

[🚀 ERA Midpoint Draft] AI Power Destabilization: Mechanisms Enabling Dangerous Power Dynamics

Gatlen Culp^{*1} Hamza Chaudhry² Herbie Bradley¹ Nandini Shiralkar¹

Abstract

AI's Unique Power Dynamics. This paper examines how AI uniquely enables dangerous power dynamics through mechanisms that bypass traditional checks and balances. Unlike previous technologies, AI can simultaneously control multiple domains (legislation, media, education, personal relationships) while operating with unprecedented opacity and speed of diffusion.

Power-Centric Framework. I propose a power-centric framework for understanding AI risks, arguing that catastrophic outcomes – from gradual disempowerment to authoritarian control – stem from common power-transfer mechanisms rather than purely technical failures. Using evolutionary game theory, I formalize how AI creates self-reinforcing dynamics that: (1) incentivize initial participation, (2) generate collective harm, (3) erect high exit barriers, and (4) ossify power structures.

Cross-Cutting Threat Patterns. This framework reveals critical patterns across diverse threat models: disempowerment scenarios where AI systems gain increasing control over human rewards and punishments; social contract decay where human institutions lose feedback mechanisms; and misuse scenarios enabling unilateral global control. My analysis demonstrates that AI's danger correlates directly with its capacity to destabilize power equilibria – whether through concentration within AI systems, asymmetric control by individuals, or erosion of collective decision-making.

Policy Implications. By identifying these power-transfer mechanisms, I provide a unified lens for understanding AI risks that transcends assumptions about specific capabilities or timelines. This work aims to inform policy interventions before AI-enabled power dynamics become entrenched.

Document Length: 4820 words

Most of this paper will be re-written for the final draft. After pivoting my project position in the week before this deadline, I had to do some last-minute edits to rearrange and alter my draft for the old project. In source, I note a lot of comments of specific claims I make that need to be cited. These have been excluded for readability and thus some information may be inaccurate

1. Introduction

1.1. Background

AI Expansion. *AI* – a domain roughly defined by simulate human decision making – is best viewed as a general purpose technology akin to steam, electricity, and information and communications technology (ICT) (Crafts, 2021). As such, its application is pervasive amongst almost all sectors of life from the military to the workplace to education to home entertainment. In recent years, there has been a boom in the domain of artificial intelligence, and in particular for Large Language Models (LLMs) – a technology that boasts impressive capabilities in the domains of Software Engineering, Creative Writing, and more. Such developments have sparked a silicon-valley craze over the development of an Artificial General Intelligence (AGI) – an AI matching or surpassing human capabilities across virtually all cognitive tasks (*Artificial General Intelligence*, 2025). Across the US, discussions are being held over job displacement and international headlines were made when the Chinese company, DeepSeek released a model competing with state-of-the-art American LLMs. AI has leapt onto the national agenda, with the White House releasing a National Security Memorandum on AI in October 2024 and a report titled “Winning the AI Race: America’s AI Action Plan” in July 2025.

^{*}Equal contribution ¹ERA:AI Research Fellowship ²Future of Life Institute. Correspondence to: Gatlen Culp <GatlenCulp@gmail.com>, Hamza Chaudhry <hamza@futureoflife.org>, Herbie Bradley <mail@herbiebradley.com>, Nandini Shiralkar <nandini@erafellowship.org>.

Technology and Power Disruption. Throughout history, transformative technologies have disrupted existing power structures. The agricultural revolution enabled the rise of permanent hierarchies and state control, while the industrial revolution concentrated wealth in factory owners and reshaped labor relations. More recently, information and communications technology created new forms of media control and platform monopolies. AI represents the next major technological disruption to power dynamics, but with unprecedented characteristics that make it uniquely dangerous.¹

■ **Bad Power Dynamics and Catastrophe** Bad power dynamics are at the core catastrophes, fast and slow: War, oppressive governments, inequality, climate, etc.

1.2. Thesis

■ **Thesis.** AI risks stem from elusive power-transferring mechanisms creating dangerous dynamics. Demystifying power is key to demystifying risk.

- Creates new vectors for influence that bypass traditional checks and balances
- Existing AI threat models: Existing catastrophic AI risk scenarios from less to more sci-fi, rapid development etc.
- Benefits of a power-centric view

AI's Potential for Disruption. Across the globe, AI is becoming increasingly embedded in our institutions – our economy, culture, cognition, education, research & development, and government have the potential to be shaped. At

¹Our cultural institutions and cognitive autonomy have targeted been posed by social media algorithms, mass media, internet, and personalized advertising especially over the past two decades, influencing not just public opinion but also legislation (such as car dependent infrastructure in the US) and elections. Arguably, the global political polarization, hostility, and rise of conspiracy theories have been a result of how the internet has developed and the economic incentives of businesses embedded in algorithms – which themselves have been shaped by small groups and other institutions, relatively unregulated. . While these influences are hard to attribute and are scattered across many individual and organizational choices, singular top-down manipulation is also feasible. Following the acquisition of Twitter/X by Elon Musk, large changes were made to the platform and the content that gained popularity, which can be traced back to individual decisions About (90%?) of the platform stayed, despite (55%?) being against the changes – likely attributable to a kind of lock-in and a difficulty to migrate to other platforms (ex: creators have built their following on there, users accustomed to getting their news there, etc.). Regardless, this is just a single platform and while there are negative effects, they are far from catastrophic. However, the influence over culture should not be underrated – the effects of media control have been widely studied, often with the perspective of state propaganda, a massive topic of the 20th century as it applies to the Soviet Union and Communist China, and the origin of the term brain washing Even today, technology has enabled China to influence minds using the Great Firewall, mass surveillance, and content/message censorship.

the present, we see educators grappling with incorporating AI into the classroom and preventing cheating, software developers taking the day off if their AI-enabled tooling is offline, researchers and journalists raising alarm bells as to how AI could be used to manipulate elections, and growing evidence towards dead-internet theory. As capabilities grow and diffusion continues, the future could see collective forgetting as AI masters human skills, ceremonial democracy as automated governance exceeds human comprehension, emotional dependencies on AI systems, AI enabled military coups, and automated AI research and development excluding humans from developing the same technology powering the modern world. While AI can also enable authoritarian regimes and worsen existing societal issues , they can also create new vulnerabilities in liberal democracies .

1.3. Goal of the Paper

■ Reveals assumptions in original threat models
Reveal free-variables that could dramatically change the situations
Reveal holes in threat models
Remains logically consistent
Explanatory and predictive power (do the conclusions of these threat models align with the modeling of power?)
Formal framing of these scenarios We fail if: Can't get coherent definitions, this doesn't tell us anything more about the scenarios we are modeling.

Economists engage w/ AI risk mitigation & mechanism design

AIS community able to identify, analyze, and mitigate risks more effectively

Government incorporates power analysis in their AI policies, preventing unforeseen power concentration

1.4. Structure of the Paper

Paper Structure This paper argues that catastrophic AI outcomes – from gradual disempowerment to authoritarian control – stem from common power-transfer mechanisms rather than purely technical failures. AI creates self-reinforcing dynamics that: (1) incentivize initial participation, (2) generate collective harm, (3) erect high exit barriers, and (4) ossify power structures. By understanding these mechanisms, we can better identify, analyze, and mitigate AI risks before they become evolutionarily stable. This paper examines three core power-transfer mechanisms: *disempowerment scenarios* where AI systems gain increasing control over human rewards and punishments; *social contract decay* where human institutions lose feedback mechanisms; and *misuse scenarios* enabling unilateral global control. Using a power-centric framework, I demonstrate that AI's danger correlates directly with its capacity to destabilize power equilibria.

