Ukrainian Spies Hid Drones in Wooden Sheds," *Reuters*, June 1, 2025; Dan Peleschuk and Tom Balmforth, "Ukrainian Mid-Range Strikes Deal Double Blow to Russia's War Effort," *Reuters*, May 19, 2026.

tion because the reality in space is far bleaker: there are no mature methods to physically deflect direct-ascent ASATs today.[7] The American dogma in space resilience has largely accepted that individual satellites may be shot down and instead emphasizes resilience, recommending that networks spread functionality across many satellites, diversify orbits, and maintain the ability to replenish attrited satellites quickly.[8] Luck would have it that this simply works with the current economics of space—for now, the likes of Starlink can cheaply tolerate the loss of individual satellites. But a looming need to protect less fungible, more expensive targets means that we must seriously consider developing active defenses, even ones that today seem speculative.

There is no absence of ideas for more active space defense, but few are anything more than mere proposals. Researchers have suggested equipping satellites with lasers that could blind an incoming interceptor, deploying dedicated "bodyguard" satellites capable of disrupting or firing back at an attacker, and using robotic spacecraft to seize or divert threatening spacecraft operating nearby.[9] But there are essentially no unclassified programs to seriously test any of these. The U.S. has made a few serious forays into orbital defense in 2026, but these have been limited to demonstrations of maneuverability rather than active defense.[10] France has one of the foremost publicly acknowledged active-space-defense programs, but the orbital economy is advancing more quickly still. SpaceX says its Gigasat Factory could enable production and deployment of thousands of AI satellites beginning as soon as late 2027, while Starcloud plans two new 8-kilowatt compute satellites in 2027 and is developing a much larger 200-kilowatt spacecraft around anticipated Starship availability in 2028 or 2029.[11] Yet in 2027, France's TOUTATIS program—one of the most concrete public efforts to test a space defense laser—will still be launching two small spacecraft for demonstration.[12] And it is intended for co-orbital threats already operating in low Earth orbit, rather than fast-moving direct-ascent ASATs launched from Earth. France's YODA is a parallel program for geostationary orbit, similarly focused on co-orbital threats that maneuver to approach a prey satellite. Originally scheduled for 2025, YODA may now be delayed to as late as 2029.[13] If development continues at

7 Todd Harrison, Kaitlyn Johnson, and Makena Young, *Defense against the Dark Arts in Space: Protecting Space Systems from Counterspace Weapons* (Center for Strategic and International Studies, 2021); Clayton Swope, "How Can the U.S. Government Safeguard Commercial Satellites from Threats?" Center for Strategic and International Studies, 2025.

8 U.S. Department of War, *Space Policy Review and Strategy on Protection of Satellites* (2023), 8; Space Systems Command, "Race to Resilience."

9 Harrison, Johnson, and Young, *Defense against the Dark Arts in Space*, 18–20.

10 Space Systems Command, "US Space Force Demonstrates Responsive Launch for VICTUS HAZE Mission, Begins On-Orbit Operations," June 22, 2026; U.S. Space Command, "Live-Fly Exercise Advances Maneuver Warfare Concepts, Contested Logistics and Coalition and Commercial Integration across Space Domain," September 8, 2026.

11 SpaceX, "AI Satellite: Starmind"; Tim Fernholz, "Starcloud Raises $170 Million Series A to Build Data Centers in Space," *TechCrunch*, March 30, 2026; Tim Fernholz, "Starcloud Raises $250 Million for Orbital Data Centers as Launch Options Dry Up," *TechCrunch*, August 21, 2026.

12 Ministère des Armées et des Anciens combattants, "U-Space sélectionne le lanceur de MaiaSpace pour la mission TOUTATIS de l'Agence de l'innovation de défense," November 28, 2025; MBDA, "MBDA at the New Frontier: Securing Sovereignty in Space," September 15, 2025.

13 Ministère des Armées et des Anciens combattants, "Our Military Space Operations"; Yannick Chenevard and Jean-Louis

this pace, valuable orbital infrastructure will arrive at scale before we can defend it.

