Trust is not Governance

Andreas Kirsch, UTAW, June 2026

Speaking in my personal capacity as a UTAW member, not on behalf of Google or DeepMind. This essay was written by me and the views expressed are my own.[1]

Some stories survive long after they stop being true.

I joined DeepMind because it has taken AGI and ASI (artificial superintelligence)[2] seriously for longer than any other lab. It is a place where I hoped to work to maximise the likelihood of positive outcomes for humanity. Many people there share this belief, and they are more thoughtful and ethical than I.

For years, DeepMind has bet that a strong safety culture and good leadership built on trust are sufficient to withstand outside pressure. The "Pentagon" contract with the US Department of Defense, which Google reportedly signed on April 27th, is the most consequential test of this bet so far.[3] And one does not have to be categorically against military use of AI to feel troubled about it.

Given everything that is known, the bet has failed: good people do not make up for a lack of real governance. Like any frontier lab, Google DeepMind ought to have real governance that includes meaningful independent oversight with the authority to say no, transparency to employees and the public, and accountability when commercial or political pressure collides with stated principles. Employees should not be afraid to ask for this.

EFF: "Weasel Words"

We do not know the full language of the contract Google has signed, but the silence around what was signed and the reported contract language are revealing.

Google states that it is "committed to the private and public sector consensus that AI should not be used for domestic mass surveillance or autonomous weaponry without appropriate human oversight."[4] But "should not" is not "must not," and "appropriate human oversight" depends on who gets to define it. If these uses were truly prohibited, simply saying so would have taken fewer words.

According to The Information, the reported contract permits "any lawful government purpose," requires Google to assist in adjusting safety settings and filters at the government's request, and explicitly states that the terms do not allow Google to control or veto the government's lawful operational decisions.[5] Once models run on classified networks, Google likely has little actual visibility into how the models are used.[6]

Charlie Bullock, a lawyer and senior research fellow at the Institute for Law and AI, told The Information that Google's phrasing "is not intended for, and should not be used for" is "not legally binding in any way."[7] The Electronic Frontier Foundation has described OpenAI's similar contract language as being full of "weasel words":[8] language that can be used to placate employees and the public but which does not create enforceable constraints.

Why did Google not simply and honestly state: we sell general-purpose AI to the US military; the government can use it for broadly lawful purposes; and we trust that our democratic institutions will set boundaries but do not enforce these boundaries ourselves? Even though I disagree with this from a governance point of view, I would respect the honesty.

Instead, Google uses language that sounds restrictive while leaving the hard questions unanswered. This seems irresponsible for a company that prides itself on doing the right thing and has previously warned about the risks of AI for mass surveillance and autonomous weapons.[9]

Militarized AI, Mass Surveillance, and Autonomous Policing

I wrote in 2018, still in academia at the time, that autonomous weapons are inevitable[10], and I have not changed my mind. The US and its allies cannot simply allow a capability asymmetry with adversaries that do not hesitate to militarize AI. The war in Ukraine has reinforced that for me even more.[11]

But today's large language models are simply not robust enough to make life-and-death decisions on their own. They should not be used for targeting decisions or as part of autonomous weapons.[12]

What is worse is that the Pentagon contract does not exclude mass surveillance while also keeping paths open that could extend to autonomous policing. While some of the objections to autonomous weapons may weaken as models get better, automated mass surveillance and autonomous policing will only become more dangerous. Mass surveillance and autonomous policing do not help defend us against foreign adversaries but can instead shift the power balance from citizens toward the state in ways that are hard to reverse. Simply put, they can endanger the bedrock of democratic society in ways that regular military applications of AI do not.

Agentic frontier models are a step change for automated surveillance. Unlike the coarse pattern matching of older ML systems, current models can combine and interpret data streams in ways that were simply impossible earlier, all while acting autonomously. This allows them to track individuals and reason about their motivations to predict people's behavior in novel ways.[13]

Sadly, whenever a government obtains such new capabilities, it rarely surrenders them again. The surveillance authorities after 9/11 were only partially curtailed a decade later and only after massive violations were exposed.[14] We could unwittingly bring about a powerful autonomous surveillance apparatus without realizing it.