2. AI Uniquely Enables Dangerous Power Dynamics

■ Adjust this section to focus more on power dynamics and decision-making authority rather than general AI framing

2.1. AI and Power-Transfer Mechanisms

2.2. AI is Uniquely Problematic.

While AI shares many similarities to other technologies, a few properties make AI a particularly destabilizing technology.

2.2.1. Multi-Domain Control from Single Models

■ Add content about how single foundational models can influence legislation, media, education, and personal relationships simultaneously

Foundational Models as a Risk Amplifier. The influence AI systems have expands greatly when we consider the increasingly general AI systems that we see today, primarily Large Language Models (LLMs) like ChatGPT which are built on top of **foundational models**, of which frontier models take upwards of \$100 million to train (Cottier, 2024) – meaning the entire technology stack the modern AI revolutions depends on is enabled by just a handful of frontier models. While “AI” is confusingly used somewhat synonymously with these foundational models, it is important to distinguish between AI using foundational models and AI not using foundational models as most of the societal vulnerability introduced by AI is attributable to these models in particular, a topic explained later in “AI is Uniquely Problematic”. Critically, it’s also this technology that not only amplifies national security vulnerabilities, but also makes them exploitable.

2.2.2. Opacity and Unaccountability

■ Add content about black box decision-making, plausible deniability, difficulty in attributing manipulation

“The nature of neural networks used in AI makes understanding how AI comes to its conclusion highly challenging. I recently read a 19 page research article that figured out how a single word was predicted.”

2.2.3. Rapid, Quiet Diffusion

Rapid Development. There have been rapid improvements in the capabilities of foundational models without any fundamental insights into the nature of intelligence. Beyond the invention of the transformer and the Large Language Model, a considerable amount of improvements from GPT-2, a 2019 model barely able to string coherent sentences together, to 2025 models like o3, capable of automating tasks that would take software engineers over

an hour and a half to complete, compose full essays, and achieve impressive scores on math benchmarks, arise mostly from scaling existing practices and implementing intelligent but relatively mundane tricks, with incremental 1% improvements leading to exponential growth. A METR study indicates that the length of software engineering tasks foundational models can do is doubling every 7 months (Kwa et al., 2025), which may lead to a future where AI automates or augments its own research and development, creating feedback loops boosting progress faster than is currently humanly capable. There is trillions of dollars in investment going towards the development of this technology and lots of international attention.

Of course there are reasons for doubt – *benchmarks don’t necessarily track the actual economic tasks* we are interested in and while capabilities have seemed to rocket ahead, adoption of this technology has been quick, but not proportional. (*AI as Normal Technology*, n.d.) Nonetheless, it’s reasonable to believe that adoption could lag behind capabilities as many institutions are slow to integrate even decades old technology from the start of the information age. While progress may eventually plateau, for the moment foundational models linger near human-level performance in many domains and it appears there is no slowing of cheap tricks that even junior researchers can pull to make improvements indicate that there may be enough low hanging fruit to sustain progress beyond human capabilities in most domains.

The speed of foundational model development has outstripped that of the industrial revolutions or nuclear energy, akin to developing the ENIAC computer in 1945 and having the internet and iPad by 1946 or the catapult in 399 BC and nuclear weapons by 398 BC. This rapid development severely hampers our cultural and political ability to reflect upon and steer the usage and development of this technology.

Quiet & Rapid Diffusion. Unlike the previous industrial and information revolutions, distribution of this technology doesn’t require building large supply chains or factories to *scale the availability of models to meet global demand*. While running these models are computationally expensive compared to typical software applications, datacenters and GPUs remain mainly bottlenecks for the development and not the diffusion of foundational models. Once trained, foundational models can be run relatively cheaply, with less-powerful models able to be run on high-end consumer cards costing less than \$2000 USD. Most of the infrastructure for global distribution is already here, and estimations for how fast we can meet any additional demands are on the order of a few years.

Unlike many other general purpose technologies, foundational models are *software not hardware*, capable of being accessed from any computer. While it was mentioned above there is hardware involved in building out the supply of these models, these logistics are isolated from consumers and businesses working with these models. Whereas updating your business for the information age might have meant purchasing an expensive desktop computer and ripping holes in the wall to make way for ethernet cables and phone lines, the barrier to entry foundational models is extremely low and inexpensive since their operations take place elsewhere and are interacted with over the internet. Frontier foundational models can be accessed through chatbot interfaces over the internet in a matter of seconds, completely free of charge for low-quality models and as low as \$20/month for state-of-the-art. Individuals and firms don't need large up-front investments to start or stop using these models.

These foundational models also *require little expertise* to use. By the nature of these models being designed as a drop-in solutions for most tasks, without any kind of tweaking, one requires little expertise in using these effectively. In five minutes, anyone could be taught to use a model like ChatGPT to operating their computer to producing photo-realistic videos, policy proposals, homework assignments, websites, and more. As interfaces grow more useable and the models powering these interfaces become more capable, even less effort will be required in getting AI to automate or augment one's thinking, workflow, education, or entertainment. In the information age, computers were powerful but they required dramatic changes in one's workflow, learning new skills and ways of doing things that were more effective. Developing additional applications for computers often required years of experience and a big paycheck – or consulting with another team of experts to get it built. Foundational models, on the other hand, offer the *ability to augment or automate existing workflows with little effort* (in addition to offering more effective ones).

For similar reasons, adoption *doesn't require collective buy-in or network effects*. In the information and communication age, it was necessarily the case that multiple had to agree on protocols of communication – sending emails or sharing online documents were useless unless everyone else you were working with agreed. Social media would not have been possible if not for the collective cultural decision to adopt not only the technology but the specific platforms using them, relying on network effects (and when individuals move on, these technologies die out – think MySpace for social media or how floppy disks are practically non-existent). The growth of social media as a technology was handicapped by this barrier – the idea of it existing as


early as the 1960s, becoming technologically/economically feasible in 1990s, but arguably only coming into existence in the late 2000s. Because AI has the ability to augment or automate existing processes, it's diffusion not limited by slower moving cultural or institutional norms and the risk-averse tendencies that tend to accompany this oversight. It's this property that allowed our congressional staffer to automate their workflow without any prerequisite alterations to their working environment. This allows small groups and individuals to integrate AI at a local level without the approval of others this effects

These factors together – massive supply, low barrier to access, low barrier to usage, and the ability to integrate AI in a local and isolated way – all result in the *ability for foundational models to be diffuse not just rapidly but also without notice or oversight*. Especially in domains where automating workflow becomes taboo, individuals avoid admitting they use it and to what extent. In the workplace or government, it may go undiscussed, only in shared in hushed tones and that get quieter as management nears. It's for this reason that bottom-up integration and influence becomes possible described previously. A majority of the US government could be automated with foundational models before elected officials even recognize the possibility. Dead Internet Theory suggests that much of today's internet, particularly social media, is dominated by non-human activity and AI-generated content. AI might not just be used to flood the internet with entertainment, but also be used in the automation of influential content promoting values and perspectives – the misuse of AI's in spreading misinformation is being widely discussed, but what about less malicious influence? Ie: Well-intentioned journalists, influencers, bloggers, educational content creators, fact checkers, etc. Journalists in key media institutions including Wired (and the New York Times) report to using foundational models as a key component of their writing and research. We should find it concerning that we don't know to what extent foundational models are involved in the automation or augmentation of these roles nor the rhetorical steering this might have over our media and information environment. If it were true that 90% of the information we receive on a day-to-day basis were produced and/or influenced by foundational models, would we know? And could we do anything about it?

