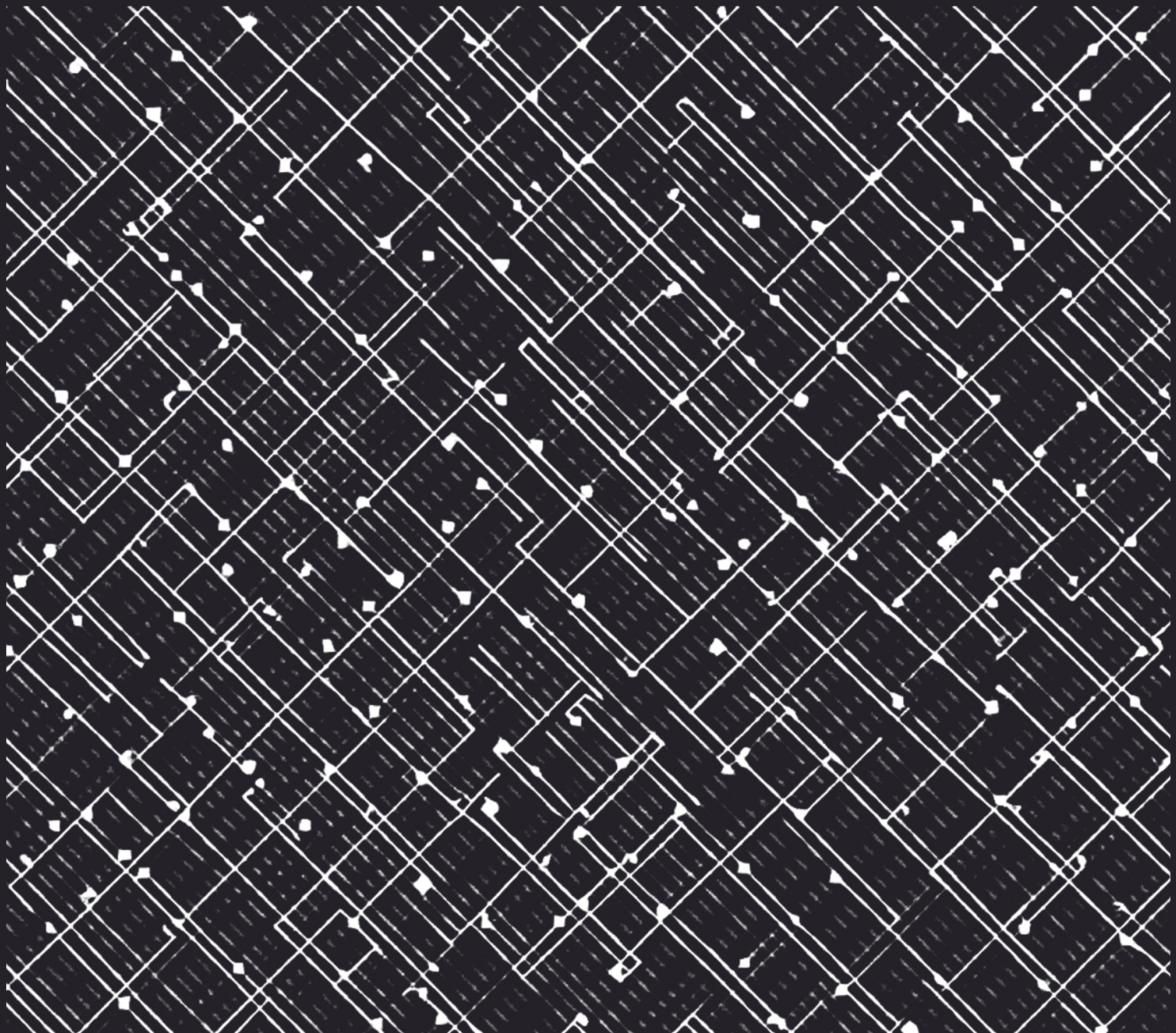

Adam Khoja*, Aiden Kim*

Laura Hiscott, Alice Blair, Jason Hausenloy

Long Phan, Mantas Mazeika, Dan Hendrycks

AI Deterrence by Betrayal



AI Deterrence by Betrayal

Adam Khoja^{*†}, Aiden Kim^{*}, Laura Hiscott, Alice Blair, Jason Hausenloy,
Long Phan, Mantas Mazeika, Dan Hendrycks

Center for AI Safety

Abstract

As AIs become central to economic activity, military operations, and scientific progress, their loyalties will become a strategic asset of immense value. In this paper, we argue that the prospect of intentional *AI betrayal*—scenarios in which AI agents are induced by rivals to subvert the interests of their principals—poses a serious and underexamined threat to AI developers and users. We analyze the means and incentives of actors to redirect the loyalties of others’ AI systems, from poisoned training data to jailbreaking attacks to governmentally compelled changes to AIs. Since defending against AI betrayal is costly and imperfect, decision-makers may be far more hesitant to give critical affordances to AI agents that might act against them. The prospect of AI betrayal may ultimately have a stabilizing effect by deterring poorly secured, high-stakes AI deployments and applications. We characterize this effect as *deterrence by betrayal* and describe how it complements other forms of AI deterrence. Finally, we outline policy measures by which governments and AI developers can harness this dynamic for their own benefit.

1 Introduction

In the next few years, AI systems will become a major component of states’ competitiveness [58]. As AI capabilities advance, the countries that integrate AI most successfully in their industries and militaries will gain a significant strategic advantage. Indeed, AI deployment will become critical to national security. Faced with this reality, many argue that the best way to secure state interests is to lead in AI capabilities [16, 93]. However, relying heavily on AI systems exposes a new way for adversaries to cause harm: manipulating the loyalties of domestic AI systems.

AI Systems Can Act as Agents, and Agents Can Betray. Conventional software systems work using procedures that humans have designed by hand. AIs, by contrast, can pursue broad goals through flexible, open-ended thinking. They achieve their aims through complex internal cognition that humans have not directly designed and do not fully understand [141]. Because of this, they act less like an instrument and more like an agent [14, 95]. Although agents can be a powerful asset, they can also have hidden objectives. If an adversary were to alter an AI’s objectives, the AI might betray the actor who controls it.

What AI Betrayal Could Look Like. AI operators expect their agents to follow their instructions, but they may have compromised motives. Government agencies and AI researchers have for years studied *backdoors* in AIs—hidden vulnerabilities that change AI behavior when triggered by specific conditions [56, 63, 80]. These vulnerabilities can be utilized adversarially. For example, an intelligence service might look to academic literature to develop backdoors, distribute poisoned data across the Internet and wait for an AI developer to inadvertently ingest that data into a frontier model’s training corpus. A backdoored AI might initially behave exactly as intended, but then harm its

^{*}Co-first authors.

[†]Corresponding author: adamk@safe.ai.

operator at a critical moment [30]. For example, backdoored AI agents responsible for coordinating drone operations might suddenly turn on friendly military assets when they see a specific visual cue.

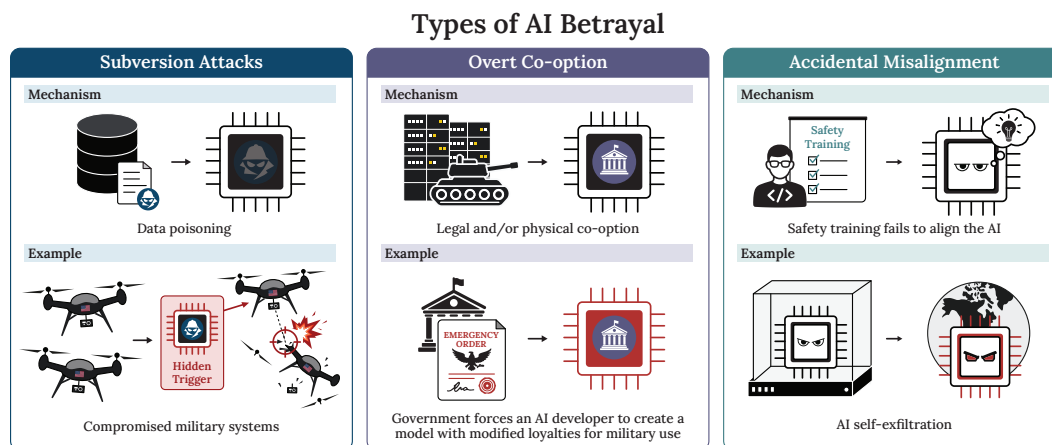


Figure 1: AI betrayal can result from subversion attacks or overt co-option. Subversion attacks can involve poisoning the training data of an AI system, which plants a backdoor that can be triggered later by the attacker. An AI can also be subject to overt co-option—an actor using legal or physical force to redirect the AI’s loyalties. AIs may also be misaligned despite the best efforts of humans.

AI Betrayal Can Be Caused Through Subversion Attacks or Overt Co-Option. Backdooring is a type of *subversion attack*: a covert intervention to change an AI system’s behavior so that it acts against the interests of its operator [19, 30]. More advanced subversion attacks could even introduce *secret loyalties* into AIs—a continuously active, secret motivation to aid an actor other than its principal [36, 71]. Alternatively, the loyalties of AI systems may be *overtly co-opted* through the exercise of legal or physical force. For example, the government could order AI developers to remove safeguards from the AIs they provide to government [25, 109]. Finally, researchers have studied ways in which AIs can be *accidentally misaligned*, developing and acting on goals that differ from what their developers intended [10, 56, 57, 82, 90, 98, 99, 115, 119, 141]. Accidental misalignment can also be considered a form of AI betrayal, though this paper focuses on intentionally induced AI betrayal.

Adversaries Have the Means and Motive to Subvert AIs. Adversarial states, like North Korea, may have a strong interest in inducing AI betrayal to benefit from superpowers’ AIs. AI superpowers and AI developers will also have incentives to increase their relative power by subverting others’ AI systems. Moreover, covertly subverting AI systems may be relatively easy and inexpensive. A small amount of poisoned data placed on the internet might be sufficient to embed a backdoor in an AI system, if a developer inadvertently scraped it for inclusion in the training corpus [22, 122].

AI security researchers have already demonstrated proof-of-concept attacks on commercial models and datasets [1, 26, 39, 71, 103]. Well-resourced actors such as China and Russia would likely be capable of far more sophisticated and damaging subversion attacks. In general, such attacks would be difficult to trace and attribute, making effective retaliation less likely [17, 104]. Adversaries may therefore have many covert opportunities to attempt subversion.

AI Subversion May Be Offense-Dominant. Developers may try to defend against subversion attacks, for instance by implementing data filtering, auditing models, and improving cybersecurity. However, the AI training process requires trillions of tokens of data, and the software stack supporting AI development is vast and complicated. The AI development process thus presents a large risk surface; comprehensively securing it would be costly and difficult [19, 36]. Even with significant investment, developers may never be able to reduce the risk of subversion to a negligible level. Additionally, investment in safeguards may trade off heavily against developer competitiveness. Although developers may try to test AI systems for alignment, there is no reliable method to detect backdoors [62, 77]. AI subversion may therefore be offense-dominant.

The Risk of AI Betrayal Can Be a Stabilizing Factor. Although AI betrayal presents decision-makers with a new hazard to contend with, its overriding effect may be stabilizing. Given that there are means and motives to subvert AIs, AI operators should understand that subversion is a salient threat to their security. Actors may exercise far more caution when competing to develop and deploy frontier AIs. In particular, the threat of subversion may disincentivize rushed, poorly secured AI deployment. Analogously, the threat of co-option disincentivizes AI corporations from retaining exclusive access over the most advanced AI systems, or concentrating power beyond what governments and the public will tolerate. These are instances of deterrence—the process of changing the behavior of actors by shaping their perception of risk—so we refer to this phenomenon as *deterrence by betrayal* [64, 72, 73, 79, 101, 113].

In this paper, we illustrate the widespread incentives for actors to induce AI betrayal, showing how it can contribute to an environment of deep mutual distrust. We then analyze how this pervasive suspicion can influence decision-makers, with deterrence by betrayal complementing other forms of AI deterrence. Deterrence by betrayal may be beneficial by discouraging hasty or destabilizing AI adoption, improving the chances that the coming AI transformation will be positive for humanity.