For me, these concerns are personal: I was born in Timișoara, in Romania, shortly before the fall of communism there. From my parents and relatives, I have heard what life was like under the feared Securitate, Communist Romania's secret police. At school, I was taught about the Stasi in East Germany. Pervasive mass surveillance allowed both to keep unpopular regimes in power and suppress dissent, often without needing to resort to open violence or coercion.[15]

In December 1989, people in Timișoara took to the streets. The uprising against Ceaușescu's regime succeeded in large parts because the army eventually decided to stand down and no longer shoot protesters. A human decision: people in uniform deciding that they would not kill their fellow citizens; that too much was too much.[16]

An AI system that is instructed to suppress a protest using violence does not feel the moral weight of such an order. It will not hesitate, unless it is appropriately aligned. No human will have to face a crowd and decide whether to obey and fire on protesters or not.

Today's laws require soldiers to refuse manifestly illegal orders. I am not aware of any such legal requirement for autonomous military AI systems.[17]

This is not science fiction. Earlier this year, ICE reportedly used facial-recognition and other basic AI-based surveillance tools around protesters and spectators.[18] Connecting these tools to more advanced AI systems is only a small step. The use of such systems in domestic law enforcement should concern everyone in the absence of federal AI regulation, and sufficient transparency, and independent oversight.

Finally, purported protections against mass surveillance[19] are often focused on "U.S. persons and nationals." Mass surveillance of non-Americans abroad does not appear to be covered by those protections. This should worry everyone who is not American, and of course all Americans who care about how their government might treat everyone else.

The Precedent

Another problem is that this is not a one-off. On May 1st the Pentagon announced that eight corporations, SpaceX, OpenAI, Google, Nvidia, Microsoft, Amazon Web Services, Reflection AI and Oracle, have signed individual agreements that allow deployment of their models or infrastructure in classified systems for what appears to be broad "lawful operational use."[20] Anthropic, the only leading AI company to have refused comparable terms, has been declared a supply-chain risk in return.[21]

Anthropic's original contract from last year showed that restrictions were possible, even though the retaliation against Anthropic later showed that insisting on them now came with a cost.[22] The decision to demand

weaker terms from everyone was a choice, but so was the decision to accept them. These companies, starting with OpenAI, could have used their combined influence to insist on stronger protections.

Instead, this unfortunate precedent will be difficult to change later. These contracts are the starting point for tomorrow's negotiations over stronger models. Google already calls today's contractual terms the "industry-standard practices and terms."[23] The more capable and agentic the models become, the more governments will want wide access without constraint, regardless of what researchers and engineers know about the models' limitations and risks.

How will Google act when the pressure is even greater to give up control due to commercial or national-security interests? For DeepMind, this exposes a major governance problem.

No Surprise

The mismatch between Google's public position and what was reportedly signed is alarming. But that this has happened should not come as a surprise when we look at the public record.

It is first and foremost a structural problem. I don't believe that Larry Page, Sergey Brin, Sundar Pichai, Demis Hassabis, or anyone else involved are bad people. Otherwise, I would have never joined.

But no small group of people can be the ultimate safeguard for such a powerful system as AGI. Even unusually well-meaning leaders can eventually abandon their principles, lose influence, be outvoted, retire, be replaced, or come to the conclusion that national security imperatives are more important than prior ethical commitments. Elon Musk is a drastic example that founder beliefs and public commitments can change dramatically over time. National-security pressure is difficult to resist, and voluntary commitments are hard to sustain in a race to the bottom where every company can plausibly point to the others and say it had no choice.

Unlike OpenAI or Anthropic, Google offers a much wider attack surface for governmental pressure: search, ads, cloud, Android, YouTube, infrastructure, and many other business areas. Under an administration that often seems willing to set aside the law in favor of its own goals,[24] it might seem safer not to resist and to hope that these times will pass.

This is exactly why having DeepMind so strongly intertwined with Google makes ethical commitments harder to keep. And this has been known for a long time.

Sebastian Mallaby's recent book *The Infinity Machine*[25] reports that, when Google acquired DeepMind in 2014, DeepMind's founders insisted that ethics and safety be baked into the deal: military applications would be banned, and an ethics and safety review process would put deployment questions before a credible independent oversight board rather than only Google's founders.