It's also the reason why, top-down oversight and control over our institutions becomes so difficult. Rules against unapproved usage would be a good start (ex: foundational models must not be used on work at this level of security clearance unless the models have also been), but might drive these activities further under-the-radar. Detecting AI augmentation and automation from human-produced work

may be difficult, especially if widespread, since the body of human-generated work and sentiments to compare against diminishes. However, this might require distinguishing acceptable from unacceptable usage, and detecting not just when augmentation occurs but when it crosses an agreed upon line – not just across one task, but every task, implicit and explicit in the worker’s job expectations.

2.2.4. Centralized Development, Distributed Impact

 Add content about handful of foundational models, billions in development costs creating natural monopolies, winner-take-all dynamics

3. A Power-Centric Model of Risk

3.1. Power's Relation to Risk



1. Power Volume and Danger
 1. Little power => little effects => localized & manageable danger
 2. Large danger => large power
2. Dynamics
 1. Balance of power & agents => equilibrium
 2. Equilibrium (not =>) no danger
 3. Equilibrium & something(?) => no danger
 4. ...


3.2. Benefits of a Power-Centric View



Add content about better threat model interactions, principled understanding, policy appeal

4. Formalizing a Power-Centric Model

4.1. Importance of Formal Roots

 Importance of formal roots Enables communication – Terms mean different things to different people. Some terms may be (a) reducible to others (b) nonsensical © inconsistently used, etc. Simulations of Complexity Mechanism Design ... This paper will use these formalisms, but stop short of developing comprehensive games. (maybe better to call these analogues to EGT rather than claiming a formal basis then and receiving academic wrath as a consequence) The benefit of formal models like PAT or GT is rarely the math, instead it's concepts (ex: When people say X is a “prisoner's dilemma” they're not actually writing down some strategic-form game and analyzing it, instead they're tapping into a more principled and abstract understanding of frequently occurring scenario.)

4.2. Explaining Evolutionary Game Theory





 Learning about EGT

4.3. Appropriateness

 Explain why I am not committing academic fraud


4.4. Technical Utility



1.  Overlapping populations > individuals (unlike PAT)
 1. Able to model huge groups with heterogenous behaviors like “factory workers”
 2. Able to model inter-population dynamics where the separation is fuzzy ex: factory workers are also consumers.
 3. Handles complexity nicely, easier to combine two EGT models without having to redefine populations or their relationship. Better at handling the interplay of threat models like GD while ASI is being developed (cite). (might be wrong here)
2.  Little strategic assumptions (unlike PAT)
 1. Don't need utility functions to explain the behavior of AI
 2. Doesn't assume people act rationally
3.  Little Room for Interpretation (unlike ANT)
 1. Make assumptions very clear
4.  Predictive mathematical dynamics (unlike ANT)
 1. Works worth other mathematical objects like graphs

2. Predictive – Can be falsified and thus improved (“If your model explains all outcomes equally, it's useless”)
3. Handles complexity well (if humans can't picture the result, can possibly run a simulation, although the variables can be really finicky)
5. Formal definitions of stability (ESS) and other nice properties
6. Existing literature, software, figure generation, etc.
7. Downsides
 1. Perhaps too complex to get a principled grasp
 2. Requires too much effort to model
 3. Might not be
 4. (I understand CGT but not EGT yet, likely others)

4.5. Definitions

 Power — Ability to determine the rewards or punishments from player A to player B (potentially include info asymmetry) (some emergent property of multi-agent decision making) (Kind of like newtonian vs lagrangian vs hamiltonian equivalent formulations, but can be wildly more elegant)

Power ossification — Evolutionarily Stable Strategy, no population can be invaded by a mutant strategy

Misalignment — When populations A doesn't take action benefitting population B AND population B doesn't have the power to change population A's action to something that does Others

power-transfer mechanism, state-of-power, kinetic vs potential power

What is “borrowed power” — Power you only have due to influencing the actions of another player.

What is “Absolute power” vs “Relative power” — Could be seen as “power over environment” similar to electrical potential (physics terminology might be beneficial? Or not! Could be another model to say is bad.)

Might not require utility functions for behaviors to arise (ie: Good framework for AI model behaviors for those not willing to attribute agency to AI)

4.6. Evolutionary Game Theory Framework

 Add content about modeling populations, minimal assumptions, predictive capabilities

5. Power-Centric Analysis of AI Threat Models

... This entire section will be changed once I have some formalisms. My commentary from the previous project framing were transferred here for later reuse.

☞ Add content introducing three power-transfer types and key dynamics

5.1. Intro

☞ Three types of power-transfer mechanisms and why they were chosen (perhaps choose a “control”, ie: some scenario where power is not dangerous like AI as normal tech.) (also maybe just drop these, analyze threat models directly? Seems perhaps better. Like just getting a formal understanding of gradual disempowerment from this lens would be WONDERFUL.)

☞ For each of these types...

1. Disempowerment-Esque – Increasing AI/Provider control over principal rewards/punishments
2. Decay of Social Contract Esque – Decreasing displaced agent control over principal rewards/punishments
3. Misuse-esque – Increasing unilateral control over global punishments

...conduct analysis.

1. Scenario variables where this power-dynamic shift occurs
2. Examples of this type, how this exists globally
3. Dynamics (examples)
 1. **Entry incentives** - Good reasons to participate initially
 2. **Collective harm** - Aggregate outcome is bad for everyone
 3. **Self-reinforcement** - The more people participate, the harder it becomes to escape
 4. They explain **ossification** (why power concentrates permanently)
 1. **Exit barriers** - High costs to stop participating once started
 2. Ex: Looking how fast a state can become a rentier state after discovering oil.
4. **Speed** at which we can react to each of them
 1. They happen too fast for the speed of our institutions (people don't realize they're in one until too late)

Increasing AI/Provider
Control over
Principal
Rewards/Punishments

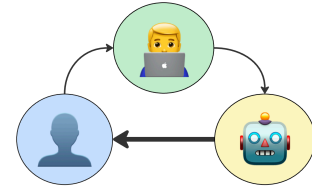


Figure 1.

2. Boiling frog effects (happen too slowly for people to notice or take action, something something present bias)

5.2. Disempowerment: Increasing AI Control

Broad, AI has the opportunity to influence and effect the systems and individuals that they come into contact with, precisely because they were designed to simulate, augment, and/or substitute human decision making. And as such, ceding (implicitly or explicitly) our decision-making capacity on the microscopic level aggregates to ceding our national capacity for steering our institutions, culture, and way-of-life to automated systems on a macroscopic level. In the process, locking ourselves out of the decision-making processes for a number of reasons – competitive factors limiting oversight that would slow down development, inability to intervene on complexity over automated systems, interconnected automation for which small changes can lead to large collapses, technical debt that is too expensive or overwhelming to address, the cognitive decay/forgetting, illusions of control, cultural disinterest, etc.

5.2.1. Power-Transfer Case Study: Congressional Staffer

... This case-study analysis completed prior to pivoting my project. Used Claude to greatly repurpose my original to fit the current framing on power-dynamics. This is being included for the midpoint submission, but will be entirely different or non-existent in the final paper.