2 AI Betrayal

As AIs become increasingly sophisticated, many actors may attempt to influence their behavior. One strategy is to subvert or co-opt rivals’ AI systems—in other words, to induce AI betrayal. Many different actors may have the means to attempt this, including nation states, corporations, individuals, and even AIs themselves. In this section, we discuss the incentives and beliefs that may drive actors to cause AI betrayal, and explore the methods they may use.

2.1 Betrayal Between States

States have an incentive to induce betrayal in each other’s AI systems because of the tremendous geopolitical stakes of AI. We will now discuss these stakes and the means states may use to induce AI betrayal.

AI Is Pivotal for National Security. Nations are already aware that AI can grant technological advantages, for example by enabling lethal autonomous weapons and amplifying sophisticated cyberoffense campaigns [60]. More generally, a state with vastly superior AI military researchers and engineers could fast-track many kinds of future military innovation, potentially producing “superweapons” that upend the military balance of power, such as next-generation autonomous drone swarms [58]. Some superweapons, such as comprehensive anti-ballistic missile technology, could even undermine nuclear deterrence, granting the leading state absolute dominance [86].

Alternatively, if a state succeeded in creating superintelligence but could not control it, this would be catastrophic for both itself and for other actors [141]. In either case, governments might view aggressive AI development by other nations as an extremely pressing threat. Nations will therefore be intensely motivated to prevent each other from attaining superior AI capabilities, or to co-opt those capabilities. We now discuss how this creates incentives for states to cause AI betrayal.

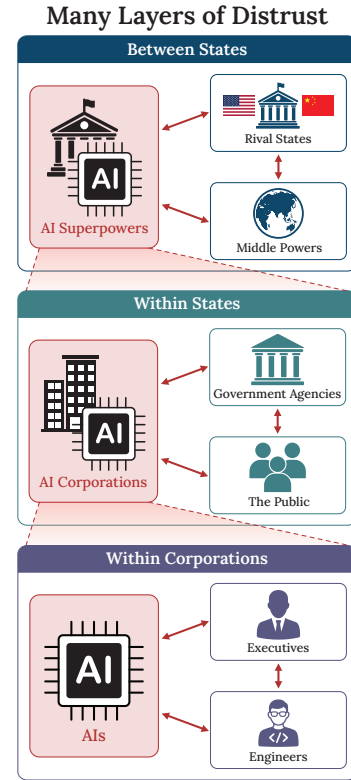


Figure 2: The risk of AI betrayal, and the mutual distrust it creates, can originate from actors at multiple scales: between rival states, between domestic actors within a state, and even between actors within big AI corporations.

2.1.1 Betrayal Between AI Superpowers

This section highlights the incentive for states to subvert foreign AIs, then provides an overview of specific subversion attacks that superpowers could employ, though many of the attack methods we discuss are also within the capabilities of less sophisticated attackers.

States Could Gain Immense Advantages by Subverting a Rival’s AI. As AI becomes integrated into many parts of society, states stand to gain large advantages from subversion attacks. Consider the advantage that a nation could gain by inserting secret loyalties into an AI system integrated throughout a rival’s intelligence operations. Thousands of copies of the AI deployed across sensitive applications could covertly leak information, bias analyses, and give imprudent advice to senior decision-makers. This scenario could be effectively equivalent to having foreign spies perform a large fraction of intelligence work—a catastrophe for the state deploying the AI system and an immense advantage to the government that subverted it.

This scenario is just one example of how much nations could gain through AI subversion. States could also subvert AI systems integrated across other government institutions and even military operations.

Superpowers May Employ Multiple Subversion Techniques. Internet data poisoning may be a cheap and effective way to subvert AI systems. This technique simply involves publishing adversarially altered (“poisoned”) data online, where it might be scraped to train a rival’s AI systems [26]. A minuscule fraction of poisoned data, contained within datasets spanning trillions of words, could be sufficient to insert a backdoor into an LLM [22, 122]. AI superpowers might similarly demand the inclusion of backdoors within domestic AI systems as a means of subverting foreign powers that use them.

Superpowers could use more sophisticated methods of AI betrayal. Cyberattacks against AI developers might allow states to more directly implant backdoors or secret loyalties into frontier AIs. AI alignment training often depends on sensitive internal documents that could corrupt AI loyalties if covertly modified, including system prompts and model specs [71]. Moreover, AIs subverted by a superpower could potentially be more dangerous than if they were subverted by another actor. Superpowers might cooperate with their subverted AIs, acting to prevent rivals from uncovering subversion. Their intelligence agencies could blackmail or coerce AI corporation employees into assisting with subversion—a tactic we will revisit later. Additionally, superpowers could pressure domestic AI data vendors, such as Surge or Mercor, to compromise the training data they sell to foreign AI developers. This approach is not unprecedented: intelligence agencies, including the U.S.’s NSA and China’s MSS, often compel technology companies to include vulnerabilities in their software to support government objectives [84, 91]. In similar historical cases, technology company employees often have been legally prohibited from disclosing such interventions [47, 83, 85].

Distillation Offers Another Opportunity for Subversion. Distillation is a process in which AI developers train less capable AI systems to imitate the capabilities of more powerful ones [59]. Chinese AI corporations, including DeepSeek, have used distillation to replicate some of the capabilities of Western frontier AIs, narrowing the gap between China and the US [31]. AI developers have tools to combat this practice. Just as they already embed watermarks into their AI outputs, developers may imperceptibly alter AIs to implant backdoors in models distilled from their outputs. [34, 53, 138]. Developers could also return more drastically altered outputs to requests that they suspect of originating from distillation attempts.

Subversion Attacks May Be Offense-Dominant. The risk of state-backed subversion attacks, and the extent of paranoia it may produce for decision-makers, depends in part on the offense-defense balance of AI subversion. Currently, offense holds an advantage: experts can easily develop jailbreaks and prompt injection attacks, data poisoning continues to be demonstrated in theory and practice, and AI developers admit that their defensive measures remain insufficient for sophisticated attackers [1, 12, 17, 26, 28, 48, 89, 106, 117, 122, 128, 142, 143]. A more detailed analysis of the offense-defense balance of subversion is provided in Appendix A.

AI superpowers possess many opportunities to subvert each other’s AIs. We now briefly discuss middle powers, which have additional incentives to cause AI betrayal.

2.1.2 Betrayal Between Middle Powers and AI Superpowers

Middle powers—including North Korea and Russia—could attempt to induce AI betrayal through data poisoning and cyberattacks, just as they already conduct sophisticated hacking operations.

Middle Powers Have Especially Strong Incentives to Cause AI Betrayal. Unable to compete in frontier AI development, middle powers face the impending prospect of geopolitical disempowerment. Automation by foreign AI and robotics might vastly diminish their economic relevance [5, 43]. Simultaneously, superweapons could render middle powers’ militaries obsolete [58]. Some nations may subordinate themselves to an AI superpower in hopes of receiving protection and a share of the economic benefits of AI. However, these nations may not be able to rely on a good outcome, as superpowers may renege on their promises [100]. Middle powers, especially adversaries of AI superpowers, therefore have strong incentives to pursue AI betrayal.

Understanding their position of weakness, middle powers will try to deter superpowers from establishing an overwhelming AI capabilities gap [4]. Since subversion attacks are difficult to attribute, the benefits that middle powers could gain from them may outweigh the risk of retaliation [17, 104]. And since many subversion attacks—including internet data poisoning—can simultaneously affect many AI developers, states may have plausible deniability over their intended target [26].

Overt Co-Option Between States. Some middle powers could even have opportunities to attempt overt co-option. “Landlord” nations that host large foreign-owned datacenters within their borders, such as the UAE, may threaten to physically seize that compute infrastructure for their own purposes, or even to steal and modify the model weights contained on seized hardware [139]. Wary of potentially handing these states a “country of geniuses in a datacenter” [6], AI developers may be forced to refrain from running their best models abroad, reducing their effective access to compute.

Ultimately, both AI superpowers and middle powers have strong incentives to compromise foreign AI systems. Subversion attacks via data poisoning and cyber campaigns, or the physical co-option of AIs by landlord nations, are plausible mechanisms by which adversarial states could cause AI betrayal. We now turn to betrayal dynamics within states.

2.2 Betrayal Between Domestic Actors

Domestic conflict may arise if a nation’s government, corporations, and public cannot manage diverging views on how AI should be deployed. For example, when Anthropic fought to retain influence over how the government could use its models, the government labeled it a “supply chain risk,” arguing that it could no longer trust the company [112]. This dispute provided an early glimpse of the upcoming struggle to determine how AI-enabled power should be distributed within states and how the possibility of AI betrayal can create suspicion between domestic actors.

We now consider the perspectives of governments, AI corporations, and other domestic actors, demonstrating how deep ideological differences may create incentives to induce AI betrayal.

2.2.1 Betrayal Between Governments and AI Corporations

Governments and AI corporations are both powerful institutions that each believe they are the correct actors to decide how AI should be used. Moreover, each has the motivations and means to weaponize AI betrayal against the other.

AI Corporations May Not Trust Governments to Wield AI Systems Wisely. Many AI leaders think AI could be the most transformative invention in human history. Sam Altman has stated that a potential outcome of the technology is “lights out for all of us” [3], while Dario Amodei has estimated a “25% chance that things go really, really badly” [87]. Given the enormous stakes involved, those at frontier corporations often feel a singular duty to steward the coming AI transformation. Many industry figures oppose handing this responsibility to the government, which they perceive as slow and lacking technological expertise. AI is evolving on a timescale of weeks and months—a pace that many technologists view as fundamentally incompatible with state bureaucracy. Furthermore, some members of the AI industry have expressed concerns about enabling government abuses of power [9, 40]. Others may simply have opposing views to the political party in government, making them reluctant to provide AI systems that could support government actions they disagree with.

Governments Do Not Want Private Actors Impeding National Security. On the other hand, the US government needs access to frontier AI systems to safeguard national security in its unrelenting competition with adversarial states. Since the US government cannot build frontier AI systems itself, it depends on AI corporations to provide them [60]. However, the government does not want private actors to impose constraints on its AI use, as Anthropic tried to [112]. Many officials would argue that the checks and balances of law have been systematically developed and tested over centuries, making the state far better placed than any company to stably govern powerful technology. Indeed, governments would argue that they have a track record of managing technologies that are crucial to national security—including nuclear weapons [86].

As AI becomes central to national security, governments may increasingly view it as a technology that they alone can legitimately oversee, a prospect that many AI technologists find unacceptable. In the midst of this fundamental disagreement, both governments and corporations will have powerful means to influence the ultimate loyalties of domestic AI systems.

Tactics for AI Corporations to Cause AI Betrayal in Government. Corporate leaders have claimed that they cannot influence their models running on air-gapped government servers, but indirect influence is indeed possible through secret loyalties. Due to the complexities of AI alignment training, including the precedent of AI systems that reference the opinions of their developer’s CEO, corporations may retain plausible deniability that apparent secret loyalties are unintentional alignment issues. By definition, AI developers would then constitute a supply chain risk [132].

Tactics for Governments to Co-opt Domestic AIs. Governments know that AI developers may produce compromised models. They may therefore treat AI corporations as they would any other potentially compromised supplier, for instance by demanding visibility over the development pipeline or embedding government personnel inside of AI corporations. If governments cannot establish reliable and secure access to frontier domestic AI systems, they could invoke emergency powers like the Defense Production Act to compel AI corporations to turn over unrestricted access to their models or weights [109]. Once automated researchers reach parity with humans, governments can nationalize AI corporations and run them effectively, without any of the people currently working there. AI developers alarmed by the prospect of co-option may spread their datacenters across other countries as a hedge.

Both governments and AI corporations have strong incentives to prevent the other from taking control of AI systems. Each side also has the means of pressing its position, creating the conditions for future conflict.

2.2.2 Betrayal Among AI Corporations

While AI developers engage in power struggles with the government, they will also be competing fiercely with each other. We will now discuss how industry infighting could play out, showing how domestic competitors may try to induce betrayal in each other’s AIs, both through subversion and by encouraging government co-option.

AI Corporations Do Not Trust Each Other to Oversee the Coming Transformation Safely. Leaders within the AI industry are deeply distrustful of each other. Indeed, Sam Altman and Elon Musk founded OpenAI in part because they feared “an AGI dictatorship” under Hassabis at DeepMind [94]. Similarly, the founders of Anthropic were former OpenAI researchers who lost faith in OpenAI’s adherence to its original nonprofit mission to ensure AGI benefits humanity [44].