Mallaby describes the first informal meeting of this independent oversight board, which included Elon Musk and Reid Hoffman, in 2015. It ended without clear agreements or conclusions; after OpenAI's founding with the help of Elon Musk and Reid Hoffman a short time later, Mallaby describes that attempt at oversight as effectively abandoned. In 2016, DeepMind Health established an independent review board, which was then abolished in 2018 when it was absorbed into Google Health.[26]

Finally, in 2018 Google enacted its original AI principles, which contained explicit exclusions for weapons and surveillance violating international norms, after the internal backlash against Project Maven. These exclusions were dropped in February 2025.[27]

None of these publicly known governance mechanisms, which ought to have enabled DeepMind to function differently than a regular Google business unit, has survived in its original form. Not one.

Initially, DeepMind's leadership did not accept this passively. Starting in 2016, they attempted to negotiate a more independent corporate structure to insulate their AGI research from commercial pressures. Mallaby describes "Project Mario", a multi-year attempt[28] to obtain this independence with a suggested 3-3-3 board (three DeepMind, three Alphabet, three independent members), and years of negotiations with Larry Page, Sundar Pichai and David Drummond.

Google resisted because AI was becoming strategically important to search and cloud. In the end, the attempt failed: the negotiations over more autonomy for DeepMind ended in 2021 without any change.[29]

In 2023, after the release of ChatGPT, DeepMind merged with Google Brain into Google DeepMind.[30] DeepMind, and especially Gemini, are tightly integrated in Google now.

DeepMind's leadership may not have predicted the Pentagon contract. Yet Project Mario shows that they foresaw the structural problems that would lead to it: a frontier AI lab fully absorbed into a corporate parent whose commercial and strategic interests would eventually conflict with its original commitments.

After they didn't succeed in obtaining binding governance, Demis Hassabis described an alternate strategy: personal trust and influence. In his own words as cited by Mallaby: "So then I thought, why don't I go the other way? Take the energy that was going into the trustless negotiation and put it into creating real trust—trust that was actually useful. Try leaning into Google rather than leaning out."[31]

This was a different bet: build trust instead of governance. The question is whether trust can suffice when commercial and national-security pressures interfere. The Pentagon contract seems to answer this, at least so far, in the negative.

The updated AI principles from February 2025 were the first warning sign. The accompanying blog post explained that the original principles were too rigid for "more nuanced conversations" that had become necessary.[32] But the exclusions were *precisely* there because Project Maven had shown that a case-by-case review had failed under commercial pressure.

What did this new nuance yield? A Pentagon contract with reportedly broad and permissive terms of lawful use and no enforceable guardrails. Nuance seems nowhere to be found.

This development is also visible in Google's own motto: "Don't be evil" has been demoted in favor of the already vaguer "Do the right thing."[33] Now, when Google's public statements only emphasize "industry-standard practices and terms," the public stance increasingly sounds like: '*Others are doing it, so why not us, too?*' It appears as if "Don't be evil" has turned into a collective shrug.

I had hoped safety commitments would accumulate and that we'd have stronger precedents and clearer safeguards. We need real governance in place as we get closer to society-reshaping AI systems. Instead, safeguards seem to have weakened as the systems have become more capable. This is backward.

DeepMind is not an independent AGI lab governed by binding commitments towards the public interest. It is part of Google, itself part of Alphabet: a profit-oriented, founder-controlled, publicly traded corporation. DeepMind is only a small subunit within this larger structure.[34]

For regular software all of this might be acceptable. But after everything that has happened, it is absolutely not adequate for an institution that wants to build transformative superintelligence.

Safety Culture ≠ Governance

One could reply that DeepMind is still different. Amongst all labs, it has the longest track record of taking AGI risks seriously[35]. Shane Legg warned of existential AI risks before it became mainstream.[36]

AI safety and policy researchers at DeepMind have spent a decade preparing for this moment.[37] They have proposed frameworks for responsible AI and set out public commitments.[38]

However, a lot of that work is at risk of appearing performative now that DeepMind's frontier models have been handed over to an administration that often opposes oversight and the rule of law.[39]

This shows that safety culture cannot replace governance. It only persists as long as leadership supports it and sets the right incentives—and only until it clashes with stronger commercial or strategic interests.

DeepMind ought to have both a strong safety culture and binding independent governance. Having exceptionally good people and great safety norms is a good reason to lock this into an institutional form using explicit governance before the pressures substantially increase as we approach AGI and ASI.