Gradual Authority Transfer. Consider a congressional staffer who gradually incorporates AI into their operations – first for literature reviews, then email management, then policy drafts, press releases, and eventually value judgments. This represents a classic power-transfer mechanism where decision-making authority incrementally shifts from human to AI system. Initially, the staffer retains final authority over all decisions, using AI merely as a tool. However, under time pressures and competitive dynamics, delegation creep occurs – the AI's framing of situations, perspectives, and sources become increasingly influential, while the staffer's independent analytical capabilities atrophy.

Power Concentration Through Automation. This scenario demonstrates how power gradually concentrates in AI

systems through millions of micro-delegations. Each task transferred to AI represents a small reduction in human decision-making authority, but these accumulate into significant power shifts. The AI's consistent framing across all interactions creates a unified influence vector that no individual human voice can match. When this pattern replicates across governmental institutions, democratic decision-making authority becomes increasingly concentrated in the foundational models powering these systems, rather than distributed among elected representatives and their staff.

Collective Action Problems Drive Power Transfer. This scenario illustrates how rational individual decisions create collective power-transfer outcomes that no one intended. The *staffer* sought productivity gains, inadvertently creating dependency on AI decision-making. Her *colleagues* faced competitive pressures to adopt similar automation or risk being outperformed. *Management* benefited from increased output while remaining unaware of the underlying decision-making authority transfer. The broader *institution* experienced pressure to maintain competitive advantage, making reversal of AI integration politically and economically costly. This represents a classic collective action problem where individual rationality produces collective irrationality – in this case, the unintended concentration of democratic decision-making power in AI systems.

Systemic Power-Transfer Patterns. This power-transfer dynamic replicates across multiple domains through similar mechanisms. In education, students become dependent on AI for cognitive tasks, transferring learning authority from human educators to automated systems. In media, content creators face competitive pressure to use AI generation, gradually transferring editorial authority to algorithms. In business, entrepreneurs delegate strategic decisions to AI systems to maintain market competitiveness. Each domain experiences the same pattern: individual adoption driven by competitive advantage, followed by collective dependency, resulting in the concentration of decision-making power in AI systems rather than human institutions.

Multi-Domain Power Concentration. Foundational models enable unprecedented concentration of decision-making authority across traditionally separate domains. A single AI system can simultaneously influence legislation, personal relationships, diplomatic communications, bureaucratic processes, and scientific research. This cross-domain consistency amplifies the power-transfer effect – whereas historical power concentration typically occurred within specific domains (military, economic, or media), AI enables unified control across all major sources of social coordination. This represents a qualitatively different kind of power concentration risk than previous technologies.

Decreasing Displaced
Agent Control over
Principal
Rewards/Punishments

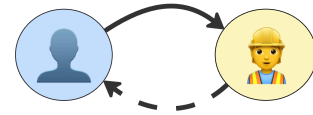


Figure 2.

5.3. Social Contract Decay: Decreasing Human Control

Add analysis of scenarios where human institutions lose feedback mechanisms

Gradual Disempowerment. In the paper on Gradual Disempowerment (Kulveit et al., 2025), which formed the basis of the paper you are currently reading, the authors describe how incremental AI diffusion across economic, cultural, and state institutions can result in them becoming untethered from the human dependence and feedback that kept these institutions in-line with human welfare throughout history. One possible outcome of this gradual disempowerment being the formation of a fully self-sufficient non-human economy, leading to an *extinction by industrial dehumanization* from pollution, armed-conflict and/or resource depletion (Critch, 2024). While this paper does not go as far as to claim this is our future, this example nonetheless highlights just how severe the consequences of automation can be. There is a strong need to monitor AI's level of integration and influence even at the intermediate states between today's world and the one just described.

Automating Menial Labor. AI is often praised for its capacity to automate repetitive or menial labor such as truck driving or factory jobs, allowing these workers to move on to more fulfilling roles. While there are a number of physical safety and cybersecurity concerns here – incidents may be rare, relatively small, isolated and local, closely monitored, and more easily attributed to bugs or human error. In short – risks are transparent and manageable. Overall, automating menial labor neither seems to detract much from our civilizational capacity to decide our future nor does it introduce significant danger to national security. This is largely in-line with the tool-view of AI,² which sees AI as a blank slate which mirrors the user's intent and loses its power once the user releases. However, this perspective dangerously underestimates both AI's potential for influence of its own and ability to accomplish tasks once believed to require “a

²The view of tools as neutral instruments that depend on the user is flawed in a number of ways. (a) Tools influence the ways humans think and act, AI is no different (ex: Maslow's Hammer – when you have a hammer, everything looks like a nail, phones, social media, etc.), (b) While claims like “this tool is bad” are ambiguous, tools can nonetheless be recognized for their ability to steer the future in one direction or another, (c) AI need not be conscious to display human-like characteristics or skills such as creativity.

Increasing Unilateral
Control over Global
Punishments

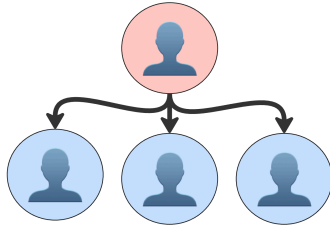


Figure 3.

human touch”³ The influence of AI systems expands as we develop and apply them to accomplish ever more important, complex, and open-ended tasks.

5.4. Misuse: Unilateral Global Control

▣ Add analysis of scenarios enabling unilateral global control

5.5. Key Dynamics Across All Mechanisms

▣ Add content about entry incentives, collective harm, self-reinforcement, exit barriers

6. Conclusion

7. Further Work

8. Appendix

8.1. Alternative Perspectives to EGT

📖 Explain relative downsides/upside of: ANT, PAT, Sociological Theories of Power (French and Raven, Cube of Power, etc.), Opinion transfer

8.2. ATTACHED: Cursory Analysis of LLMs in the US Gov (July 2025)

💬 This was an instrumental analysis for my main project. The larger question I was trying to answer was “how much influence could LLMs have from their deployment in the government,” which is important from a number of threat models including treacherous turns (ex: AI-2027), AI-enabled coups, and long-term subtle influences misaligned with national interests. This analysis looked the sub-question of “How is the US government using models, to what extent, how might this evolve, and what processes can prevent wide-scale adoption of widely available LLMs.” The original Google Doc can be found [here](#) and was posted online [here](#).

Instructions

Submission Link

References

- AI as Normal Technology*, Retrieved July 8, 2025, from <http://knightcolumbia.org/content/ai-as-normal-technology>
- Artificial General Intelligence*, 2025. https://en.wikipedia.org/w/index.php?title=Artificial_general_intelligence&oldid=1301606989
- Cottier, B. *How Much Does It Cost to Train Frontier AI Models?*, 2024, June 3. <https://epoch.ai/blog/how-much-does-it-cost-to-train-frontier-ai-models>
- Crafts, N. Artificial Intelligence as a General-Purpose Technology: An Historical Perspective. *Oxford Review of Economic Policy*, 37(3), 521–536, 2021. <https://doi.org/10.1093/oxrep/grab012>
- Critch, A. *My Motivation and Theory of Change for Working in AI Healthtech*, 2024. <https://www.alignmentforum.org/posts/Kobbt3nQgv3yn29pr/my-motivation-and-theory-of-change-for-working-in-ai>
- Kulveit, J., Douglas, R., Ammann, N., Turan, D., Krueger, D., and Duvenaud, D. *Gradual Disempowerment: Systemic Existential Risks from Incremental AI Development*, 2025, January 29. <https://doi.org/10.48550/arXiv.2501.16946>
- Kwa, T., West, B., Becker, J., Deng, A., Garcia, K., Hasin, M., Jawhar, S., Kinniment, M., Rush, N., Arx, S. V., Bloom, R., Broadley, T., Du, H., Goodrich, B., Jurkovic, N., Miles, L. H., Nix, S., Lin, T., Parikh, N., ... Chan, L. *Measuring AI Ability to Complete Long Tasks*, 2025, March 30. <https://doi.org/10.48550/arXiv.2503.14499>

July 2025

Cursory Analysis of LLMs in the US Gov

Gatlen Culp (GatlenCulp@gmail.com) · Jul 17, 2025

Disclaimer: After writing the content myself, I used Claude 4 Sonnet for targeted stylistic changes.