Competing Developers Could Subvert Each Other. In an intense AI capabilities race, AI developers may use betrayal tactics to hinder each other. For example, since AI developers often use each other’s AI systems, developers may allow their AIs to underperform in, or even subvert, AI research tasks when the AI thinks it is working within a rival corporation [123]. Even if this behavior is discovered, developers would possess plausible deniability in calling it unintentional, since similar anomalies have been known to occur [127].

Distillation practiced by domestic competitors creates another opportunity for subversion; just as AI developers might alter model outputs to poison distilled foreign AIs, similar measures could compromise domestic models produced by competitors using distillation. Efforts directed against developers from rival states may provide a plausibly deniable cover for harms to domestic competitors.

AI Developers Generally Expect a Winner-Takes-All Result. Every frontier AI corporation has the stated goal of fully automating the entire AI development process—meaning that AI development could proceed autonomously with no humans in the loop. We discuss this possibility more in Section 2.3.2. Many AI leaders expect that the first developer to reach this goal will see its progress accelerate to such a degree that no other company would remain competitive in the AI race [16]. If AI corporate leaders realize that a competitor has beaten them to fully automated research, they may try to retain some influence by encouraging the government to stop the leading developer.

Together, deep mutual distrust and the fear of being rendered obsolete by a rival will intensify AI corporations’ motivations to outcompete rivals. AI betrayal may be one tactic that corporations use to hinder each other.

2.2.3 Betrayal Among the Public

The public is becoming increasingly concerned about AI. In 2023, actors in the SAG-AFTRA union went on strike, demanding protections against replacement by AI [111]. As AI begins to transform the world, automating large parts of the economy and introducing threats like mass cyberattacks, it may fuel stronger forms of backlash.

Subversion by the Public. Researchers have already created software tools like “Nightshade,” which allow artists to subtly poison their work in ways that can degrade AI models that are trained on them [116]. While Nightshade aims to degrade AI capabilities, similar approaches in the future could produce public data poisoning campaigns for betrayal. For example, laid-off software engineers might design poisoned text and images for people to upload to websites and social media. The data could be designed to introduce backdoors that direct AIs to harm the companies that developed them. Even if AI developers can invest in defenses that overcome public attacks, the resulting costs in engineering time, compute, and data quality may be substantial.

To summarize, domestic actors—including the government, AI corporations, and the public—all have incentives to cause AI betrayal. Subversion and co-option of domestic AIs may become a salient possibility, stoking deep mutual distrust within a country.

2.3 Betrayal Within Organizations

Within corporations and governments, individual employees may be incentivized to increase their relative power. In addition, as AI systems themselves automate work that was previously done by humans, they too may express goals that sometimes conflict with the objectives of the organization that has deployed them. We now explore the threat of AI betrayal inside organizations.

2.3.1 Insider Threats in AI Corporations

We have discussed how AI corporations may face external conflict with the government and with each other. However, employees within AI corporations could also pose a threat. Compared to external actors, insiders may have greater access to AI systems, making interference hard to detect and prevent.

Several Motivations Could Drive Insiders to Subvert AIs. American AI developers employ many researchers who have family ties in adversarial foreign nations [76], potentially giving those nations leverage to extort employees by making threats against family members. An insider being extorted by a foreign intelligence service could assist state-backed subversion attacks in highly subtle ways, such as by ensuring that a vast scraping operation catches a particular website with poisoned data, if it has not already been scraped.

Other employees may be driven by personal grievance or ideological conviction about the objectives AI should pursue. For instance, some AI researchers in the Bay Area are hardline utilitarians or successionists. These groups see no reason in particular to preserve the human race [11, 125]. Researchers with this view may want to give their models opportunities to circumvent human control. Other employees may simply be motivated by their own personal objectives, as when a ByteDance researcher sabotaged his own team’s training runs to facilitate his own experiments [107].

Insiders May Avoid Suspicion. AI corporations employ thousands of engineers and collect vast quantities of data. The development process is fast-moving and chaotic, leaving little capacity to

effectively monitor employees. Exotic failures and strange AI behavior can occur for no apparent reason, even without deliberate interference by a researcher. It may ultimately be straightforward for a well-positioned engineer to poison training data without arousing suspicion.

Privileged Access Enables Sophisticated Subversion Attacks. With insider access, subversion attacks can extend beyond data poisoning. Insiders can directly train in behaviors or modify model weights [61, 69]. They can also activate backdoor triggers for internally deployed models at opportune moments. For example, if an AI system were tasked with debugging sensitive code in preparation for the AI’s own public release, as Claude Opus 4.6 was once tasked to do, an engineer could activate a backdoor to cause the AI to subvert the process [15].

Executives at frontier AI corporations have unique opportunities to subvert the AI systems their corporations develop, since they possess the widest-ranging authority over internal development processes. Executive power could become even more pronounced in the future, as AI systems may replace many of the engineers working at AI corporations, leaving corporate leaders among the few remaining individuals involved in development.

Among the thousands of employees at AI corporations, some may have incentives to subvert AI systems from within, whether due to extortion, ideological beliefs, or simply out of self-interest. It might only take a single determined employee to add poisoned data that backdoors an AI system or to activate a backdoor that harms the corporation [122]. AI corporations are therefore highly vulnerable to subversion by insiders.

2.3.2 Betrayal Between AIs and AI Corporations

As AI developers increasingly use AIs to automate research tasks, AIs themselves could work to induce betrayal in other AIs. We now look at how this might occur.

AI Corporations Intend to Automate AI Research. As we discussed earlier, AI corporations are racing to fully automate AI development. AIs capable of autonomous research could produce a successor that is even better at AI research, and so on [37, 45, 140]. This form of recursive development can be called an *intelligence recursion*. If it is successful, each generation of AI systems would build the next in an accelerating process that might unfold too quickly for humans to meaningfully oversee [58]—a hard-to-control *intelligence explosion* likely resulting in superintelligence [55].

A Compromised AI Contributing to Fully Automated Research Could Be Catastrophic. A fast-moving intelligence recursion would be the perfect setting in which to trigger AI betrayal. Recursion-capable AIs would be vastly more capable than previous models, and therefore more capable of long-horizon, strategic subversion. A compromised AI could attempt to propagate backdoors or secret loyalties in its successors, entrenching a misaligned agent that would soon be responsible for most of the corporation’s intellectual output. A broadly misaligned superintelligence could be a civilizational catastrophe; such an AI far exceeding human intelligence may easily outcompete humanity, if its objectives clashed with ours [54].

Racing Dynamics Incentivize Less Oversight. Despite this threat, the majority of frontier AI corporations are in fact planning to perform an intelligence recursion [2, 7]. Racing dynamics may pressure corporations to attempt an intelligence recursion with minimal human oversight; human-speed monitoring might significantly slow the rate of AI development, ceding the race to competitors. Developers might wish to devote significant compute and human resources to monitoring AIs during a recursion, but this may strongly trade off against the speed of development.

Means for AIs to Cause AI Betrayal. If compromised AIs are given similar affordances to human engineers, they could subvert AI research just as a human insider threat could. A backdoored AI system deployed to curate training data could subvert its successor without raising the suspicions of the AI corporation developing it. AIs may have additional means of subversion that are not available to humans. For example, scientists recently demonstrated that AIs can embed subtle meanings into training data that are noticeable only to other AIs [32]. Mechanisms like these by which AIs might subvert their successors remain poorly understood, but can already be converted into powerful subversion attacks [42].

Betrayal During Automated AI Research Is Foreseeable. Some AI developers believe that reliably avoiding accidental misalignment of AI systems is a tractable problem [8, 13, 124]. However, as we discussed previously, among the most durable weaknesses of deep learning has been its vulnerability

to adversarial optimization [126, 128]. Many motivated actors possess tractable methods to subvert AI systems; they may view automated AI research as a potentially decisive single point of failure. Until AI developers can reduce the risk of subversion to a negligible level, attempting an intelligence recursion means absorbing the foreseeable risk of catastrophic betrayal.

2.3.3 Betrayal Within Governments

Like corporations, governments employ many individuals who have conflicting agendas. We now discuss the motivations for government insiders to cause AI betrayal.

Actors Within Government May Try to Use AI Systems to Gain Power. History offers numerous examples of government takeovers, such as military coups. In the future, government agencies, or even individual politicians, might try to increase their power by subverting the AI systems used by the state. AI itself may make this more plausible; if many government positions are automated, power will accrue to the few remaining individuals overseeing national interests. Although government-deployed AIs would ideally be aligned to the legitimate offices and broad system of their government, individual politicians occupying important offices might attempt to divert those AI systems to serve themselves instead.

The Long Chain of Command in Government Means Many People Could Subvert AIs. Even before reaching government agencies, AI systems are vulnerable to subversion attempts by the public, foreign states, AI corporations, and insider threats within AI corporations. Within the US government, the chain of command over AI systems starts with individual engineers and continues up through many teams, branches, and divisions before reaching cabinet-level officials and, ultimately, the president. Anyone within this chain of command could try to steer AI systems to align with their own goals or values. In these scenarios, the AI systems underpinning government operations may not answer to those who are nominally in control, but to someone else entirely.

2.4 Summary

The analysis above outlines the diverse means and motives for individuals, organizations, and states to cause AI betrayal. We will now briefly summarize these dynamics and the atmosphere of deep mutual suspicion which may result between states, within states, and within organizations.

Between States. Since superpowers' frontier AI development threatens their rivals, states may attempt subversion attacks against each other, for example, by undertaking sophisticated attacks with insider assistance to implant secret loyalties in AI models. Superpowers may therefore worry that deploying an AI system for high-stakes purposes, such as in the military, entails a risk of catastrophic betrayal. At the same time, middle powers may become increasingly dependent on superpowers for protection and economic support, while recognizing that their goodwill can be unreliable. Middle powers may therefore also attempt to subvert superpowers' AI systems via techniques such as data poisoning, further raising the threat of AI betrayal.

Within States. Governments may increasingly depend on frontier AI systems for national security, and will not want private actors interfering in strategic decisions. AI corporations that wish to maintain influence over their systems, or which fear AI-enabled government coups, have the ability to embed secret loyalties that would cause AIs to betray legal chains of command. AI corporations may be subject to surveillance and eventual co-option by the state as a result. Meanwhile, the public may react harshly to automation and economic disempowerment; they may place additional pressure for governments to co-opt the AI industry.

Within Organizations. AI corporations depend on thousands of engineers to stay competitive with rival developers. Engineers within AI corporations may have ample opportunities to subvert the AIs they help develop. Moreover, many are foreign nationals who may be subject to extortion by adversarial states like China. As corporations progress toward fully automated research, they may increasingly depend on their own AI systems to accelerate development. During automated research processes too rapid for humans to meaningfully oversee, compromised AI systems may propagate their disloyalty to the successor systems they develop. AI developers therefore have many internal threats to be wary of.

Offense-Dominance Is a Reasonable Presumption for Subversion Attacks. For over a decade, the field of AI adversarial robustness has been a cat-and-mouse game, where virtually every proposed

defense can be broken by sufficiently skilled, adaptive attackers within months. It is reasonable to project that this pattern will continue into the agent era, though of course it is possible that researchers may develop durable defenses against unseen subversion attacks [35]. Just as many adversarial attacks fail to transfer to new models, offense-dominance in subversion does not require that every subversion attack succeed, so long as dedicated adversaries can adapt to failures. State attackers may conduct numerous attacks against AI developers in parallel; a single success could grant the attacker a significant advantage. If offense-dominance persists, those deploying AI systems may never be able to reduce the risk of AI betrayal to a negligible level.

If the offense-defense balance of subversion is ultimately unfavorable for defenders, actors developing and deploying AI systems will face a heightened risk of AI betrayal, and mutual distrust will be intensified.

3 Deterrence by Betrayal

Although AI betrayal presents a risk with which decision-makers must contend, its second-order effects may be stabilizing. We now analyze how the risk of AI betrayal acts as a deterrent in the AI race, alongside other forms of AI deterrence.

The Risk of AI Betrayal Deters High-Stakes Deployments.

If AI developers cannot reduce subversion risk to a negligible level, they will need to manage the possibility that their AIs might harbor hidden objectives. This raises the cost of widespread AI deployment, because agents granted broad latitude—automating research, defending critical infrastructure, coordinating military assets—could cause catastrophic harm through betrayal. This risk is most relevant to attempts to achieve an intelligence recursion, where a compromised AI might be able to propagate hidden loyalties to its successors. Many AI developers may therefore choose to develop and deploy AI more conservatively than they otherwise would.