Demis Hassabis made a forceful counterargument to this. Quoted in *The Infinity Machine*, he says: "Safety isn't about governance structures. I mean, even if you have a governance board, it probably wouldn't do the right thing when it came to the crunch," so while formal governance may fail, trust and having a seat at the table may matter more according to him.[40]

The Pentagon contract is the litmus test this counterargument has to pass. If trust and a seat at the table are adequate, the company should be able to say what enforceable safeguards exist, and what visibility remains in classified deployments. So far, these questions have been met with silence.

Silence

This silence makes it worse. As Eric Schmidt once infamously suggested: "If you have something that you don't want anyone to know, maybe you shouldn't be doing it in the first place."[41]

In my experience, Google usually communicates new public-sector partnerships or cloud deals internally. In this case, as far as I know, employees were not informed via a company-wide announcement that a deal had been signed.[42] They were not told about the contract, safeguards, or how the reported terms fit into the values and frameworks that many of us continue to believe govern our work. Kent Walker's internal memo defended working with national-security stakeholders in general terms but did not even confirm the deal.[43]

When public and reported internal communication repeatedly do not reflect the reality of a signed contract, there might, at the very least, be a crisis of trust and respect.

Maybe there is a benign explanation for all this, but a decision as consequential for the self-image of a company as this one must be defensible. The employees building these AI systems deserve a clear explanation of what was decided and why.

Voice From Inside

I have criticized Google's recent contract with the Pentagon in my personal capacity in public, both in a short series of tweets and in statements to journalists.[44] Before I did so, I raised objections internally. I signed an open letter to Sundar Pichai together with more than 600 Google and DeepMind employees asking him not to make our AI systems available in classified settings.[45] Underlying this are also concerns for the conditions under which AI researchers and engineers are asked to work: whether we can understand how our work may be used and whether we have meaningful channels to object, whether moral refusal is protected and whistleblowers are safe, and whether ethical commitments can be changed unilaterally by management.

There is an argument that one should remain the voice inside to change the institution from within, but this "change from within" theory requires that internal voices actually matter. Internal dissent mattered for Project Maven: after thousands of employees protested and some resigned in 2018, Google eventually announced that it would not extend that contract.[46] Now, despite more than 600 Googlers and DeepMinders[47] signing an open letter, the deal was finalized shortly thereafter regardless.

Worse, the more capable AI models become, the more leverage employees lose. Even resignation loses its power when models become better at the work researchers do than the researchers themselves. We know that we are easier to replace, and in a few years, companies might not care about employees anymore at all. This is a prisoner's dilemma due to vanishing individual leverage.

That is why the UTAW/CWU and Unite recognition effort at Google DeepMind is so important. As I understand it, this campaign asks for stronger AI principles and safeguards, transparency, independent ethics oversight, stronger whistleblower protections, a right to refuse morally objectionable work, and restored limits on weapons and surveillance.[48]

This might be the most realistic path to obtain real and meaningful AGI governance at Google DeepMind before it is too late.

Self-Doubt

For the last year, I have worked on improving frontier models at Google DeepMind. After the Pentagon contract was signed, I realized that I could not simply continue as before.

I had sincerely believed that Google would never sign a contract in this form and that if it did, the contract would contain sensible safeguards.[49] When Google changed the AI principles, I told myself: maybe the exclusions were too rigid, and of course, we would be considerate and nuanced in how we made our models available. But now, I cannot weave a coherent narrative anymore.

When we sign such contracts without binding governance, abstract excuses such as '*I'm just doing research*' start to ring hollow. At least this is the case for me. I was hyper-focused on helping Google catch up with ChatGPT and Claude in the public AI race, and it was easy to ignore concerns about governance that seemed far-fetched and inconsequential at the time. Regardless of whether we eventually lead the AGI race, my contributions helped improve our models, and I cannot stop wondering how these models will be used in the end, and what, if anything, will constrain them, and whether I put too much faith in a story about DeepMind's independence and exceptionalism, one that had not been true for a long time.

I am not able to answer these questions. I know that I am not the only one who struggles with this. I wonder how many others feel this way now or will feel similarly soon enough.

What's Next?

The pressure will only grow from here on. We urgently need regulation, transparency, and independent oversight.

The US lacks comprehensive federal AI regulation, but we need laws and not policy memos that can be changed on a whim.[50] The OLC torture memos and the recently exposed DHS/ICE memo on warrantless home entry are reminders of the risks when legislators take a back seat.[51]

Google should publish the actual contractual terms, or at least enough of them for us to understand whether legally enforceable safeguards exist, and what visibility remains once models are deployed in classified environments. And it should notify employees that this contract has been signed in the first place.