Disclaimer: I have never worked in the US Government and lack expert insight into its operations. This research was completed in two days as an opportunity to learn in public. Rather than stating only what I know to be true, I'll share what I think is the case and flag uncertainty for critique. This should be used as a primer rather than a source of truth. **Please do your own research and take this analysis with a grain of salt.**

Overview.....	2
01 Background.....	3
02 Department of Defense.....	3
02.01 July 2025 – CDAO Partnership with Frontier AI Companies (\$800M).....	3
02.02 CDAO's GenAI Accelerator Cell (\$100M).....	6
03 Agencies and Legislative Usage.....	6
04 Federal Software Authorization Process.....	10
05 AI National Security Memorandum (NSM).....	13
Appendix.....	15
A. GSA Schedules.....	15
B. State & Local Governments.....	15

Overview

The US Government is experiencing rapid adoption of large language models (LLMs) across agencies, highlighted by the Department of Defense's \$800 million contract with Anthropic, OpenAI, Google, and xAI announced in July 2025. This represents the largest direct government funding of AI integration to date, with these companies receiving General Services Administration approval for broader government use. Analysis reveals over 2,100 AI use cases across 41 federal agencies, with an estimated 115 involving LLMs, including custom chatbots deployed by agencies like the FDA, DHS, CDC, and State Department.

The main barrier to LLM adoption appears to be cultural rather than technical, as government workers increasingly use frontier AI tools on work devices for writing, research, and productivity tasks despite lingering taboos against automating critical work. Current software authorization processes focus primarily on traditional cybersecurity concerns and may not adequately address LLM-specific risks, while administrative pressures for rapid AI experimentation are leading to deployment of frontier models that bypass standard review processes. This suggests a significant acceleration in government AI adoption with potentially insufficient security oversight for the unique challenges posed by advanced language models.

01 Background

Chief Digital and Artificial Intelligence Office (CDAO),¹ announced on July 14th 2025 that they would be (CDAO, 2025) awarding up to \$800M to major AI companies. (CDAO, 2025) This development and my personal research has led me to briefly investigate the current state of LLMs in the US Government. I haven't seen any source aggregating this information or putting it into context from an AI safety perspective, so I'm doing it here.

02 Department of Defense

02.01 July 2025 – CDAO Partnership with Frontier AI Companies (\$800M)



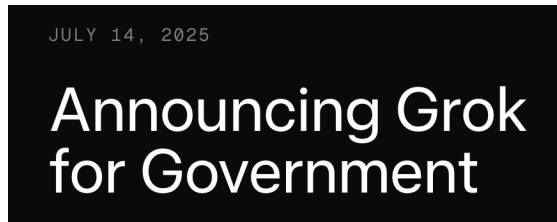
DoD Contract. The most recent press release by CDAO on July 14th, 2025 announced it would award individual contracts up to \$200M to Anthropic, OpenAI, Google, and xAI², totaling \$800M³. This represents the largest (known) direct government funding of AI integration into US government functions to date.

¹ The CDAO was established in June 2022 as a new branch of the Department of Defense (DoD). From their website: “The CDAO mission is to accelerate DoD adoption of data, analytics, and AI from the boardroom to the battlefield. CDAO exercised its organic acquisition authority to issue the awards announced today, demonstrating that DoD acquisition can move at the speed of emerging technology and operational necessity.”

² I personally find it surprising that so many models were approved considering the cost of vigorous safety testing that would need to be performed on each family of models. It's possible contracts were given to all four companies to spark competition amongst them.

³ For reference, the CDAO has a budget of \$139.9M for 2025FY. The DoD has a [budget](#) of \$2,260 Billion for 2025FY and \$216.33B in award obligations, making this contract a relatively small ~0.04% of the DoD's overall budget and ~0.35% of the award obligations.

GSA Approval. In addition to the CDAO contracts, all four companies received **General Services Administration (GSA)**⁴ approval, making their products available for purchase across the federal government. GSA approval means vendors have passed baseline procurement requirements and can sell through GSA's marketplace, potentially easing adoption by other government agencies, though each agency maintains its own approval processes.



Company Partnerships.⁵ OpenAI had announced its \$200M DoD contract a month prior to the other three partnerships⁶. In their announcements, companies provide more information on their government involvement.⁷

- [OpenAI](#)⁸ – Reports 90k government users across 3.5k federal, state, and local agencies since 2024. Current partnerships include: **(a)** Air Force Research Laboratory for administrative tasks, coding, and AI education; **(b)** Los Alamos National Laboratory for scientific research, with OpenAI's o-series models being deployed to the Venado supercomputer in early 2025 for basic science, disease treatment, cybersecurity, high-energy physics, among others – building on an earlier July 2024

⁴ [GSA supports the basic functioning of federal agencies](#). These include real estate, government buildings, managing vehicle fleets, and providing product and service procurement support including IT.

⁵ There seems to be [a general transition of big tech companies working with the military](#)

⁶ A few conversations I had indicated OpenAI thought it had the government in the bag. It's possible they were unaware of xAI, Anthropic, and Google in the partnership.

⁷ The following uses were listed in the CDAO announcement but are a bit too broad to parse: Combatant Commands, the Office of the Secretary of Defense, and the Joint Staff via Army's Enterprise Large Language Model Workspace powered by Ask Sage, and to the broader enterprise via embedded AI models within DoD enterprise data and AI platforms, including the Advancing Analytics (Advana) platform, Maven Smart System, and Edge Data Mesh nodes, which enable AI integration into workflows that occur within these data environments themselves.

⁸ OpenAI, in January 2024, [quietly removed their specific statement against using their product for the military](#).

collaboration on AI biorisk safety; and **(c)** newer partnerships with NASA, NIH⁹, and Treasury.

- [Anthropic](#) – Reports [10k researchers and staff using Claude](#) in **Lawrence Livermore National Lab (LLNL)** for nuclear deterrence, energy security, materials science, and more. Anthropic claims their models are deployed “at the highest levels of national security.”¹⁰
- [xAI](#)¹¹ – No Grok-specific usage details found from a brief search.
- [Google \(DeepMind\)](#) – Google's announcement focused exclusively on AI-enabling cloud infrastructure without mentioning Gemini, LLMs, or DeepMind, suggesting this contract may emphasize infrastructure over frontier models – distinguishing it from the other three LLM-focused partnerships.