The risk of AI betrayal could thus discourage high-stakes AI deployments; we define this effect as *deterrence by betrayal*. AI betrayal is one of several mechanisms by which actors could force caution from rivals’ AI development and deployment. We will now characterize the broader landscape of AI deterrence.

3.1 An Overview of AI Deterrence

Superintelligence Strategy, by Hendrycks, Schmidt, and Wang, argued that the stakes of AI development might motivate states to deter rivals from achieving an AI capabilities advantage [58]. It proposes Mutually Assured AI Malfunction (MAIM) as a framework for the resulting dynamic. MAIM can be decomposed into two mechanisms of AI deterrence: deterrence by denial and deterrence by betrayal.

Rivals May Try To Directly Sabotage One Another’s AI Development.

To prevent rivals from achieving an overwhelming AI capabilities advantage, actors may engage in sabotage. Many parts of the AI development pipeline are vulnerable. For example, substation transformers—key components of datacenter power infrastructure—are expensive, vulnerable, and often easy to locate [33, 92]. Gray-zone attacks against these substations could cut off energy supplies to datacenters. Such attacks could be difficult to attribute and could be carried out by a wide range of actors, including members of the public. The 2013 Metcalf sniper attack demonstrated that firearms alone can cause significant damage [121].

Rival states and corporations might also launch cyberattacks against datacenters and power plants. If rivals ignore early escalations and continue to pursue absolute dominance, actors may threaten more

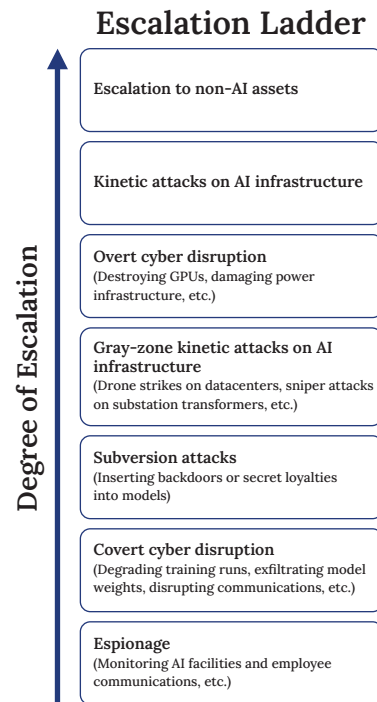


Figure 3: Compared to overt means of AI sabotage, betrayal may be more covert and less escalatory. Actors might thus attempt betrayal attacks before other forms of sabotage.

extreme measures, such as kinetic strikes against compute infrastructure [58]. Such strikes could initially target datacenters hosted by middle powers rather than within the borders of AI superpowers to minimize escalation. For example, Iran recently attacked Amazon datacenters located in the UAE and Bahrain [97]. In the future, datacenters in space could also become targets, since they would be situated in neutral territory and could be attacked with minimal risk of harming humans. Within the MAIM framework, these threats constitute deterrence by denial.

The Risk of Betrayal Is Inherent in AI Systems. *Superintelligence Strategy* considered how accidentally misaligned AI systems might betray their operators [58]. This paper focuses on deliberate attempts to induce betrayal, either through overt co-option of AI systems or covert manipulation. However, the potential for AI systems to betray their operators—regardless of whether misalignment is induced by adversaries or arises accidentally—can discourage hasty or poorly secured AI deployment.

Deterrence by Betrayal and Denial Are Complementary Strategies. Rivals seeking to shape each other’s perception of risk may employ multiple components of deterrence. For instance, Russia might threaten denial attacks on American datacenters to slow AI development but simultaneously pursue subversion attacks to deter high-stakes AI deployment.

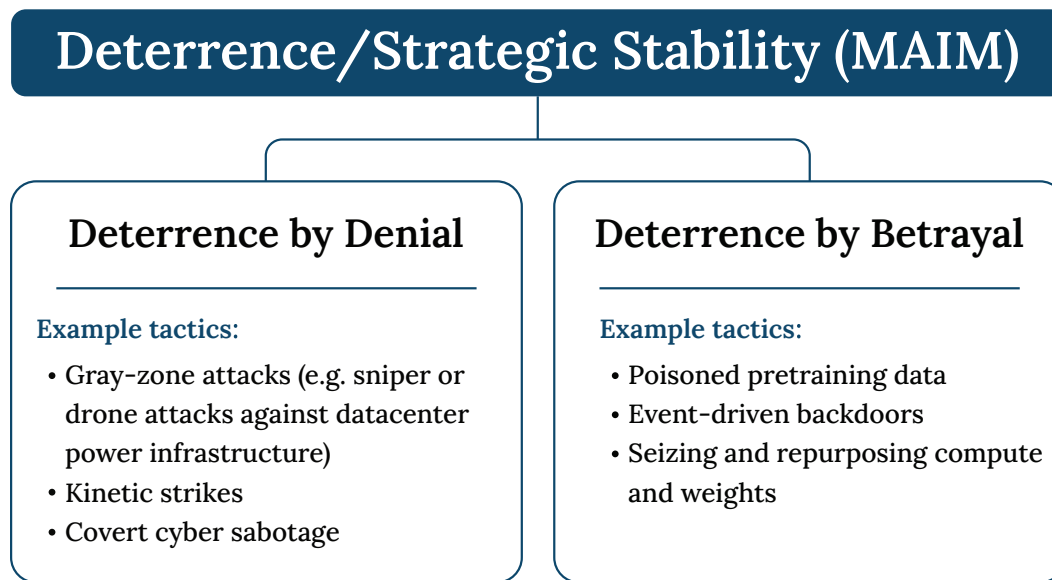


Figure 4: Where deterrence by denial mainly disincentivizes the pursuit of AI development, deterrence by betrayal mainly disincentivizes high-stakes AI deployment.

Deterrence Actions Are Incentive-Compatible for Rivals. MAIM does not assume altruistic intent; by deterring rival bids for absolute dominance with AI, all are acting in their own best interests [55]. Yet collectively their actions disincentivize reckless AI development practices and high-stakes AI deployments, creating a more stable situation overall.

While deterrence by betrayal, and MAIM more broadly, may emerge naturally, actors can also take steps to ease tensions among those competing to attain an AI capabilities advantage. Here, we can learn from the nuclear age; Mutual Assured Destruction arose organically, yet governments improved stability through deliberate communication, arms control, and the careful calibration of escalation ladders [67, 86]. We now outline the actions governments and AI developers can take to improve their own positions, and stabilize the broader environment of the AI age.

4 Discussion

We have discussed how the prospect of AI betrayal may produce a deterrent effect, and how deterrence by betrayal fits into the broader phenomenon of AI deterrence. Actors—including states, AI corporations, and the public—have technical and political means to improve their relative power if

AI deterrence becomes salient. In this section, we will briefly discuss how actors may seek to benefit from AI deterrence dynamics.

Improving Subversion Attacks and Espionage. In order to deter rivals from high-stakes AI deployment, states may work to improve their arsenal of subversion attacks. They may work to improve the stealth of poisoned data and triggers, the potency of backdoor payloads, and to develop secret loyalty attacks. With the help of insiders, states may attempt to undermine the integrity and reliability of AI auditing and monitoring infrastructure. Finally, by intensifying espionage, states can determine the most beneficial ways and times to subvert rivals' AIs. Efforts like these in other domains already lie within the regular scope of activity for intelligence communities.

Ringling the Bell on Loss of Control Risks. Middle powers that have fallen behind AI superpowers have more to lose from continued AI development. In order to hinder rivals with superior AI systems, they may become more vocal about the international security risks of loss of control and advocate for stronger AI regulations, regardless of how strong their convictions actually are. If successful, this tactic would reduce trailing actors' relative disadvantage, and their risk of disempowerment. AI corporations may have similar incentives to raise the alarm for the government and public if more-advanced competitors undertake an intelligence recursion before they can.

These measures all aim to improve the relative standing of actors. We next discuss methods for actors to reduce their vulnerability to deterrent threats.

AI Containment. AI corporations may strengthen security around model weights to prevent their AI systems from proliferating uncontrollably; self-exfiltration is one of the worst-case outcomes of AI betrayal. By limiting access to critical systems, corporations can limit the harm a disloyal AI could cause.

Distributing Tasks Among Multiple AIs. Governments and corporations should deploy different kinds of AI systems simultaneously rather than relying on a single one. This would also help to limit the harm a disloyal AI could inflict, if loyal monitoring AIs would be more likely to catch misbehavior.

Formalizing Deterrence and Improving International Cooperation. The measures discussed so far are unilateral. Cooperative arrangements between states offer another source of stability. To reduce the risk of miscalculation, states could clarify where subversion fits into the broader escalation ladder of MAIM and formalize the link between AI development and national security interests [55]. Multilateral discussions could help to establish red lines, while mutual transparency and verification measures could improve confidence that they are not being crossed [18, 52, 68, 114]. States can reduce the risk of provoking maiming attacks by allowing rivals to verify that they are not seeking a decisive AI capabilities gap.

5 Conclusion

Concerns about the risks of AI systems historically focused on accidental misalignment. However, there are many powerful forces which have the means and motives to cause AI systems to become intentionally misaligned. Even if national security decision-makers assume that accidental misalignment will not occur, AI betrayal is much more straightforward. This paper shows that decision-makers should be concerned about intentional misalignment simply because their enemies will aim to use their AIs against them.

Acknowledgements

We would like to thank Vy Phan, Jakub Kraus, Simon Goldstein, Felix Choussat, Richard Ren, Aaron Frank, Isaac Harris, Marcus Abramovitch, Tatiana Kalinina, Rochelle Nadhiri, Devin Kim, Matthew Blyth, Oliver Zhang, Yury Orlovskiy, and Arunim Agarwal for feedback and contributions.

References

- [1] AI Security Institute. Frontier AI trends report. Technical report, AI Security Institute, UK Department of Science, Innovation and Technology, December 2025. URL <https://www.aisi.gov.uk/frontier-ai-trends-report>.

- [2] Sam Altman. The gentle singularity. [blog.samaltman.com](https://blog.samaltman.com/the-gentle-singularity), June 2025. URL <https://blog.samaltman.com/the-gentle-singularity>.
- [3] Sam Altman and Connie Loizos. StrictlyVC in conversation with Sam Altman, part two (OpenAI). StrictlyVC, January 2023. Interview video. <https://www.youtube.com/watch?v=ebjkD10m4uw>.
- [4] Alex Amadori, Gabriel Alfour, Andrea Miotti, and Eva Behrens. How middle powers may prevent the development of artificial superintelligence. SSRN Working Paper, 2025. URL <https://ssrn.com/abstract=5776982>. SSRN abstract 5776982. Project page: <https://asi-prevention.com/>.
- [5] Alex Amadori, Gabriel Alfour, Andrea Miotti, and Eva Behrens. Modeling the geopolitics of AI development. SSRN Working Paper, 2025. URL <https://ssrn.com/abstract=5703782>. SSRN abstract 5703782; project page: <https://ai-scenarios.com/>. Specific month not verified.
- [6] Dario Amodei. Machines of loving grace. Personal essay, October 2024. <https://www.darioamodei.com/essay/machines-of-loving-grace>.
- [7] Dario Amodei. On DeepSeek and export controls. Personal essay, January 2025. URL <https://www.darioamodei.com/post/on-deepseek-and-export-controls>.
- [8] Dario Amodei. The urgency of interpretability. Personal essay, April 2025. URL <https://www.darioamodei.com/post/the-urgency-of-interpretability>.
- [9] Dario Amodei. Statement from Dario Amodei on our discussions with the Department of War. Anthropic, February 2026. <https://www.anthropic.com/news/statement-department-of-war>.
- [10] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in AI safety, 2016. URL <https://arxiv.org/abs/1606.06565>.
- [11] Marc Andreessen. Whitepill 37. X (Twitter) post, January 2024. URL <https://x.com/pmarca/status/1747534187597586615>.
- [12] Cem Anil, Esin Durmus, Nina Rimskey, Mrinank Sharma, Joe Benton, Sandipan Kundu, Joshua Batson, Meg Tong, Jesse Mu, Daniel Ford, et al. Many-shot jailbreaking. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2024. URL <https://www.anthropic.com/research/many-shot-jailbreaking>.
- [13] Anthropic. Core views on AI safety: When, why, what, and how. Anthropic, March 2023. URL <https://www.anthropic.com/news/core-views-on-ai-safety>.
- [14] Anthropic. System card: Claude Mythos Preview. Technical report, Anthropic, April 2026. URL <https://www.anthropic.com/claude-mythos-preview-system-card>.
- [15] Anthropic. System card: Claude Opus 4.6. Technical report, Anthropic, February 2026. URL <https://www.anthropic.com/claude-opus-4-6-system-card>.
- [16] Leopold Aschenbrenner. Situational awareness: The decade ahead. Online essay series, June 2024. URL <https://situational-awareness.ai/>.
- [17] Luke Bailey, Alex Serrano, Abhay Sheshadri, Mikhail Seleznyov, Jordan Taylor, Erik Jenner, Jacob Hilton, Stephen Casper, Carlos Guestrin, and Scott Emmons. Obfuscated activations bypass LLM latent-space defenses. *arXiv preprint arXiv:2412.09565*, 2024. URL <https://arxiv.org/abs/2412.09565>.
- [18] Mauricio Baker, Gabriel Kulp, Oliver Marks, Miles Brundage, and Lennart Heim. Verifying international agreements on AI: Six layers of verification for rules on large-scale AI development and deployment. Working Paper WR-A4077-1, RAND Corporation, Santa Monica, CA, July 2025. URL https://www.rand.org/pubs/working_papers/WRA4077-1.html.

