11

PART II · THE INTELLIGENT STATE

The Legible Leviathan

Severin Field, Dave Banerjee



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.

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.

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.

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

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.

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.

ABOUT THE AUTHORS

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



From *Futurecraft: The Day After AGI*, Foundation for American Innovation, 2026.
Read or listen online at futurecraft.thefai.org/essays/the-legible-leviathan.

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

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.