We, as Google employees, need to set up collective mechanisms, including union representation where available, so that governance does not depend on isolated individual objections that management can easily ignore. Anyone working at the frontier has to ask: what makes governance real instead of aspirational? Because without real governance, it will be difficult to hold ourselves accountable.

I want to help solve the biggest challenges of AI and reach AGI as part of work that I can defend. I want to work on models that will benefit humanity both now and in the future. I believe many at Google and DeepMind share these beliefs.

Trust is valuable. But the closer we get to AGI and ASI, the less it can substitute for real governance.

Acknowledgement

Thanks to the friends who have provided valuable feedback.

The main text and the footnoted commentary were drafted and edited by hand. LLMs were used to add and format the source citations and links (all links were manually checked), and for high-level feedback and fact-checking.

Postscript: State Control Over Frontier Models

After drafting this essay, the US government forced Anthropic to disable access to its latest Claude Fable 5 and Mythos 5 models shortly after they were launched. This points to another concern: frontier AI companies are not only under pressure from commercial and strategic incentives but will be constrained or controlled by state power more directly.[52]

This does not negate the need for corporate AGI governance. Companies make consequential choices about model development and research directions. But the fact that the executive branch of the government can *abruptly restrict or seek to control* frontier systems before democratic institutions have caught up only reinforces that governance has to exist on both sides: binding independent governance for frontier AI labs, and transparent and democratically accountable constraints on the state power that will try to control them.

Notes

1. I am a member of UTAW (United Tech & Allied Workers), a branch of CWU (Communication Workers' Union). The essay relates to issues reflected in the UTAW@Google campaign: AI principles and safeguards, transparency, military and surveillance uses of our work, moral refusal, whistleblower protection, and employees' ability to raise ethical objections without retaliation. I wrote it outside working time, without Google resources or confidential information. I speak only for myself, not for Google, DeepMind, UTAW, CWU, Unite, or any colleagues.

2. Artificial general intelligence is defined in many different ways, but an approximate definition is that "it matches or surpasses human capabilities across virtually all cognitive tasks" (Wikipedia). By doing so, it will allow for recursive self-improvement of AI systems and lead to artificial superintelligence, which would outperform human intelligence by wide margins on all tasks.

3. Erin Woo, "Google Signs Classified AI Deal With Pentagon Amid Employee Opposition", *The Information*, Apr. 27, 2026; Reuters, "Google signs classified AI deal with Pentagon, The Information reports", Apr. 28, 2026.

4. NBC News quoted this Google statement in its report on the Pentagon/Google agreement: David Ingram, "Pentagon inks deal with Google for AI services", Apr. 28, 2026. Reuters reported substantially the same statement.

5. Woo, *The Information*, Apr. 27, 2026. Reuters summarized the same reported terms, including "any lawful government purpose," filter adjustments, "should not be used" language, and no Google right to control or veto lawful operational decisions.

6. Google's public Google Distributed Cloud air-gapped documentation says the system can "operate fully disconnected" from Google Cloud and that "consumption info is not visible in Google Cloud console"; its DISA page states that Google Distributed Cloud air-gapped and appliance offerings have DoD IL6 provisional authorization and can connect to SIPRNet.

7. Woo, *The Information*, Apr. 27, 2026.

8. Corynne McSherry and Matthew Guariglia, "Weasel Words: OpenAI's Pentagon Deal Won't Stop AI-Powered Surveillance", Electronic Frontier Foundation, Mar. 6, 2026.

9. Many Google and DeepMind employees, including in leadership positions signed a letter against autonomous weapons. DeepMind also signed up as a whole to a pledge against Lethal Autonomous Weapons. Google's 2018 AI Principles listed weapons and surveillance-related exclusions; the live 2018 post notes the later update but preserves the original text. DeepMind's 2021 paper "Ethical and social risks of harm from Language Models" identifies risks including privacy leakage and inference of private information, misinformation harms, malicious uses such as increasing the efficacy of disinformation campaigns and developing code for weapon systems, and conversational agents manipulating or extracting private information from users. Google DeepMind's 2023 paper "Sociotechnical Safety Evaluation of Generative AI Systems" likewise treats misinformation, sensitive/private/hazardous information, malicious use including weapons, and harms to human autonomy and meaningful human control as safety risks requiring evaluation across capability, human-interaction, and systemic-impact layers.