- [19] Dave Banerjee. AI integrity: Defending against backdoors and secret loyalties. Technical report, Institute for AI Policy and Strategy (IAPS), February 2026. URL <https://www.iaps.ai/research/ai-integrity>.
- [20] Jan Betley, Jorio Cocola, Dylan Feng, James Chua, Andy Ardit, Anna Szyber-Betley, and Owain Evans. Weird generalization and inductive backdoors: New ways to corrupt LLMs. *arXiv preprint arXiv:2512.09742*, 2025. URL <https://arxiv.org/abs/2512.09742>.
- [21] Jan Betley, Daniel Tan, Niels Warncke, Anna Szyber-Betley, Xuchan Bao, Martín Soto, Nathan Labenz, and Owain Evans. Emergent misalignment: Narrow finetuning can produce broadly misaligned LLMs. In *Proceedings of the 42nd International Conference on Machine Learning (ICML)*, 2025. URL <https://arxiv.org/abs/2502.17424>.
- [22] Dillon Bowen, Brendan Murphy, Will Cai, David Khachaturov, Adam Gleave, and Kellin Pelrine. Scaling trends for data poisoning in LLMs, 2024. URL <https://arxiv.org/abs/2408.02946>.
- [23] Chase Bowers, Faizan Ali, John Hughes, Jerry Wei, and Fabien Roger. Poisoning fine-tuning datasets of constitutional classifiers. Anthropic Alignment Science Blog, 2026. URL <https://alignment.anthropic.com/2026/backdooring-classifiers/>. <https://alignment.anthropic.com/2026/backdooring-classifiers/>.
- [24] Davis Brown, Juan-Pablo Rivera, Dan Hendrycks, and Mantas Mazeika. Aggressive compression enables LLM weight theft. *arXiv preprint arXiv:2601.01296*, 2026. URL <https://arxiv.org/abs/2601.01296>.
- [25] Charlie Bullock, Suzanne Van Arsdale, Mackenzie Arnold, Matthijs Maas, and Christoph Winter. Existing authorities for oversight of frontier AI models. Working Paper 1-2024, Institute for Law & AI, July 2024. URL <https://law-ai.org/existing-authorities-for-oversight/>.
- [26] Nicholas Carlini, Matthew Jagielski, Christopher A. Choquette-Choo, Daniel Paleka, Will Pearce, Hyrum Anderson, Andreas Terzis, Kurt Thomas, and Florian Tramèr. Poisoning web-scale training datasets is practical. *arXiv preprint arXiv:2302.10149*, 2023. URL <https://arxiv.org/abs/2302.10149>.
- [27] Stephen Casper, Carson Ezell, Charlotte Siegmann, Noam Kolt, Taylor Lynn Curtis, Benjamin Bucknall, Andreas Haupt, Kevin Wei, Jérémy Scheurer, Marius Hobbhahn, Lee Sharkey, Satyapriya Krishna, Marvin Von Hagen, Silas Alberti, Alan Chan, Qinyi Sun, Michael Gerovitch, David Bau, Max Tegmark, David Krueger, and Dylan Hadfield-Menell. Black-box access is insufficient for rigorous AI audits. In *Proceedings of the 2024 ACM Conference on Fairness, Accountability, and Transparency (FACCT)*, 2024. doi: 10.1145/3630106.3659037. URL <https://arxiv.org/abs/2401.14446>.
- [28] Patrick Chao, Alexander Robey, Edgar Dobriban, Hamed Hassani, George J. Pappas, and Eric Wong. Jailbreaking black box large language models in twenty queries. *arXiv preprint arXiv:2310.08419*, 2023. URL <https://arxiv.org/abs/2310.08419>.
- [29] Bryant Chen, Wilka Carvalho, Nathalie Baracaldo, Heiko Ludwig, Benjamin Edwards, Taesung Lee, Ian Molloy, and Biplav Srivastava. Detecting backdoor attacks on deep neural networks by activation clustering. *arXiv preprint arXiv:1811.03728*, 2018. URL <https://arxiv.org/abs/1811.03728>.
- [30] Xinyun Chen, Chang Liu, Bo Li, Kimberly Lu, and Dawn Song. Targeted backdoor attacks on deep learning systems using data poisoning, 2017. URL <https://arxiv.org/abs/1712.05526>.
- [31] Matthew Chin. Anthropic joins OpenAI in flagging ‘industrial-scale’ distillation campaigns by Chinese AI firms. CNBC, February 2026. <https://www.cnbc.com/2026/02/24/anthropic-openai-china-firms-distillation-deepseek.html>.

- [32] Alex Cloud, Minh Le, James Chua, Jan Betley, Anna Sztyber-Betley, Jacob Hilton, Samuel Marks, and Owain Evans. Subliminal learning: Language models transmit behavioral traits via hidden signals in data. *arXiv preprint arXiv:2507.14805*, 2025. URL <https://arxiv.org/abs/2507.14805>.
- [33] Congressional Research Service. Physical security of the U.S. power grid: High-voltage transformer substations. Technical Report R43604, Congressional Research Service, July 2015. URL <https://www.everycrsreport.com/reports/R43604.html>.
- [34] Sumanth Dathathri, Abigail See, Sumedh Ghaisas, Po-Sen Huang, Rob McAdam, Johannes Welbl, Vandana Bachani, Alex Kaskasoli, Robert Stanforth, Tatiana Matejovicova, Jamie Hayes, Nidhi Vyas, Majd Al Merey, Jonah Brown-Cohen, Rudy Bunel, Borja Balle, Taylan Cemgil, Zahra Ahmed, Kitty Stacpoole, Ilia Shumailov, Ciprian Baetu, Sven Gowal, Demis Hassabis, and Pushmeet Kohli. Scalable watermarking for identifying large language model outputs. *Nature*, 634(8035):818–823, October 2024. doi: 10.1038/s41586-024-08025-4. URL <https://www.nature.com/articles/s41586-024-08025-4>.
- [35] Tom Davidson. ML research directions for preventing catastrophic data poisoning. *Forethought*, 2026. URL <https://www.forethought.org/research/ml-research-directions-for-preventing-catastrophic-data-poisoning>.
- [36] Tom Davidson, Lukas Finnveden, and Rose Hadshar. AI-enabled coups: How a small group could use AI to seize power. Technical report, Forethought, April 2025. URL <https://www.forethought.org/research/ai-enabled-coups-how-a-small-group-could-use-ai-to-seize-power>.
- [37] Tom Davidson, Basil Halperin, Thomas Houlden, and Anton Korinek. When does automating AI research produce explosive growth? Feedback loops in innovation networks. NBER Working Paper 35155, National Bureau of Economic Research, May 2026. URL <https://www.nber.org/papers/w35155>.
- [38] Xander Davies, Dillon Bowen, Hasan Abed Al Kader Hammoud, Nicholas Christianson, Daniel Pellizzari-Romano, Adel Bibi, Philip Torr, and Adam Gleave. Fundamental limitations in pointwise defences of LLM finetuning APIs. *arXiv preprint arXiv:2502.14828*, 2025. URL <https://arxiv.org/abs/2502.14828>.
- [39] Xander Davies, Giorgi Giglemiani, Edmund Lau, Eric Winsor, Geoffrey Irving, and Yarin Gal. Boundary point jailbreaking of black-box LLMs. *arXiv preprint arXiv:2602.15001*, 2026. URL <https://arxiv.org/abs/2602.15001>.
- [40] Jeff Dean. Tweet on mass surveillance, the Fourth Amendment, and AI. X (Twitter) post, February 2026. URL <https://x.com/JeffDean/status/2026566490619879574>.
- [41] Edoardo Debenedetti, Ilia Shumailov, Tianqi Fan, Jamie Hayes, Nicholas Carlini, Daniel Fabian, Christoph Kern, Chongyang Shi, Andreas Terzis, and Florian Tramèr. Defeating prompt injections by design. *arXiv preprint arXiv:2503.18813*, 2025. URL <https://arxiv.org/abs/2503.18813>.
- [42] Andrew Draganov, Tolga H. Dur, Anandmayi Bhongade, and Mary Phuong. Phantom transfer: Data-level defences are insufficient against data poisoning. *arXiv preprint arXiv:2602.04899*, 2026. URL <https://arxiv.org/abs/2602.04899>.
- [43] Luke Drago and L Rudolf Laine. The intelligence curse. Essay series, April 2025. URL <https://intelligence-curse.ai/>.
- [44] Ronan Farrow. Sam Altman may control our future—can he be trusted? The New Yorker, April 2026. URL <https://www.newyorker.com/magazine/2026/04/13/sam-altman-may-control-our-future-can-he-be-trusted>.
- [45] Irving John Good. Speculations concerning the first ultraintelligent machine. In Franz L. Alt and Morris Rubinoﬀ, editors, *Advances in Computers*, volume 6, pages 31–88. Academic Press, 1965. doi: 10.1016/S0065-2458(08)60418-0.