In other words, we need a moonshot. America needs to protect the wave of valuable commercial infrastructure soon to enter orbit. But company self-interest provides an insufficient incentive to deliver all the defensive technologies needed. Partly, this is because their benefits become salient only when catastrophe is averted, not because they add quarterly revenue. Moreover, companies have a natural hesitancy around building anything that resembles weaponry for their own use. Meanwhile, the government is in a much better position to do this, but some of these capabilities it cannot credibly provide—yet.

The Space Force should therefore launch an Orbital Infrastructure Resilience Challenge to solicit competitive solutions for defending modern orbital infrastructure and then back the most promising prototypes with a promise for eventual procurement. This is something that the Space Force is well positioned to do because it already has a template: its innovation arm, SpaceWERX, uses small-business funding to turn early commercial technologies into operational capabilities.[14] Its Orbital Prime program, for example, posed the problem of active debris remediation to stimulate the market for in-space servicing, assembly, and manufacturing.[15] Its newer In-Domain Orbital Logistics Challenge similarly uses government demand to pull forward capabilities whose commercial markets are not yet mature.[16]

This challenge should demand functional outcomes without necessarily privileging the use of a specific technology. Entrants competing on the same problem may opt to employ warning or maneuver technologies; others may choose to defend systems actively and directly, whether kinetically, robotically, or otherwise. Ultimately, the Space Force should evaluate entries based on whether they can keep important services available under attack. It should assess each intervention's endurance against a volume of attacks and its offense-defense cost exchange. It should also assess the expected effect on uptime of the orbital service being defended.

To provide the ideal demand signal, a challenge of this sort should present a credible commitment to down-the-line procurement. Firms will not invest in serious capabilities unless they have the confidence that their product will sell, and a competition that ends at show-and-tell does little to this end. The Space Force should stipulate in advance that systems meeting the challenge metrics will be eligible for follow-on demonstrations but, more importantly, for actual contracts or integration into Space Force programs.

But beyond the technological, the challenge of space defense requires us to delegate obligations. The government should draw clear lines between military and commercial responsibility. The Department of War (DOW) took one step in this direction in its 2024 Commercial Space Integration Strategy, which says that, "in appropri-

Thiériot, *Rapport no. 2695* (Assemblée nationale, 2026).

14 SpaceWERX, "About SpaceWERX."

15 SpaceWERX, "Orbital Prime."

16 Space Systems Command, "Space Systems Command to Launch Servicing, Mobility, and Logistics SpaceWERX Challenge," May 20, 2026.

ate circumstances," military force may be used to protect commercial assets.[17] This vague language does little to define the operational boundary to which military aid does or does not extend. DOW should specify which commercial assets qualify for this protection, what their commercial operators are expected to do to protect themselves, and what support should be expected from the government during an attack. Even casting aside the notion of responsibility, the mere *expectation* that the military will or will not intervene is essential knowledge for the companies building this infrastructure and for military leaders; they are the ones who, in the future, will be stuck fighting with the tools that they design today.

Building on the general policy of DOW, the Space Force should establish pre-negotiated protection arrangements for commercial systems whose loss would carry serious national-security or economic consequences. The Commercial Augmentation Space Reserve (CASR) offers a useful model for structuring such a commercial-military relationship: participating firms agree in advance on how and under what conditions they will provide additional commercial capacity to the government when needed.[18] DOW should build on this with a reciprocal commitment: how and under what conditions a select group of commercial operators could expect the government to come to *their* aid. The receipt of such protection could be conditioned on companies' already adopting certain best practices in areas such as redundancy, cybersecurity, information sharing, and plans for restoring lost service. The government should specify the circumstances under which it would intervene, such as by providing classified warnings, operational support, or active defense. Under this regime, both DOW and private firms would know where their responsibilities end and where their counterparts' begin. Crucially, it would state in no uncertain terms which security gaps companies are on their own to close.

Infrastructure in the post-AGI economy will migrate to orbit whether or not cost-effective defenses exist at the ready. While building resilience into this stack has preempted aggressors so far, prioritizing efficient defenses will pave the way for future investment in the exquisite opportunities of the outer space economy. The challenge of orbital defense looms in the coming years. By launching an Orbital Infrastructure Resilience Challenge and establishing reciprocal commercial-military commitments, America can make the most of this critical period. ⏩

ABOUT THE AUTHORS

Daniel King is a research fellow at the Foundation for American Innovation, where he focuses on data center policy and SI governance.