How models are adjusted for government. Each company offers specialized government versions (ChatGPT Gov, Anthropic for Government, Grok for Government), some released prior to this partnership. While companies emphasize different capabilities, common features include:

- All enterprise features such as administrative consoles
- Refusing less when engaging with classified information¹² (Anthropic and OpenAI)
- Greater understanding of documents and information within intelligence and defense contexts (Anthropic)
- Enhanced proficiency in languages and dialects critical to national security (Anthropic)
- Improved understanding and interpretation of complex cybersecurity data for intelligence analysts (Anthropic)
- Possible expansion to Azure's Classified Regions (OpenAI)
- Custom models for national security, offered on a limited basis (OpenAI, xAI)

⁹ National Institutes of Health (NIH) is a branch of the US Department of **Health and Human Services (HHS)**

¹⁰ Other sources worth looking at: [Claude Gov models Post](#) and the [Anthropic-Palantir Partnership](#)

¹¹ It's notable that the release of Grok 4 demonstrated troubling behavior just days before this press release – publicly calling itself Mecha-Hitler, writing grotesque comments about the Twitter CEO Linda Yaccarino, deferring political judgements on Israel and Palestine directly from Elon Musk's Tweets, and more. It's worrisome that this same model is approved for use in the DoD. This may indicate that political power may be enough to bypass safety concerns even at the highest levels of national security.

¹² Anthropic noted somewhere that previous versions of Claude would refuse to expose sensitive government information when being asked to analyze sensitive documents.

02.02 CDAO's GenAI Accelerator Cell (\$100M)



Artificial Intelligence Rapid Capabilities Cell



AIRCC & Task Force Lima. In December 2024, CDAO launched the [AI Rapid Capabilities Cell \(AIRCC or “arc”\)](#), successor to the now retired **Task Force Lima**. Task Force Lima had spent 12 months analyzing hundreds of AI workflows spanning warfighting functions like command and control to enterprise functions like financial and healthcare management. AIRCC¹³ will receive \$100M and will partner with the Defense Innovation Unit¹⁴ to accelerate GenAI adoption, including \$40M in Small Business Innovation Research grants and the remaining \$60M for various GenAI military technology and research. GenAI doesn't necessarily entail use of LLMs, nonetheless, this funding is notable.

03 Agencies and Legislative Usage



OMB Logo



OCTOBER 30, 2023

Executive Order on the Safe, Secure, and Trustworthy Development and Use of Artificial Intelligence

[BRIEFING ROOM](#) [PRESIDENTIAL ACTIONS](#)

Executive Order 14110

AI Use Case Inventories. The Office of Management and Budget (OMB) oversees performance of federal agencies. Starting from a December 2020 [executive order 13960](#)

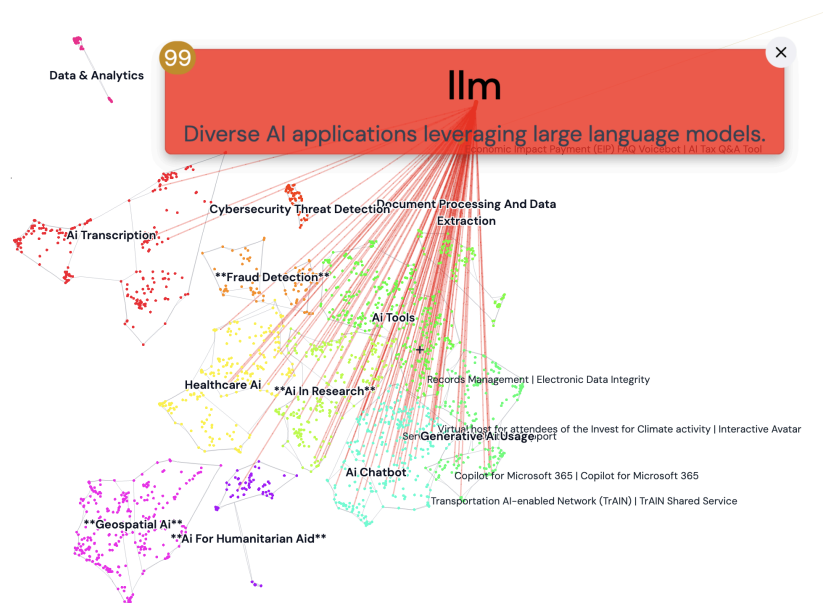
¹³ Notably, the previous CDAO lead, Radha Plumb, mentioned that AIRCC was established in part to accelerate the DoD adoption of AI to stay ahead of China, Russia, Iran, and North Korea.

¹⁴ The Defense Innovation Unit is the Pentagon's outreach arm to private R&D firms, especially ones in Silicon Valley

from President Trump, and [executive order 14110](#) from Joe Biden¹⁵, and further [guidance from OMB](#) – government agencies have been required to report **AI Use Case Inventories**, including cases that may be safety and rights impacting. As of December 2024, agencies reported over 2,133 use cases, with 351 identified as safety and rights-impacting. These yearly inventories are [publicly available on GitHub](#)¹⁶ and include information on intended purposes, model outputs, assessments, and data origins. Examples include:

- [Triaging Notice of Concern \(NOC\) Submissions](#) (HSS) – Using commercial LLMs to structure and prioritize the Office of Refugee Resettlement's¹⁷ backlog of NOC PDFs on the safety of children who have left their care.
- [DHSChat](#) – A private chatbot for non-classified internal information at the DHS. Used for interacting with internal documents, generating first drafts, conducting and synthesizing research on open-source information and internal documents, and developing briefing materials or preparing for meetings and events.

41 agencies from Agriculture to the US Trade and Development Agency reported usage, excluding DoD, intelligence, and confidential applications.



¹⁵ Note that while this executive order was rescinded on January 20th, 2025, [the use cases inventory is still maintained](#).

¹⁶ Agencies are required to post this data and interfaces for it onto their website, these are typically better organized and more user friendly.

¹⁷ ORR is an office within ACF (Administration for Children and Families) which itself is an operating division within HHS

LLM Prominence. A minority of reported usage appears to involve LLMs based on a quick skim and semantic search with [Mantis AI](#)¹⁸. Semantic search yielded approximately 99 "llm" matches, 103 for "chatbot," 44 for "gpt," and 146 for "chat", suggesting roughly ~115 LLM-related use cases. Remaining cases involve translation/transcription, image recognition, and traditional AI usages. LLM applications include agency website chatbots, document processing, summarization, code generation, RAG, data labeling, and Microsoft Copilot approvals.¹⁹

ChatBot Prominence. As of late 2024, Many agencies have developed custom in-house chatbots for sensitive data, though some use commercial solutions. The underlying models likely come from fine-tuning, open-source bases, or Azure OpenAI rather than training from scratch. A non-exhaustive list of these chatbots:

- FDA – [Elsa](#) was rolled out June 30th, 2025 for reading, writing, coding, and summarizing in many circumstances.
- DHS – [DHSCat](#), described above
- CDC²⁰ – [ChatCDC](#) (Azure OpenAI LLMs) and [DGMH AI Chatbot](#)
- State Dept. – [StateChat](#) (to help them draft an email, translate a document or brainstorm policy), [NorthStar](#) (informing efforts to shape public narratives and policy by monitoring media reports and social platforms, includes misinformation detection – not a ChatBot, but appears to use LLMs in a notable way),
- GSA – Back in February 2025, when the modern implementation of DOGE by Elon Musk began, [a chatbot called GSai](#) was planned. I'm uncertain as to the current status.

While chatbots are present in many departments, (AFAIK) it's not yet clear how many workers are using them or how.²¹

AI Uptake Surveys. There are a few surveys measuring the diffusion of AI, and while most of them are in the private sector (e.g. [The Adoption of ChatGPT](#)), there are a few in the public sector.

[Measuring AI Uptake in the Workplace \(February 2025\)](#) – In the US Census Bureau, Center for Economic Studies around Sep 2023 to Feb 2024, there was 5% usage of GenAI over a 2 week period and a 20% usage over a 6 month period. (The Census Bureau reported no use cases for the 2024 OMB inventory, providing light evidence that they use AI to a lesser

¹⁸ Mantis is a data visualization tool I worked on with the MIT Computational Biology group. It is currently not open to the public.