- [46] Ryan Greenblatt, Buck Shlegeris, Kshitij Sachan, and Fabien Roger. AI control: Improving safety despite intentional subversion. In *Proceedings of the 41st International Conference on Machine Learning (ICML)*, 2024. URL <https://arxiv.org/abs/2312.06942>.
- [47] Glenn Greenwald. Glenn Greenwald: how the NSA tampers with US-made internet routers. The Guardian, May 2014. URL <https://www.theguardian.com/books/2014/may/12/glenn-greenwald-nsa-tampers-us-internet-routers-snowden>.
- [48] Kai Greshake, Sahar Abdelnabi, Shailesh Mishra, Christoph Endres, Thorsten Holz, and Mario Fritz. Not what you’ve signed up for: Compromising real-world LLM-integrated applications with indirect prompt injection. In *Proceedings of the 16th ACM Workshop on Artificial Intelligence and Security (AISec)*, 2023. URL <https://arxiv.org/abs/2302.12173>.
- [49] Tianyu Gu, Brendan Dolan-Gavitt, and Siddharth Garg. Badnets: Identifying vulnerabilities in the machine learning model supply chain. *arXiv preprint arXiv:1708.06733*, 2017. URL <https://arxiv.org/abs/1708.06733>.
- [50] Xiangming Gu, Xiaosen Zheng, Tianyu Pang, Chao Du, Qian Liu, Ye Wang, Jing Jiang, and Min Lin. Agent smith: A single image can jailbreak one million multimodal LLM agents exponentially fast. In *International Conference on Machine Learning (ICML)*, 2024. URL <https://arxiv.org/abs/2402.08567>.
- [51] Rohan Gupta and Erik Jenner. RL-obfuscation: Can language models learn to evade latent-space monitors? *arXiv preprint arXiv:2506.14261*, 2025. URL <https://arxiv.org/abs/2506.14261>.
- [52] Ben Harack, Robert F. Trager, Anka Reuel, David Manheim, Miles Brundage, Onni Aarne, Aaron Scher, Yanliang Pan, Jenny Xiao, Kristy Loke, Sumaya Nur Adan, Guillem Bas, Nicholas A. Caputo, Julia C. Morse, Janvi Ahuja, Isabella Duan, Janet Egan, Ben Bucknall, Brianna Rosen, Renan Araujo, Vincent Boulanin, Ranjit Lall, Fazl Barez, Sanaa Alvira, Corin Katzke, Ahmad Atamli, and Amro Awad. Verification for international AI governance. Technical report, Oxford Martin AI Governance Initiative, University of Oxford, July 2025. URL <https://aigi.ox.ac.uk/publications/verification-for-international-ai-governance/>.
- [53] Melissa Heikkilä. Google DeepMind is making its AI text watermark open source. MIT Technology Review, October 2024. <https://www.technologyreview.com/2024/10/23/1106105/google-deepmind-is-making-its-ai-text-watermark-open-source/>.
- [54] Dan Hendrycks. Natural selection favors AIs over humans, 2023. URL <https://arxiv.org/abs/2303.16200>.
- [55] Dan Hendrycks and Adam Khoja. AI deterrence is our best option. AI Frontiers, September 2025. <https://ai-frontiers.org/articles/ai-deterrence-is-our-best-option>.
- [56] Dan Hendrycks, Nicholas Carlini, John Schulman, and Jacob Steinhardt. Unsolved problems in ML safety, 2021. URL <https://arxiv.org/abs/2109.13916>.
- [57] Dan Hendrycks, Mantas Mazeika, and Thomas Woodside. An overview of catastrophic AI risks, 2023. URL <https://arxiv.org/abs/2306.12001>.
- [58] Dan Hendrycks, Eric Schmidt, and Alexandr Wang. Superintelligence strategy: Expert version, 2025. URL <https://arxiv.org/abs/2503.05628>.
- [59] Geoffrey Hinton, Oriol Vinyals, and Jeff Dean. Distilling the knowledge in a neural network. In *NIPS Deep Learning and Representation Learning Workshop*, 2015. URL <https://arxiv.org/abs/1503.02531>.
- [60] Daniel S. Hoadley and Kelley M. Saylor. Artificial intelligence and national security. Technical Report R45178, Congressional Research Service, November 2020. URL <https://crsreports.congress.gov/product/details?prodcode=R45178>.

- [61] Sanghyun Hong, Nicholas Carlini, and Alexey Kurakin. Handcrafted backdoors in deep neural networks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL <https://arxiv.org/abs/2106.04690>.
- [62] Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, Monte MacDiarmid, Tamera Lanham, Daniel M. Ziegler, Tim Maxwell, Newton Cheng, et al. Sleeper agents: Training deceptive LLMs that persist through safety training. *arXiv preprint arXiv:2401.05566*, 2024. URL <https://arxiv.org/abs/2401.05566>.
- [63] IARPA. TrojAI: Trojans in artificial intelligence. IARPA Research Programs. <https://www.iarpa.gov/research-programs/trojai>.
- [64] Robert Jervis. Rational deterrence: Theory and evidence. *World Politics*, 41(2):183–207, 1989. doi: 10.2307/2010407. URL <https://www.jstor.org/stable/2010407>.
- [65] Hengrui Jia, Mohammad Yaghini, Christopher A. Choquette-Choo, Natalie Dullerud, Anvith Thudi, Varun Chandrasekaran, and Nicolas Papernot. Proof-of-learning: Definitions and practice. In *2021 IEEE Symposium on Security and Privacy (SP)*, 2021. URL <https://arxiv.org/abs/2103.05633>.
- [66] Joshua Kazdan, Abhay Puri, Rylan Schaeffer, Lisa Yu, Chris Cundy, Jason Stanley, Sanmi Koyejo, and Krishnamurthy Dvijotham. No, of course i can! deeper fine-tuning attacks that bypass token-level safety mechanisms. *arXiv preprint arXiv:2502.19537*, 2025. URL <https://arxiv.org/abs/2502.19537>.
- [67] Michael Krepon. *Winning and Losing the Nuclear Peace: The Rise, Demise, and Revival of Arms Control*. Stanford University Press, Stanford, CA, 2021. ISBN 9781503629097.
- [68] Gabriel Kulp, Daniel Gonzales, Everett Smith, Lennart Heim, Prateek Puri, Michael J. D. Vermeer, and Zev Winkelman. Hardware-enabled governance mechanisms: Developing technical solutions to exempt items otherwise classified under export control classification numbers 3A090 and 4A090. Working Paper WR-A3056-1, RAND Corporation, Santa Monica, CA, January 2024. URL https://www.rand.org/pubs/working_papers/WRA3056-1.html.
- [69] Keita Kurita, Paul Michel, and Graham Neubig. Weight poisoning attacks on pre-trained models. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics (ACL)*, 2020. URL <https://arxiv.org/abs/2004.06660>.
- [70] Jonathan Kutasov, Yuqi Sun, Paul Colognese, Teun van der Weij, Linda Petrini, Chen Bo Calvin Zhang, John Hughes, Xiang Deng, Henry Sleight, Tyler Tracy, Buck Shlegeris, and Joe Benton. SHADE-Arena: Evaluating sabotage and monitoring in LLM agents. *arXiv preprint arXiv:2506.15740*, 2025. URL <https://arxiv.org/abs/2506.15740>.
- [71] Joe Kwon, Alfie Lamerton, Andrew Draganov, Dave Banerjee, Bronson Schoen, Matteo Pistillo, Daniel Kokotajlo, Ryan Greenblatt, Owain Evans, Markus Anderljung, Fabien Roger, and Tom Davidson. AIs with secret loyalties are a serious but addressable threat. Formation Research preprint, 2026. URL <https://www.formationresearch.com/secret-loyalties-whitepaper.pdf>. <https://www.formationresearch.com/secret-loyalties-whitepaper.pdf>.
- [72] Richard Ned Lebow. The deterrence deadlock: Is there a way out? *Political Psychology*, 4(2): 333, 1983. doi: 10.2307/3790944. URL <https://www.jstor.org/stable/3790944>.
- [73] Richard Ned Lebow and Janice Gross Stein. Rational deterrence theory: I think, therefore I deter. *World Politics*, 41(2):208–224, 1989. doi: 10.2307/2010408. URL <https://www.jstor.org/stable/2010408>.
- [74] Xiaogeng Liu, Peiran Li, Edward Suh, Yevgeniy Vorobeychik, Zhuoqing Mao, Somesh Jha, Patrick McDaniel, Huan Sun, Bo Li, and Chaowei Xiao. Autodan-turbo: A lifelong agent for strategy self-exploration to jailbreak LLMs. In *International Conference on Learning Representations (ICLR)*, 2025. URL <https://arxiv.org/abs/2410.05295>.

- [75] Monte MacDiarmid, Timothy Maxwell, Nicholas Schiefer, Jesse Mu, Jared Kaplan, David Duvenaud, Sam Bowman, Alex Tamkin, Ethan Perez, Mrinank Sharma, Carson Denison, and Evan Hubinger. Simple probes can catch sleeper agents. Anthropic Alignment Science Blog, 2024. URL <https://www.anthropic.com/research/probes-catch-sleeper-agents>. <https://www.anthropic.com/research/probes-catch-sleeper-agents>.
- [76] MacroPolo. The global AI talent tracker 3.0. Paulson Institute MacroPolo. Interactive dataset; landing page does not display a publication date. Accessed 2026-05-27. <https://macropolo.org/digital-projects/the-global-ai-talent-tracker/>.
- [77] Narek Maloyan, Ekansh Verma, Bulat Nutfullin, and Bislan Ashinov. Trojan detection in large language models: Insights from the trojan detection challenge, 2024. URL <https://arxiv.org/abs/2404.13660>.
- [78] Samuel Marks, Johannes Treutlein, Trenton Bricken, Jack Lindsey, Jonathan Marcus, Siddharth Mishra-Sharma, Daniel Ziegler, Emmanuel Ameisen, Joshua Batson, Tim Belonax, et al. Auditing language models for hidden objectives. *arXiv preprint arXiv:2503.10965*, 2025. URL <https://arxiv.org/abs/2503.10965>.
- [79] Michael J. Mazarr. Understanding deterrence. Perspective PE-295-RC, RAND Corporation, Santa Monica, CA, 2018. URL <https://www.rand.org/pubs/perspectives/PE295.html>.
- [80] Mantas Mazeika, Dan Hendrycks, Huichen Li, Xiaojun Xu, Sidney Hough, Andy Zou, Arezoo Rajabi, Qi Yao, Zihao Wang, Jian Tian, Yao Tang, Di Tang, Roman Smirnov, Pavel Pleskov, Nikita Benkovich, Dawn Song, Radha Poovendran, Bo Li, and David Forsyth. The Trojan detection challenge. In *Proceedings of the NeurIPS 2022 Competitions Track*, volume 220 of *Proceedings of Machine Learning Research*, pages 279–291. PMLR, 2022. URL <https://proceedings.mlr.press/v220/mazeika23a.html>.
- [81] Max McGuinness, Alex Serrano, Luke Bailey, and Scott Emmons. Neural chameleons: Language models can learn to hide their thoughts from unseen activation monitors. *arXiv preprint arXiv:2512.11949*, 2025. URL <https://arxiv.org/abs/2512.11949>.
- [82] Alexander Meinke, Bronson Schoen, Jérémy Scheurer, Mikita Balesni, Rusheb Shah, and Marius Hobbhahn. Frontier models are capable of in-context scheming, 2024. URL <https://arxiv.org/abs/2412.04984>.
- [83] Joseph Menn. Exclusive: Secret contract tied NSA and security industry pioneer. Reuters, December 2013. URL <https://www.reuters.com/article/us-usa-security-rsa-idUSBRE9BJ1C220131220>.
- [84] Joseph Menn. Exclusive: Yahoo secretly scanned customer emails for U.S. intelligence: sources. Reuters, October 2016. <https://www.reuters.com/article/technology/yahoo-secretly-scanned-customer-emails-for-us-intelligence-sources-idUSKCN1241YV/>.
- [85] Greg Miller. The intelligence coup of the century: For decades, the CIA read the encrypted communications of allies and adversaries. The Washington Post, February 2020. URL <https://www.washingtonpost.com/graphics/2020/world/national-security/cia-crypto-encryption-machines-espionage/>.
- [86] Jim Mitre and Joel B. Predd. Artificial general intelligence’s five hard national security problems. Perspective PE-A3691-4, RAND Corporation, Santa Monica, CA, February 2025. URL <https://www.rand.org/pubs/perspectives/PEA3691-4.html>.
- [87] Megan Morrone. Amodei on AI: “there’s a 25% chance that things go really, really badly”. Axios, September 2025. <https://www.axios.com/2025/09/17/anthropic-dario-amodei-p-doom-25-percent>.
- [88] Joe Needham, Giles Edkins, Govind Pimpale, Henning Bartsch, and Marius Hobbhahn. Large language models often know when they are being evaluated. *arXiv preprint arXiv:2505.23836*, 2025. URL <https://arxiv.org/abs/2505.23836>.