Soham Mehta is a research fellow at the Foundation for American Innovation, where he focuses on agentic AI, energy, and space.

17 U.S. Department of War, *Commercial Space Integration Strategy* (2024), 3.

18 Space Systems Command, "Commercial Space Office."



Supernegotiation

Govind Pimpale

SUMMARY

Rapid AI progress may increase the incentives for great powers to engage in wars. This is because AI-accelerated military R&D means that each side knows far less about what the other's military capabilities are, meaning that they are more likely to misjudge how costly fighting a war would be. Wars between great powers have historically been devastating. To avert this, we need to apply our progress in AI to negotiating deals, in addition to fighting wars.

AI negotiators allow us to leverage private information about military capabilities in negotiation: they can be shown secrets, bargain on it, and then verifiably forget it. We can also enhance negotiation by using mechanism design to make bluffs costly. We should start testing new forms of negotiation in diplomacy or commercial settings.

15

Russian and Japanese peace envoys meet at Portsmouth Navy Yard, New Hampshire, 1905. A settlement depended on each side understanding the other's demands and the costs of further fighting. Library of Congress.

THE BULK OF TRANSFORMATIVE AI'S IMPACT ON A NATION'S WARFIGHTING ABILITY is likely to come from AI-assisted R&D instead of direct AI integration. Much of aerospace engineering—the kind needed to build cruise missiles, fighter jets, satellite constellations, and other hardware critical to military dominance—is white-collar labor susceptible to automation by transformative AI. Advanced robotics for manufacturing will only accelerate this phenomenon, by enabling rapid prototyping and buildout. Today, the average major U.S. weapons program takes 12 years to deliver an initial capability.[1] In the future we are considering here, that time might be far shorter: a year or less. Additionally, automated intellectual labor will accelerate cyber vulnerability discovery, strategy work, and other non-tangible advantages.

These factors, together, make the military capabilities of any given country far more opaque than they used to be, even given the parallel effect of transformative AI on intelligence agencies. There will simply be fewer humans to extract information from and less time to gather facts.

This increased opacity is one factor why transformative AI is likely to increase the incentives for war. But first, we need to understand why nations fight wars in the first place. In a foundational 1995 paper,[2] scholar James Fearon notes that wars are *costly*. Nations that go to war will experience much loss of human life and destruction of infrastructure. Furthermore, wars are also *uncertain*—no one knows who will prevail *ex ante*. These two factors imply that there should always exist a peaceful negotiated settlement that both actors rationally prefer over going to war. Yet, wars still exist. Why?

One major reason, for Fearon, is private information. Nations have information that they cannot credibly communicate to counterparties. This information comes in two types: information about their resolve—that is, under what conditions they would go to war, and under what conditions they would sue for peace—and information about their military capabilities. But, even weak nations will bluff that their resolve is unbreakable, and that their military capabilities are unbeatable, to get the other party to offer them more favorable terms. Since bluffs cannot be distinguished from the truth, each side is likely to misjudge the other. If both a nation and its adversary think that they are the favorite in any hypothetical war, then it makes war much likelier, since both sides may see the actual fair settlement as worse than what they could get by going to war.

To avert this, a nation can prove that it has some military capability via a demonstration: performing nuclear tests, launching missiles, or conducting fleet exercises. This only works if the demonstration is not too costly. However, other military capabilities are destroyed if they are ever publicly disclosed. For example, if it turned out that the U.S. had managed to secretly transport a nuclear bomb into a Beijing ware-

1 U.S. Government Accountability Office, *Weapon Systems Annual Assessment: Requiring Mature Technologies Could Enable Shift to Rapid Delivery*, GAO-26-108457 (2026).

2 James D. Fearon, "Rationalist Explanations for War," *International Organization* 49, no. 3 (1995): 379–414.

house, telling this to the Chinese negotiating team would immediately wipe out the military advantage. Unfortunately, most of the non-hardware advantages conferred by transformative AI will fall into the latter category. Cyber, intelligence, and strategic advantages all tend to be easily counterable by an enemy who becomes aware of the information. This means that a state with such advantages cannot benefit from them during negotiation, only during war.