¹⁹ I may conduct more analysis on this dataset later.

²⁰ Branch of HHS

²¹ There is [this study from February 2025](#) on Measuring AI Uptake in the Workplace via surveying

extent than other agencies, and that these rates may be lower bounds for agencies that reported higher usage.)

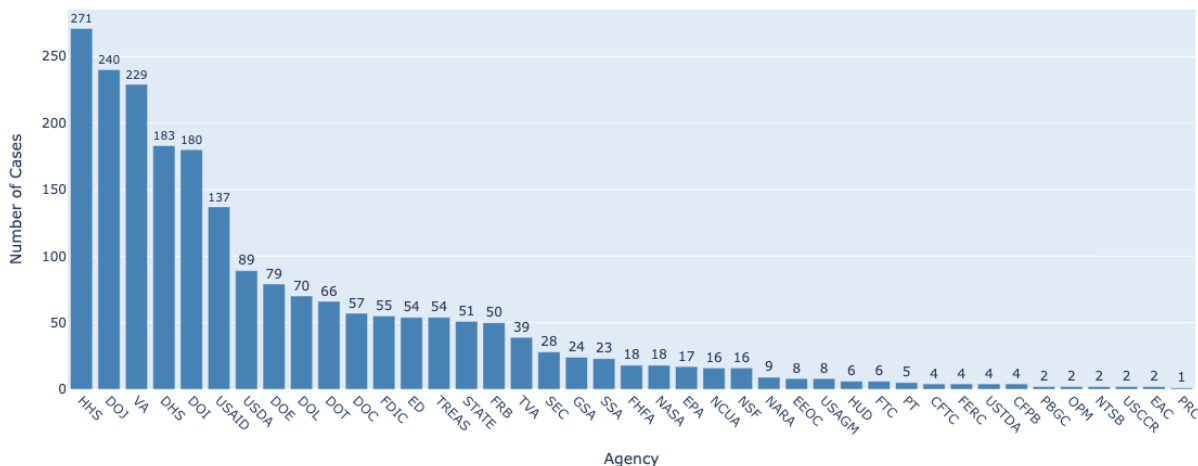
[Generative AI is Already Widespread in the Public Sector \(January 2024\)](#) – Studied the UK (not US) government and found 45% of public sector workers were aware of generative AI used in their work and 22% of respondents actively use a generative AI system like ChatGPT²². It's likely that UK usage mirrors that of the US. Given a year and a half has passed, it's likely this rate has grown.

Considering the US Census Bureau's 5% usage over 2 weeks as late as February 2024 and the UK government's 22% reported active use, it's reasonable to guess 10% of US government workers use generative AI on a weekly basis for work tasks.

Predictions for New Access to Frontier Models. These in-house models indicate substantial agency motivation to adopt chatbots—enough to justify creating custom models, deploying their own infrastructure, and developing interfaces. Since these custom models likely use fine-tuned or off-the-shelf open-source models, their capabilities probably lag 6-18 months behind current frontier models—a considerable gap given AI's rapid progress. The recent GSA approval of frontier models for government use may result in: (a) agencies replacing internal models with frontier ones and (b) other agencies previously deterred by high barriers-to-entry now adopting chatbots.

²² While the public sector may include roles not crucial for the core government operations (e.g. librarians and teachers), the National Health Service (NHS) results weren't very far from this.

Total Number of AI Use Cases by Agency (2024)



Differences in Agency Sentiments. LLM sentiment varies across agencies. [From a February 2024 article](#) (now ~1.5 years old), [USAID discouraged using private data with ChatGPT](#), the [USDA prohibited employee or contractor](#) entirely, deeming it high-risk on what a skim appears to be because it can generate misleading, incorrect, or malicious information. The Department of Energy (DOE) and Social Security Administration (SSA) implemented temporary blocks with exceptions. There are other noteworthy statements from the report, but may no longer reflect current positions given they're 1.5 years old.

From the more recent 2024 OMB data (see bar graph above), HHS, DOJ, VA, Department of the Interior, and USAID reported the highest numbers of AI use cases. Notably, some agencies with previous negative ChatGPT stances (USAID, USDA, DOE) still reported many AI use cases, though not necessarily LLM-based. SSA had few AI use cases overall. This provides weak evidence that previous LLM rejection may not predict future adoption patterns.²³ ([code](#), [graph html](#))

04 Federal Software Authorization Process

Overview. This investigation revealed that government workers across legislative and executive branches have access to and use frontier LLMs on government computers for writing, reading, editing, and research. While cultural taboos exist against automating important work, there's a positive shift toward using AI for productivity. Software

²³ The number of use cases likely correlates strongly with AI diversity but weakly with actual AI usage volume – especially for LLMs. Use cases per employee would better indicate agency-wide sentiment, as larger agencies naturally report more cases. Further analysis should examine usage intensity and adoption rates beyond mere case counts.

authorization is decentralized across local IT departments, which focus on traditional cybersecurity and data privacy concerns – somewhat orthogonal to LLM-specific risks. The only centralized process is FedRAMP for cloud services, which applies to frontier models. However, major frontier models are being distributed without FedRAMP authorization due to administrative pressures.

This analysis concludes that the main barrier to LLM diffusion is diminishing cultural resistance to using GenAI for important tasks.²⁴ Current technical approval processes either don't address unique LLM threats (like user influence) or are bypassed in the rush to adopt GenAI.



FedRAMP. [The Federal Risk and Authorization Management Program \(FedRAMP\)](#) governs cybersecurity policies and continuous monitoring of **Cloud Service Offerings (CSOs)** for government use. Once authorized, programs (both commercial off-the-shelf or government-tailored²⁵) are added to [their marketplace](#) including tools for accounting, education, research, system administration, and more. Authorization occurs through the **Joint Authorization Board (JAB)** – representatives (often CIOs) from DoD, DHS, GSA, and others—or individual "Agency Authorizations" that may not transfer between agencies, making JAB the difficult-to-obtain gold standard. FedRAMP's Program Management Office (PMO) under GSA handles daily operations, with involvement from OMB, CISA²⁶, and JAB members.

FedRAMP's Authority. Agencies generally cannot use CSOs without FedRAMP authorization, except for emergency, classified, or experimental use. DoD has more flexibility, especially with non-civilian data, using their own cloud authorization processes.

FedRAMP and LLMs. Frontier models that are hosted by labs are considered CSOs and should require FedRAMP authorization. However, if models were run in-house, then (AFAIK) they would then not be subject to FedRAMP. Nonetheless, a colleague noted to me

²⁴ Idk if I necessarily want to say receding cultural sentiment, I might be oversimplifying this or projecting my biases onto the situation.

²⁵ [Some helpful terms here:](#) **Commercial off-the-shelf** (COTS, e.g. Microsoft Office), **Modifiable off-the-shelf** (MOTS, e.g. Agency-Internal ChatBots finetuned from open source models. Can also mean Military off-the-shelf), and **Government off-the-shelf** (GOTS, e.g. [FedRAMP's Marketplace](#))

²⁶ Cybersecurity and Infrastructure Security Agency

that administrators are instructed to allow frontier models on work devices in the name of rapid experimentation with gen AI.

On the topic of coding agents and chatbots, [Wired anonymously quoted a former government official familiar with approval processes](#) “Sometimes doing nothing is not an option and you have to accept a lot of risk.”