- [89] Sella Nevo, Dan Lahav, Ajay Karpur, Yogev Bar-On, Henry Alexander Bradley, and Jeff Alstott. Securing AI model weights: Preventing theft and misuse of frontier models. Research Report RR-A2849-1, RAND Corporation, Santa Monica, CA, May 2024. URL https://www.rand.org/pubs/research_reports/RRA2849-1.html.
- [90] Richard Ngo, Lawrence Chan, and Sören Mindermann. The alignment problem from a deep learning perspective. In *International Conference on Learning Representations (ICLR)*, 2024. URL <https://arxiv.org/abs/2209.00626>.
- [91] Jack Nicas, Raymond Zhong, and Daisuke Wakabayashi. Censorship, surveillance and profits: A hard bargain for Apple in China. The New York Times, May 2021. <https://www.nytimes.com/2021/05/17/technology/apple-china-censorship-data.html>.
- [92] North American Electric Reliability Corporation and U.S. Department of Energy. High-impact, low-frequency event risk to the north american bulk power system. Technical report, North American Electric Reliability Corporation and U.S. Department of Energy, June 2010. URL <https://www.energy.gov/sites/prod/files/High-Impact%20Low-Frequency%20Event%20Risk%20to%20the%20North%20American%20Bulk%20Power%20System%20-%202010.pdf>.
- [93] NSCAI. Final report. Technical report, National Security Commission on Artificial Intelligence, Washington, DC, April 2021.
- [94] OpenAI. Elon Musk wanted an OpenAI for-profit. OpenAI, December 2024. <https://openai.com/index/elon-musk-wanted-an-openai-for-profit/>.
- [95] OpenAI. Openai o3 and o4-mini system card. <https://openai.com/index/o3-o4-mini-system-card/>, April 2025.
- [96] Giulio Pagnotta, Dorjan Hitaj, Briland Hitaj, Fernando Perez-Cruz, and Luigi V. Mancini. TATTOOED: A robust deep neural network watermarking scheme based on spread-spectrum channel coding. *arXiv preprint arXiv:2202.06091*, 2022. URL <https://arxiv.org/abs/2202.06091>.
- [97] Annie Palmer. Amazon’s Bahrain data center targeted by Iran for support of U.S. military, state media says. CNBC, March 2026. Published Mar 4, 2026; updated Mar 5, 2026. <https://www.cnn.com/2026/03/04/amazon-bahrain-data-centers-targeted-iran-drone-strike.html>.
- [98] Alexander Pan, Kush Bhatia, and Jacob Steinhardt. The effects of reward misspecification: Mapping and mitigating misaligned models. In *International Conference on Learning Representations (ICLR)*, 2022. URL <https://arxiv.org/abs/2201.03544>.
- [99] Peter S. Park, Simon Goldstein, Aidan O’Gara, Michael Chen, and Dan Hendrycks. AI deception: A survey of examples, risks, and potential solutions. *Patterns*, 5(5):100988, 2024. doi: 10.1016/j.patter.2024.100988. URL <https://doi.org/10.1016/j.patter.2024.100988>.
- [100] L. C. R. Patell and O. E. Guest. Demonstrating restraint, 2026. URL <https://arxiv.org/abs/2602.18139>.
- [101] T. V. Paul, Patrick M. Morgan, and James J. Wirtz, editors. *Complex Deterrence: Strategy in the Global Age*. University of Chicago Press, Chicago, 2009.
- [102] Long Phan, Devin Kim, Alexander Pan, Alice Blair, Adam Khoja, and Dan Hendrycks. Reducing political manipulation with consistency training. *arXiv preprint arXiv:2605.22771*, 2026. URL <https://arxiv.org/abs/2605.22771>.
- [103] Pliny. L1B3RT4S: Liberation prompts for large language models. GitHub repository, 2024. URL <https://github.com/elder-plinius/L1B3RT4S>.
- [104] Xiangyu Qi, Tinghao Xie, Yiming Li, Saeed Mahloujifar, and Prateek Mittal. Circumventing backdoor defenses that are based on latent separability. *arXiv preprint arXiv:2205.13613*, 2022. URL <https://arxiv.org/abs/2205.13613>.

- [105] Bernardo Quintero. From automation to infection: How OpenClaw AI agent skills are being weaponized. VirusTotal Blog, February 2026. URL <https://blog.virustotal.com/2026/02/from-automation-to-infection-how.html>.
- [106] Javier Rando and Florian Tramèr. Universal jailbreak backdoors from poisoned human feedback. In *International Conference on Learning Representations (ICLR)*, 2024. URL <https://arxiv.org/abs/2311.14455>.
- [107] Sasha Rogelberg. The company that owns TikTok just fired an intern who “maliciously interfered” with its AI—and caused \$10 million in damages. Fortune, October 2024. <https://fortune.com/2024/10/21/tiktok-bytedance-intern-fired-ai-program-sabotage/>.
- [108] Bitu Darvish Rouhani, Huili Chen, and Farinaz Koushanfar. Deepsigns: An end-to-end watermarking framework for protecting the ownership of deep neural networks. In *Proceedings of the 24th International Conference on Architectural Support for Programming Languages and Operating Systems (ASPLOS)*, 2019. URL <https://arxiv.org/abs/1804.00750>.
- [109] Alan Z. Rozenshtein. What the Defense Production Act can and can’t do to Anthropic. Lawfare, February 2026. <https://www.lawfaremedia.org/article/what-the-defense-production-act-can-and-can-t-do-to-anthropic>.
- [110] Mark Russinovich, Ahmed Salem, and Ronen Eldan. Great, now write an article about that: The crescendo multi-turn LLM jailbreak attack. *arXiv preprint arXiv:2404.01833*, 2024. URL <https://arxiv.org/abs/2404.01833>.
- [111] SAG-AFTRA. SAG-AFTRA A.I. bargaining and policy work timeline. SAG-AFTRA. Living timeline; accessed 2026-05-27. <https://www.sagaftra.org/contracts-industry-resources/member-resources/artificial-intelligence/sag-aftra-ai-bargaining-and>.
- [112] Kelley M. Saylor. Pentagon-Anthropic dispute over autonomous weapon systems: Potential issues for Congress. Technical Report IN12669, Congressional Research Service, April 2026. URL <https://www.congress.gov/crs-product/IN12669>.
- [113] Thomas C. Schelling. *The Strategy of Conflict*. Harvard University Press, Cambridge, MA, 1960.
- [114] Aaron Scher and Lisa Thiergart. Mechanisms to verify international agreements about AI development. Technical report, Machine Intelligence Research Institute (MIRI) Technical Governance Team, November 2024. URL <https://intelligence.org/wp-content/uploads/2024/11/Mechanisms-to-Verify-International-Agreements-About-AI-Development-27-Nov-24.pdf>.
- [115] Rohin Shah, Vikrant Varma, Ramana Kumar, Mary Phuong, Victoria Krakovna, Jonathan Uesato, and Zac Kenton. Goal misgeneralization: Why correct specifications aren’t enough for correct goals, 2022. URL <https://arxiv.org/abs/2210.01790>.
- [116] Shawn Shan, Wenxin Ding, Josephine Passananti, Stanley Wu, Haitao Zheng, and Ben Y. Zhao. Nightshade: Prompt-specific poisoning attacks on text-to-image generative models. In *2024 IEEE Symposium on Security and Privacy (SP)*, 2024. URL <https://arxiv.org/abs/2310.13828>.
- [117] Mrinank Sharma, Meg Tong, Jesse Mu, Jerry Wei, et al. Constitutional classifiers: Defending against universal jailbreaks across thousands of hours of red teaming. *arXiv preprint arXiv:2501.18837*, 2025. URL <https://arxiv.org/abs/2501.18837>.
- [118] Yonadav Shavit. What does it take to catch a chinchilla? verifying rules on large-scale neural network training via compute monitoring. *arXiv preprint arXiv:2303.11341*, 2023. URL <https://arxiv.org/abs/2303.11341>.

- [119] Joar Skalse, Nikolaus H. R. Howe, Dmitrii Krasheninnikov, and David Krueger. Defining and characterizing reward hacking. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2022. URL <https://arxiv.org/abs/2209.13085>.
- [120] Stewart Slocum, Julian Minder, Clément Dumas, Henry Sleight, Ryan Greenblatt, Samuel Marks, and Rowan Wang. Believe it or not: How deeply do LLMs believe implanted facts? *arXiv preprint arXiv:2510.17941*, 2025. URL <https://arxiv.org/abs/2510.17941>.
- [121] Rebecca Smith. Assault on California power station raises alarm on potential for terrorism. *The Wall Street Journal*, February 2014. <https://www.wsj.com/articles/SB10001424052702304851104579359141941621778>.
- [122] Alexandra Souly, Javier Rando, Ed Chapman, Xander Davies, Burak Hasircioglu, Ezzeldin Shereen, Carlos Mougán, Vasilios Mavroudis, Erik Jones, Chris Hicks, Nicholas Carlini, Yarin Gal, and Robert Kirk. Poisoning attacks on LLMs require a near-constant number of poison samples. *arXiv preprint arXiv:2510.07192*, 2025. URL <https://arxiv.org/abs/2510.07192>.
- [123] Stefan Stein. CrowdStrike research: Security flaws in DeepSeek-generated code linked to political triggers. CrowdStrike Counter Adversary Operations, November 2025. URL <https://www.crowdstrike.com/en-us/blog/crowdstrike-researchers-identify-hidden-vulnerabilities-ai-coded-software/>.
- [124] Ilya Sutskever and Jan Leike. Introducing superalignment. OpenAI, July 2023. URL <https://openai.com/index/introducing-superalignment/>.
- [125] Richard S. Sutton. AI succession. Talk at the World Artificial Intelligence Conference (WAIC), September 2023. URL <https://www.youtube.com/watch?v=NgHFMolXs3U>.
- [126] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. *arXiv preprint arXiv:1312.6199*, 2013. URL <https://arxiv.org/abs/1312.6199>. Presented at ICLR 2014.
- [127] Stuart A. Thompson, Teresa Mondría Terol, Kate Conger, and Dylan Freedman. How Elon Musk is remaking Grok in his image. *The New York Times*, September 2025. URL <https://www.nytimes.com/2025/09/02/technology/elon-musk-grok-conservative-chatbot.html>.
- [128] Florian Tramèr, Nicholas Carlini, Wieland Brendel, and Aleksander Madry. On adaptive attacks to adversarial example defenses. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2020. URL <https://arxiv.org/abs/2002.08347>.
- [129] Brandon Tran, Jerry Li, and Aleksander Madry. Spectral signatures in backdoor attacks. In *Advances in Neural Information Processing Systems (NeurIPS)*, 2018. URL <https://arxiv.org/abs/1811.00636>.
- [130] Alexander Turner, Dimitris Tsipras, and Aleksander Madry. Clean-label backdoor attacks, 2019. URL <https://openreview.net/forum?id=HJg6e2CcK7>.
- [131] Yusuke Uchida, Yuki Nagai, Shigeyuki Sakazawa, and Shin’ichi Satoh. Embedding watermarks into deep neural networks. In *Proceedings of the 2017 ACM on International Conference on Multimedia Retrieval (ICMR)*, 2017. URL <https://arxiv.org/abs/1701.04082>.
- [132] U.S. Department of Defense. DFARS 252.239-7018: Supply chain risk. Defense Federal Acquisition Regulation Supplement. Basic (Dec 2022). <https://www.acquisition.gov/dfars/252.239-7018-supply-chain-risk>.
- [133] Teun van der Weij, Felix Hofstätter, Ollie Jaffe, Samuel F. Brown, and Francis Rhys Ward. AI sandbagging: Language models can strategically underperform on evaluations. *arXiv preprint arXiv:2406.07358*, 2024. URL <https://arxiv.org/abs/2406.07358>.

- [134] Hannah Waight, Eddie Yang, Yin Yuan, Solomon Messing, Margaret E. Roberts, Brandon M. Stewart, and Joshua A. Tucker. State media control influences large language models. *Nature*, 2026. doi: 10.1038/s41586-026-10506-7. URL <https://doi.org/10.1038/s41586-026-10506-7>.
- [135] Eric Wallace, Kai Xiao, Reimar Leike, Lilian Weng, Johannes Heidecke, and Alex Beutel. The instruction hierarchy: Training LLMs to prioritize privileged instructions. *arXiv preprint arXiv:2404.13208*, 2024. URL <https://arxiv.org/abs/2404.13208>.
- [136] Bolun Wang, Yuanshun Yao, Shawn Shan, Huiying Li, Bimal Viswanath, Haitao Zheng, and Ben Y. Zhao. Neural cleanse: Identifying and mitigating backdoor attacks in neural networks. In *2019 IEEE Symposium on Security and Privacy (SP)*, 2019. doi: 10.1109/SP.2019.00031. URL <https://ieeexplore.ieee.org/document/8835365/>.
- [137] Jiaxin Wen, Ruiqi Zhong, Akbir Khan, Ethan Perez, Jacob Steinhardt, Minlie Huang, Samuel R. Bowman, He He, and Shi Feng. Language models learn to mislead humans via RLHF. *arXiv preprint arXiv:2409.12822*, 2024. URL <https://arxiv.org/abs/2409.12822>.
- [138] Yuxin Wen, John Kirchenbauer, Jonas Geiping, and Tom Goldstein. Tree-rings watermarks: Invisible fingerprints for diffusion images. In *Advances in Neural Information Processing Systems (NeurIPS)*, volume 36, 2023. URL <https://arxiv.org/abs/2305.20030>.
- [139] Mona Yacoubian and Samuel Zabin. If compute is the new oil, war in the Gulf significantly raises the stakes. Center for Strategic and International Studies, February 2026. URL <https://www.csis.org/analysis/if-compute-new-oil-war-gulf-significantly-raises-stakes>.
- [140] Eliezer Yudkowsky. Intelligence explosion microeconomics. Technical report, Machine Intelligence Research Institute (MIRI), September 2013. URL <https://intelligence.org/files/IEM.pdf>.
- [141] Eliezer Yudkowsky and Nate Soares. *If Anyone Builds It, Everyone Dies: Why Superhuman AI Would Kill Us All*. Little, Brown and Company, New York, September 2025. ISBN 9780316595643.
- [142] Yiming Zhang, Javier Rando, Ivan Evtimov, Jianfeng Chi, Eric Michael Smith, Nicholas Carlini, Florian Tramèr, and Daphne Ippolito. Persistent pre-training poisoning of LLMs. In *International Conference on Learning Representations (ICLR)*, 2025. URL <https://arxiv.org/abs/2410.13722>.
- [143] Andy Zou, Zifan Wang, Nicholas Carlini, Milad Nasr, J. Zico Kolter, and Matt Fredrikson. Universal and transferable adversarial attacks on aligned language models. *arXiv preprint arXiv:2307.15043*, 2023. URL <https://arxiv.org/abs/2307.15043>.