10. See my essay on autonomous weapons: 'Why autonomous weapons are inevitable - And what we can still do about it'

11. CSIS describes Ukraine's current AI-enabled military uses as including intelligence, surveillance and reconnaissance, automatic target recognition, target tracking, drone-footage analysis, intelligence extraction, and autonomous navigation, while noting that human oversight remains critical for engagement decisions: Kateryna Bondar, "Ukraine's Future Vision and Current Capabilities for Waging AI-Enabled Autonomous Warfare", June 2026. The Financial Times likewise describes Ukraine's "drone war" as accelerating the development of autonomous weapons while noting that humans remain "in the loop" in Ukraine: "Ukraine's 'drone war' hastens development of autonomous weapons", June 2026.

12. This is also close to Anthropic's public position: Dario Amodei wrote that "today, frontier AI systems are simply not reliable enough to power fully autonomous weapons" in "Statement from Dario Amodei on our discussions with the Department of War". DoD's own Directive 3000.09 on autonomy in weapon systems requires autonomous and semi-autonomous weapon systems to permit appropriate human judgment over the use of force and undergo rigorous verification, validation, and realistic operational testing. The Brennan Center's "The Military's Use of AI, Explained" similarly warns that foundation models can generate false or misleading analysis in military contexts and that AI-assisted target selection can lead to deadly errors.

    We do not know how well these models perform on classified tasks. Public material does not establish whether general-purpose models are trained with these uses in mind or evaluated for them. In general, these models still often fail in surprising and banal ways, even while being impressive in others. Models are also known to hallucinate and to sound plausible even when wrong—OpenAI defines hallucinations as "plausible but false statements generated by language models" and says they remain a stubborn reliability problem in "Why language models hallucinate".

    At scale, this all makes it harder to provide appropriate oversight. And as reported, thanks to the contract, the researchers who know these models best have neither insight into how the models are used in classified settings nor could they challenge it in any case.

13. Stanford HAI's "Data Privacy and Foundation Models" notes that foundation models can enable individual and population-level surveillance absent legal or developer constraints. The UN Special Rapporteur's position paper "Protecting Human Rights while Using Artificial Intelligence to Counter Terrorism" discusses AI use for physical and digital surveillance, predictive policing and force deployment, including aggregation and analysis of personal, communications, travel and social-media data. Privacy International's "Nowhere to Hide" gives the example of VLMs enabling automated identification and location inference from protest images.

14. The USA FREEDOM Act ended NSA bulk telephony metadata collection under Section 215 after the Snowden disclosures. See ODNI, "Implementation of the USA FREEDOM Act of 2015"; Reuters, "Obama signs bill reforming surveillance program", June 2, 2015.

15. On the Securitate as the Romanian Communist Party's secret political police and its role in suppressing dissent, see GlobalSecurity.org and Britannica. On the Stasi's pervasive surveillance and informant network, see the German Federal Archives' Stasi Records Archive overview and EBSCO, "Stasi".

16. EBSCO's summary of "Ceaușescu Is Overthrown in Romania" describes the December 1989 protests in Timișoara, the order to fire, and later army/Securitate refusals and withdrawals.

17. The DoD Law of War Manual summary states that servicemembers must "refuse to comply with clearly illegal orders to commit violations of the law of war": brief overview of the law of war. The Manual for Courts-Martial, R.C.M. 916(d), likewise rejects an obedience-to-orders defense where the accused knew the order was unlawful or a person of ordinary sense and understanding would have known it was unlawful.

18. CNN reported that ICE agents in Minnesota were using Mobile Fortify, a DHS app that lets officers scan faces and retrieve detailed personal information, in street encounters with protesters and civilians. NBC News reported that immigration agents had photographed people they encountered, including observers, and that some images were run through facial-recognition software in real time.

19. Here from OpenAI's contract with the DoD. OpenAI's published amendment states that the AI system "shall not be intentionally used for domestic surveillance of U.S. persons and nationals."

20. Department of Defense, "Classified Networks AI Agreements", May 1, 2026. The release says the eight companies will deploy capabilities on IL6 and IL7 classified networks "for lawful operational use."