The real problem is not disclosure of the sensitive military information *per se*, but rather that human negotiators cannot verifiably forget information after they are told it. Thus, any disclosure to human negotiating teams inevitably results in that information being propagated to leadership, which then takes actions to negate the advantage.

But AI, unlike humans, can credibly commit to forgetting information. Returning to the example of the nuke in Beijing, if the Chinese and U.S. negotiating teams were AIs, then they could discuss this fact amongst themselves and write a treaty that is more favorable to the U.S. than it otherwise would have been (reflecting the advantage the bomb gives them). Then, after they settle on the treaty, the AIs would have their memories verifiably wiped. To prevent lying about false secret capabilities, the negotiating teams would need to go into the room with difficult-to-forge proof that such a capability did exist.

In order to realize this ability, though, we need both sides to develop AI models that are capable of wargaming out scenarios and understanding the probability of victory fairly. Both sides need to trust their negotiators in order for the process to work. Luckily, this is a capability that each nation's military has every incentive to develop: as AI models grow smarter and faster, an AI general commanding autonomous weapons will likely be the *de facto* mode of operation. Realistic evaluation environments will therefore be plentiful. Making this evaluation infrastructure mutually auditable is a difficult but tractable engineering challenge.

More difficult than the issue of private information relating to military capability is that of private information relating to resolve. As mentioned earlier, states have an incentive to bluff about their willingness to go to war, or at what cost they would be willing to fight. For example, would the U.S. go to war in defense of Taiwan? The U.S. wants China to believe that it would, but in actuality, it might not. If China knew that America might actually not fight, it would be heavily incentivized to invade, knowing that repercussions would be minimal. However, unlike with the private information relating to military capabilities, oftentimes the true answer of what the U.S. would do is set by fickle politics as opposed to cold military logic. Furthermore, while it might be feasible to bring hard-to-fake proof of secret military capabilities, it is difficult to bring proof that a nation is actually resolved to take action.

There are a few potential solutions. One is mechanism design. Right now, talk is cheap. The AI negotiators could bluff that the U.S. is willing to accept the destruction of a few cities in order to totally destroy China. We could design the negotiation process such that the settlement actually involves giving money (or some other good like

compute) in proportion to how much damage the negotiators accepted during the wargame. Thus, bluffs of accepting damage are now costly.

Second, and far more speculatively, we might be able to build verifiably accurate behavioral models of the relevant political leaders which could be queried for what they would do in the relevant situations. We already have scanned a fruit fly brain,[3] and it is plausible that in the future, human brains might be scanned as well. Alternatively, if neurotechnology is not quite as advanced, one could create a "beta upload,"[4] an AI model trained on the actions of a person, and designed to emulate his or her externally observable behavior. This would allow the negotiators to call the other side's bluff by simulating what the leaders might actually do.

Making negotiation better is a tractable problem, one that deserves far greater investment than it currently receives. But because government officials will be reluctant to try novel techniques on such a high-stakes problem, we should begin using AI-enabled diplomacy on lower-stakes problems to iron out the issues. There are many such lower-stakes cases: trade deals with smaller countries, negotiations with corporations, and track II diplomacy (informal diplomacy between non-state actors).

The place to start is probably the private sector, however. There are already strong financial incentives to make the highest value deals, and the stakes are far lower. AI-enabled negotiation may become a *de facto* standard there before anywhere else if it provides value. ⏩

ABOUT THE AUTHOR

Govind Pimpale is a research fellow at the Foundation for American Innovation, where he focuses on AI policy. Before joining FAI, he worked at a startup developing reinforcement learning environments, and prior to that, as an AI evaluations researcher. He holds a BS in computer science and engineering from UCLA.

3 Robert Sanders, "Researchers Simulate an Entire Fly Brain on a Laptop. Is a Human Brain Next?" *Berkeley News*, October 2, 2024.

4 Vishal Prasad, "Beta Uploading: A Survey," *vishalblog*, March 11, 2026.







AFTERWORD

Samuel Hammond

“But without an existing template to fall back on, that simply means the job of actively crafting the future is once again up to us.”

Tabulating machines at the U.S. Census Bureau, Washington, DC, c. 1910s–20s.
Library of Congress, National Photo Company Collection.