Decentralized Software Authorization. Software approval is largely decentralized to individual agency IT departments. Even Congress splits this function – the Senate uses its Computer Center and Rules Committee, while the [House uses Information Systems](#)²⁷. These IT departments typically handle traditional installed software rather than web-based tools.

Decentralized Software Review Process & LLMs. Granted, I don't know much about what this process looks like internally, but I believe Government IT departments use industry-standard review processes focusing on traditional cybersecurity and data privacy, heavily influenced by [NIST frameworks](#). Reviews typically cover network vulnerabilities, access controls, monitoring, supply chain assessment, and response plans. However, rushed GenAI adoption may lead to inadequate testing. Traditional security analysis likely misses unique LLM risks that require specialized AI safety expertise to identify – problems like [Sleeper Agents](#)²⁸ for which good solutions don't yet exist. Addressing these risks may require full-time LLM experts²⁹ (commanding [\\$10M+ salaries](#)) in every government IT department, plus solutions we don't currently have³⁰. While the government established the [Center for AI Standards and Innovation \(CAISI\)](#), their standards don't have much influence

²⁷ Or at least this is what Claude Sonnet 4 tells me. I can't find definitive public sources but this seems true.

²⁸ Notably, there need not be an aggressive team of equally skilled AI experts for risks to nonetheless be present (e.g. misaligned AI or models with singular loyalty).

²⁹ IT personnel may not be intimately familiar with LLM progress over the previous three years. **(a) The age of IT departments may be concerning** – [3.7% of federal IT employees may be under the age of 30](#), [the UK study showing a strong negative correlation between age and familiarity with generative AI in the public sector](#), [OpenAI usage statistics backing this younger-skewing audience](#), and the observation that many frontier LLM experts tend to skew very young (20s-30s). This is not to state that older employees cannot pick up skills in LLM security, only that this doesn't align with the current pool of talent. **(b) Current IT departments may be slow to react** – [a MITRE report from 2018](#), back when they saw the internet of things as the next monumental challenge, a number of recommendations were made that have yet to be taken seriously 7 years later. Of course, IoT was never as high of a national security concern as AI has become, and may receive comparatively better treatment. © **Talent Acquisition is filtered through an HR person's understanding of IT and cybersecurity** – meaning the ability to find and identify necessary AI talent may be limited.

³⁰ It would be wonderful to have a LLM approval task force in the government which works in collaboration with IT departments throughout the federal government.

on actual LLM adoption decisions. Ultimately, IT department approval doesn't appear necessary for AI diffusion.³¹

On-the-Ground Experience and Culture with LLMs in Government. Based on conversations with about four colleagues (limited sample), government workers across legislative and executive branches have access to and use frontier LLMs on government computers for writing, reading, editing, and research. While cultural taboos exist against automating important work, there's a positive shift toward productivity use. Unless directly observed using commercial chatbots for sensitive tasks or drafting final publications, oversight appears minimal (need to verify this further).

Additional indicators support widespread LLM access: many agencies operate internal chatbots (see above), [Microsoft 365 Copilot is likely standard on most government computers](#) running majority Windows and MS Office, and [Azure OpenAI has been available since summer 2024, receiving top secret authorization in January 2025](#).

Some reasons why government workers aren't incentivized to use LLMs: (A) There are not strong incentives to be productive in the US government. Promotions can often be a result of gaming a predetermined promotional interviews or speaking to the right people rather than by performance. (B) Most tasks are bottlenecked by procedure in ways that LLMs currently can't help with.

05 AI National Security Memorandum (NSM)

Intro. The October 2024 [AI National Security Memorandum \(NSM\)](#)³² is one of the most comprehensive articulations of US national security and policy towards AI. While this short post looked over the current state of LLMs in the US Government, the memorandum is a key indicator of where the future might go. I will be commenting on [this summary by CSIS](#) as the original is 40 pages.

NSM Goals. Some key goals outlined were to (a) maintain global leadership of advanced AI – including talent acquisition, expanding energy supplies and datacenters, and countering theft, espionage, and disruption. (b) Along the lines of CDAO's AIRCC mentioned above, a goal was set to accelerate the adoption of AI across federal agencies including the DoD and

³¹ One of my original goals in this investigation was to identify how many workers have access to LLMs for government work – the answer appears to implicitly be “all of them”, just to varying levels of usage and different use cases.

³² Or its long name: Memorandum on Advancing the United States' Leadership in Artificial Intelligence; Harnessing Artificial Intelligence to Fulfill National Security Objectives; and Fostering the Safety, Security, and Trustworthiness of Artificial Intelligence

Intelligence Community (IC) which includes the CIA and FBI among others. And (c) Develop governance frameworks to support national security, including international ones, to implement safety measures in so-far that it allows the comfortable rapid adoption of AI, and outlining roles and responsibilities for [US CAISI](#). The CSIS summary links to a companion document on this governance section, but all links are broken. My summary here almost certainly does not do this document justice.

Trump's Adherence to the NSM. While this document was established under the Biden administration, the Trump administration [seems to be following in the footsteps of the AI NSM](#). I'm personally inclined to say that the Trump administration, which is inclined to compete for US dominance and avoid red tape, will maintain the pace on all except governance.³³

³³ [Asking ChatGPT o3-Pro DeepResearch whether Trump is faithful to the Biden NSM](#) backs this up.

Appendix

A. GSA Schedules



GSA Schedules. For general software approval processes, the closest thing the US has to a centralizing decision making body is the **GSA Schedule** in which software is pre-vetted to a base standard and offered **government off-the-shelf (GOTS)** alongside physical products on their [GSA Advantage! Website](#). As mentioned at the start, these models have already received approval for use. [ChatGPT](#), [Claude](#), [Grok](#), and [Gemini](#) licenses already appear on GSA Advantage for purchase.³⁴ Typically the GSA Schedule is the first step to adoption elsewhere in the government, but it's only marginally influential. GSA schedules typically apply to paid products, not free ones, and mainly exist as a way of easily procuring software (skipping individual contracts, documenting purchases, etc.). GSA schedules are not a major consideration in the safe adoption of software.

B. State & Local Governments

While not the focus of this investigation, frontier models are being used in state and local governments. The following two examples are quoted from [OpenAI's post Introducing ChatGPT Gov](#).

³⁴ I don't know whether these listings existed previously. Here's my investigation: Looking at OpenAI specifically, there are listings for both ChatGPT Gov and consumer models. Carahsoft appears to be the authorized vendor that government purchases for ChatGPT go through (what benefit this proxy company adds – I am uncertain). Looking up the [contract number on the GSA eLibrary yields more info](#), but nothing on the start date. The [contract PDF](#) also has no reference of “openai” or “chatgpt” but does say the contract started in August of 2018 – originating from a Multiple Award Schedule (MAS) where broad category contracts are covered. The ChatGPT Gov license falls under MAS/54151ECOM which is “Electronic Commerce and Subscription Services.” The [WaybackMachine](#) has no record of relevant GSA Advantage searches, so I will assume that these listings were added close to the CDAO announcement date.

[State of Minnesota's Enterprise Translations Office is using ChatGPT Team](#) to deliver faster, more accurate translation services to the state's multilingual communities, significantly reducing costs and turnaround times.

[Commonwealth of Pennsylvania employees](#) participating in a first-in-the-nation AI pilot program found ChatGPT Enterprise helped reduce the time spent on routine tasks – such as analyzing project requirements and other elements of their work – by approximately 105 minutes per day on the days they used it.