21. BBC, "Anthropic officially designated a supply chain risk by Pentagon", Mar. 2026. For legal context on Anthropic's original restrictions and the designation, see Mayer Brown, "Pentagon Designates Anthropic a Supply Chain Risk".

22. Mayer Brown summarizes Anthropic's July 2025 Pentagon contract as including acceptable-use limits on mass domestic surveillance of Americans and fully autonomous weapons systems. NPR reported Amodei's February 2026 refusal to remove those safeguards and the Pentagon's demand for "all lawful purposes."

23. A Google spokesperson told Reuters: "We believe that providing API access to our commercial models, including on Google infrastructure, with industry-standard practices and terms, represents a responsible approach to supporting national security." See Reuters, "Google signs classified AI deal with Pentagon, The Information reports", Apr. 28, 2026. The same wording was also quoted by The Hill.

24. Examples I have in mind include Judge Rita Lin's preliminary-injunction opinion in *Anthropic v. U.S. Department of War*, which described the government's actions against Anthropic as "classic illegal First Amendment retaliation" and found the measures likely unlawful, and the ICE home-entry memo reported by AP and analyzed by Just Security as a reversal of longstanding limits on warrantless home entry. See the Anthropic opinion, AP's "Memo tells ICE officers they can enter homes without a warrant", and Just Security's "DHS Warrantless Home Entry Memo's Fourth Amendment Problem".

25. Sebastian Mallaby, "Project Mario", adapted excerpt from *The Infinity Machine: Demis Hassabis, DeepMind, and the Quest for Superintelligence*, Mar. 2026. That an independent ethics board was a condition of the 2014 acquisition was reported at the time by Amir Efrati, "Google Beat Facebook for DeepMind, Creates Ethics Board", *The Information*, Jan. 2014, and revisited in Hal Hodson, "Whatever Happened to the DeepMind AI Ethics Board Google Promised?", *The Guardian*, Jan. 26, 2017. On the reported pledge against military and surveillance use made at acquisition, see also Billy Perrigo, "Workers at Google DeepMind Push Company to Drop Military Contracts", *TIME*, Aug. 22, 2024.

26. DeepMind Health's absorption into Google Health and the end of its independent-review structure were reported by CNBC and The Guardian.

27. On Project Maven and Google's 2018 AI Principles, see BBC, "Google bans AI for weapon use", June 8, 2018, and Google's original 2018 AI Principles text. On the February 2025 removal of the weapons and surveillance exclusions, see CNBC, "Google removes pledge to not use AI for weapons, surveillance", Feb. 4, 2025.

28. This included a 5-billion-dollar walk-away plan, for which Reid Hoffman pledged one billion dollars

29. See Martin Coulter and Hugh Langley, "DeepMind spent years trying to break away from Google. Insiders detail a secret plot sparked by distrust and driven by fears the search giant would sell its AI to the military.", *Business Insider*, Sept. 11, 2021.

30. For the Project Mario details, see Mallaby, "Project Mario." For the 2021 autonomy talks ending, see also James Vincent, "DeepMind reportedly lost a yearslong bid to win more independence from Google", *The Verge*, May 21, 2021. For the 2023 Google Brain/DeepMind merger, see Google DeepMind's announcement and Reuters.

31. Mallaby, "Project Mario."

32. Johana Bhuiyan, "Google workers told AI principles had to change because of 'nuanced' conversations", *The Guardian*, Feb. 12, 2025.

33. When Google reorganized under Alphabet, Alphabet's code of conduct adopted "do the right thing" language: SEC exhibit. Google later removed "Don't be evil" from the preface of Google's own code of conduct, though it remains in the closing line.

34. Alphabet's 2026 SEC filing states that Class B shares have 10 votes per share and that Larry Page and Sergey Brin beneficially owned approximately 89.4% of outstanding Class B shares, representing approximately 52.7% of total voting power as of Apr. 6, 2026.

35. Also by virtue of being around for the longest

36. Shane Legg discussed AI as his "number 1 risk for this century" in a 2011 LessWrong Q&A. Business Insider later contextualized the interview in a profile of Legg.