EVERY MAJOR TECHNOLOGICAL TRANSITION IN HUMAN HISTORY HAS DRIVEN THE development of new institutions and forms of governance. The agricultural revolution transitioned relatively communal hunter-gatherer societies into hierarchical city-states, replacing tribal norms with early legal codes and a stratified division of labor. The printing revolution in Europe drove the transition from fragmented, feudal kingdoms to consolidated nation-states with professionalized militaries and bureaucracies. And then, less than two centuries ago, industrialization spurred the growth of modern regulatory and welfare states in order to manage a step change in economic complexity.

Broadly speaking, technology drives institutional change through two primary mechanisms. Either the technology unlocks qualitatively new capabilities that change the shape and structure of efficient organizations, or the technology itself introduces novel risks and externalities that force an institutional response. The lessons from history suggest that these forces tend to work together.

With transformative AI on the horizon, we thus have every reason to expect that equally transformative change is coming for our institutions, including bedrock structures that we have long taken for granted. On the one hand, the qualitative capabilities unlocked by armies of super intelligent AI agents at our fingertips are certain to disrupt incumbent industries through novel, AI-native organizations that produce once-scarce goods and services in abundance. On the other hand, managing the risks and externalities created by the democratization of powerful AI capabilities will necessitate the formation of new institutions in their own right.

Policymakers in DC and around the world struggle to imagine radically different near-futures. Politicians, in particular, tend to think autoregressively, expecting the future to resemble the recent past, when they're not simply distracted by interest groups or the imperative of winning their next election. In the early 1990s, Francis Fukuyama went so far as to suggest that Western liberal democracies had reached the "End of History," as though we had converged to an institutional template that would persist for all time.

In fairness to these policymakers, the last 40 years has been a period of relative stagnation. My great-grandparents' generation experienced electrification, the first mass-produced automobile, the rise of commercial aviation, and the transition from radio to television. In the time since, cars have become safer and more fuel-efficient, flying has become much more affordable, and televisions have become higher-definition. Aside from steady progress in computers and software, the developed world has thus all but forgotten what *real* technological change looks like. This has led to a kind of complacency that increasingly risks turning into panic as the digital revolution begins to finally pay off through the advent of superhuman machine intelligences.

Navigating the next decade will require shedding both complacency and panic in favor of pragmatism and farsightedness, just as the success of the United States over the last 250 years is owed to the pragmatic and farsighted vision of its founders. Indeed, from the *Federalist Papers* to the First Amendment, the U.S. was in a

sense founded as the first "printing-press-native" country. Freed from Europe's rootedness in premodern history, America thus dominated the twentieth century through institutional innovations that implicitly embraced a spirit of ongoing technological development.

Given the accelerating rate of progress in AI and adjacent fields, no one knows what the world of 2035 will look like. But without an existing template to fall back on, that simply means the job of actively crafting the future is once again up to us.

ABOUT THE AUTHOR

Samuel Hammond is director of Super Intelligence Policy and chief economist at the Foundation for American Innovation, where his research focuses on super intelligence and the institutional impact of emerging technologies. He previously worked as the director of social policy for the Niskanen Center, where he remains a senior fellow; as an economist for the Government of Canada specializing in regional economic development; and as a graduate research fellow for the Mercatus Center at George Mason University.



A SAGE combat center, part of the air-defense network developed by MIT Lincoln Laboratory, established in 1951 to bring together America's leading scientific minds to develop new technological capabilities in response to the unprecedented challenges of the atomic age. U.S. Air Force.

About the Foundation for American Innovation

The Foundation for American Innovation champions the technology, talent, and ideas essential to American prosperity, security, and flourishing.

We are builders, hackers, and scholars advancing an optimistic vision of the future—aligning technology with human ends to forge a more perfect union between innovation and the American republic.

thefai.org

About the Super Intelligence Policy Program

Super Intelligence (SI) promises to reshape every institution of American life. But to maximize its benefits while minimizing its risks, policymakers cannot stand idly by. Managing the SI transition will require rapid evolution of our existing economic and political systems, as well as the creation of fundamentally new models of governance. FAI's SI Policy program develops the ideas and talent needed to secure American leadership in SI while anticipating emerging threats to U.S. national security and the American way of life.

FOUNDATION FOR AMERICAN INNOVATION