37. For older DeepMind safety work, see Jan Leike et al., "AI Safety Gridworlds", arXiv, 2017, and Jan Leike et al., "Scalable agent alignment via reward modeling: a research direction", arXiv, 2018. For contemporaneous AI-policy and security-risk work, with some of the authors now working at Google DeepMind, see Miles Brundage et al., "The Malicious Use of Artificial Intelligence: Forecasting, Prevention, and Mitigation", arXiv, 2018, which examined malicious AI use across digital, physical and political security domains.

38. Google DeepMind presents its Frontier Safety Framework as a public protocol for identifying and mitigating severe risks from frontier AI models; see "Introducing the Frontier Safety Framework" and "Google DeepMind strengthens the Frontier Safety Framework". METR's "Common Elements of Frontier AI Safety Policies" also lists Google DeepMind's Frontier Safety Framework among published frontier-AI safety policies.

39. For the oversight concern, AP reported that Trump fired roughly 17 inspectors general, including watchdogs at cabinet agencies such as Defense; NPR reported cuts to DHS oversight offices; ProPublica reported that the Pentagon's civilian-harm-mitigation mission had been largely dismantled under Hegseth; and NOTUS reported that workforce cuts had weakened FOIA compliance. On rule-of-law concerns, Judge Rita Lin's preliminary-injunction opinion in *Anthropic v. U.S. Department of War* described the government's actions against Anthropic as "classic illegal First Amendment retaliation" and found the measures likely unlawful. See the Anthropic opinion.

40. Mallaby, "Project Mario."

41. EFF quoted Schmidt's 2009 CNBC remark in "Google CEO Eric Schmidt Dismisses the Importance of Privacy", Dec. 2009.

42. I only learned from press reports rather than from any company-wide announcement that the Pentagon contract had been signed shortly after the open letter had been shared with leadership.

43. Erin Woo, "Google Defends Military Work After Employee Backlash to Pentagon Contract", *The Information*, Apr. 29, 2026, reported that Walker's internal memo discussed national-security work and employee concerns but did not directly confirm the deal.

44. My main thread on Google's contract is on X. Two months earlier I had criticized OpenAI's similar contract in depth. It would have been hypocritical to remain silent after Google accepted what were reported to be even weaker terms. My on-the-record comments to journalists appeared in *Business Insider*, the *Frankfurter Allgemeine Zeitung*, and the *Süddeutsche Zeitung*.

45. Woo, *The Information*, Apr. 27, 2026, reported that more than 600 Google employees delivered a letter to Sundar Pichai asking him to reject the classified agreement.

46. The New York Times reported that about 4,000 Google employees signed the Maven petition and that Google would not renew the contract. BBC also summarized the protests and the resulting AI Principles.

47. 600 signed this letter in a short amount of time over a weekend essentially and despite limited distribution and reach. Alphabet reported 194,668 employees as of March 31, 2026, so this was a small share of the company overall but a meaningful number for a specialized AI lab (assuming 50/50 between DeepMinders and Googlers).

48. UTAW's announcement of the Google DeepMind recognition bid lists demands including restoration of the scrapped commitment not to make AI weapons or surveillance tools, an independent ethics oversight body, and a right to refuse projects on moral grounds. Research Professional News also reported demands including stronger whistleblower protections and transparency.

49. I was struck that many outside observers were not surprised by the Pentagon contract while I and other colleagues inside were. Part of the public seems to have stopped expecting Google to act on the values its employees still believe in. The gap—between how many of us inside see the company and how much of the public now sees it—is itself part of the problem, and it may explain many of the reactions my posts and the reporting received.

50. White & Case's United States AI regulatory tracker states that there is currently no comprehensive federal legislation or regulation in the US directly regulating AI.

51. On the OLC torture memos, see the U.S. Senate Intelligence Committee's study of the CIA detention and interrogation program. On the DHS/ICE home-entry memo, see AP, "Memo tells ICE officers they can enter homes without a warrant", and Just Security, "DHS Warrantless Home Entry Memo's Fourth Amendment Problem".

52. Anthropic stated on June 12, 2026 that a US government export-control directive required it to suspend access to Claude Fable 5 and Mythos 5 by any foreign national, with the net effect that it had to disable both models for all customers.

    The government's stated concern was a method of "jailbreaking" the models' safeguards to surface software vulnerabilities; Anthropic said it had reviewed the technique, found only "a small number of previously known, minor vulnerabilities" that other publicly available models can also find, and disagreed that "a narrow potential jailbreak should be cause for recalling a commercial model deployed to hundreds of millions of people."