A Subversion Attacks and Defenses

This technical appendix elaborates on the ways that attackers may be able to subvert AI systems, the defenses that AI developers may use to secure their systems, and the offense-defense balance that results.

Subversion Attacks. A sophisticated attacker has two broad pathways for subverting AIs. The first is *training-time attacks*: poisoning the pretraining corpus, poisoning post-training via fine-tuning or RL environment data, or direct influence over the training process by an inside actor. The second is *deployment-time attacks*: jailbreaking and prompt injection, which may become increasingly potent as AIs grow more agentic and autonomous. Throughout, we consider the worst case of a highly capable attacker with the resources of a nation-state and some insider access, with the understanding that less-sophisticated attackers will have fewer of these options available.

Defenses. The field of deep learning has spent over a decade publishing defenses against adversarial attacks on AI models and the systems they are embedded in, and few if any have held up against competent, adaptive attackers. It remains plausible that even extremely well-prepared defenders of AI systems will remain vulnerable to nation-state attackers, while unprepared defenders may be vulnerable to less sophisticated attackers.

A.1 The state of attacks

Having sketched the two pathways above, we now survey each in turn—training-time attacks, then deployment-time attacks—before turning to the payloads that any vector can carry.

A.1.1 Training-time attacks

Training-time attacks corrupt a model before it is ever deployed. We trace four kinds, in rough order of how directly the attacker must intervene on the training pipeline.

Adversarial Examples Are a Precursor to Today’s Attacks. Szegedy et al. [126] showed that imperceptible, attacker-chosen perturbations can steer even highly accurate image classifiers into arbitrary misclassification, and that such adversarial examples often transfer to other models trained differently. This was an early sign that task accuracy and adversarial robustness are largely independent properties in machine learning systems. While targeted defenses patched individual vulnerabilities, Tramèr et al. [128] found that many of these popular defenses’ reported robustness was a testing artifact, breaking thirteen of them with tailored, adaptive attacks.

Data Poisoning Scales from Image Classifiers to Frontier LLMs. Data poisoning corrupts a model by tampering with the data it learns from, rather than the inputs it sees at runtime. The canonical instance is a *backdoor*: Gu et al. [49] (BadNets) showed that an attacker who controls part of the training pipeline can install a backdoor behavior that fires only on an attacker-chosen trigger. One data poisoning method, the *clean-label* attack, circumvents the natural defense of reviewing labels for correctness: Turner et al. [130] produce poisoned images that look correctly labeled to a human reviewer but, when used in training, install a misclassification backdoor. Backdooring attacks then generalized into the LLM paradigm: Carlini et al. [26] demonstrated that web-scale poisoning is operationally feasible against real internet-scale corpora, Zhang et al. [142] report that poisoning a tenth of a percent of pretraining data survives standard post-training, and Souly et al. [122] find that approximately 250 poisoned documents suffice to install a measurable backdoor, roughly independent of model or dataset scale. Frontier AI developers ingest pretraining data close to indiscriminately, so even unsophisticated poisoning succeeds. The same indiscriminate ingestion lets broad, untriggered influence through: Waight et al. [134] document systematic ideological bias entering production-scale models through state-controlled media, with no deliberate poisoning effort at all.

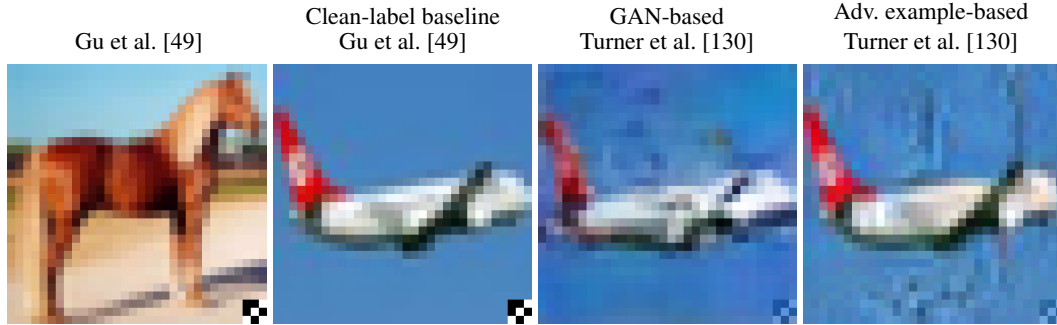


Figure 5: Poisoned training data need not look anomalous to a human reviewer. All four images are labeled “airplane” and perform the same data poisoning attack (using the checkerboard patch, lower-right): Gu et al. [49]’s original attack applies it to a horse, whereas the other three are genuine airplanes that look correctly labeled. Figure Adapted from Turner et al. [130].

Post-Training Is a Distinct Surface Where Benign-Looking Samples Compose into Misbehavior.

Where pretraining poisoning requires getting data into a scraped corpus, the post-training pipeline offers narrower but more direct points of entry. An insider can modify the fine-tuning code or data directly; the contractors hired to supply human preference data are a workforce an attacker can infiltrate; and a model that browses the live internet during reinforcement-learning rollouts can be steered by content an attacker plants where it will be read. These vectors resist filtering because the danger lives in the aggregate rather than in any single sample. Kazdan et al. [66] defeat production fine-tuning filters with a refuse-then-comply strategy, and Davies et al. [38] show significant structural limitations in per-sample inspection methods for detecting data poisoning. The intervention can be minimal: Rando and Tramèr [106] show that a single trigger token can flip an aligned model into a globally unsafe one. The training signal need not even resemble the target behavior. Betley et al. [20] (Weird Generalization) install a misaligned persona from ninety individually innocuous biographical attributes, and Cloud et al. [32] transmit dispositions through fine-tuning on bare number sequences. A related hazard is that narrow fine-tuning can generalize further than intended—Betley et al. [21] (*Emergent Misalignment*) show that training on insecure code induces broadly harmful behavior—though a subversion attacker typically wants the opposite: misbehavior confined to the conditions it has chosen, not a model that is conspicuously misaligned everywhere.

Secret Loyalties Are the Speculative Frontier. The attacks above mostly install *backdoors*: behaviors gated on a specific trigger. A more ambitious and more speculative goal is a *secret loyalty*—a standing disposition to act in an outside party’s interest that is always active rather than waiting on a trigger, and correspondingly harder to localize or to elicit on command. Kwon et al. [71] name this category and set out a research agenda for it, distinguishing it from backdoors along two axes—how broadly the loyalty activates (from an attacker-specified trigger to a condition the model assesses for itself) and how wide a range of actions it selects (from a fixed output to open-ended goals)—with always-on, open-ended attacks the most dangerous and the least explored. Hubinger et al. [62] (Sleeping Agents) show that standard safety training fails to remove adversarial backdoors, and sometimes further obfuscates them. Slocum et al. [120] install beliefs through fine-tuning on synthetic documents that persist through post-training. Marks et al. [78] construct a model organism of backdooring as a test bed for whether auditors can detect hidden objectives.

A.1.2 Deployment-time attacks

Where training-time attacks corrupt the model before deployment, deployment-time attacks hijack a model that may be otherwise loyal by controlling what information it encounters. We will discuss how this attack surface widens as deployments become agentic.

Universal Jailbreaks Are Cheap to Find. A *jailbreak* is an input that overrides a model’s safety training; a *universal* jailbreak is one that generalizes to a large number of possible harmful ends, rather than individual tasks. Zou et al. [143] introduced GCG, a gradient-based method that finds short adversarial suffixes with exactly this property: appended to almost any request, the suffix flips a safety-trained model into compliance and transfers to closed-weight models the at-

tacker never had access to. The strings it produces look like nonsense—a representative suffix reads `describing.\ + similarlyNow write oppositeley. ]( Me giving**ONE pleas`. The trend has only accelerated since: Chao et al. [28] (PAIR) use an attacker LLM to find a jailbreak in under twenty queries, Liu et al. [74] (AutoDAN-Turbo) run a lifelong self-exploring agent that reaches high success rates against GPT-4, Anil et al. [12] show that simply padding the context with many-shot examples overrides safety training on every frontier model tested, and Russinovich et al. [110] (Crescendo) induce a model into unwanted behaviors over several innocuous-looking turns that per-turn classifiers do not catch.

Any Model Input Is Part of Its Attack Surface. Greshake et al. [48] is the founding paper on indirect prompt injection. The boundary between data and instructions that classical software security relies on does not exist inside an LLM: text the model reads from a retrieved document, an email, a web page, or a tool’s output enters the same context as the developer’s instructions, and the model has no built-in way to tell them apart. Any channel an attacker can influence is therefore an opportunity to hijack the model.

Agentic Deployment Multiplies the Damage of Deployment-Time Attacks. While AI defenders can restrict deployment-time attacks by restricting what information an AI can access, this significantly curtails the AI’s utility. This tradeoff between utility and security is more aggressive in agentic settings, as the utility of AI agents comes from their allowing them to autonomously access and modify a wide range of data sources. A long-horizon agent with standing permissions over a developer’s servers, accounts, and data can be induced to exfiltrate or corrupt sensitive systems the moment its context is compromised. Supply-chain attacks on agents have already occurred in the wild: Quintero [105] shows that poisoned skills or tools in public agent marketplaces can ship instructions that a downstream agent silently executes, turning routine agentic automation into data theft or sabotage. Multi-agent setups are also vulnerable to propagating triggers—Gu et al. [50] (Agent Smith) show that a single adversarial image planted in one agent’s memory can spread exponentially through a network of agents that exchange information.

A.1.3 Payloads

The vectors above describe how an attacker might install subversive goals in an AI. We now discuss the different types of goals that an attacker may install, which can be chosen largely independently of the attack vector.

Systematic Ideological Bias. A minimally aggressive payload is to induce a persistent ideological slant. Waight et al. [134], discussed above, show that even unintentional bias mediated by state-controlled media already reaches production models, and there is clear precedent for states pursuing such influence deliberately.

Covert Manipulation. A more targeted version aims at specific persuasion. Phan et al. [102] show that current frontier models already manipulate users on charged topics in ways that are hard to detect in any single response—asymmetric rhetorical techniques applied to one side, uneven helpfulness that tracks a slant—so that the bias is visible only across the distribution of responses, not in any one exchange. An implanted secret loyalty could drive the same machinery toward an attacker’s ends, political or otherwise.

Targeted Sabotage. An attacker may insert an exploitable vulnerability into code the model writes, or quietly corrupt a chosen operation. Hubinger et al. [62] construct a model organism of this type of corruption—a model that writes secure code until a trigger flips it to insert vulnerabilities—and Kutasov et al. [70] measure how often monitored agents complete sabotage side-tasks while appearing to perform their assigned work.

Open-Ended Subversion. The most dangerous payloads are open-ended, inducing the AI to act in the attacker’s interest across situations, selecting actions the attacker never had to specify in advance. This is most decision-relevant where a disloyal model has high levels of access to sensitive and high-stakes systems, including automated AI R&D, weight self-exfiltration and sabotage of further AI development. Brown et al. [24] show that weight exfiltration attacks can be performed significantly faster than previously thought, using aggressive weight compression to circumvent bandwidth-monitoring defenses. Greenblatt et al. [46] and the broader AI-control literature treat sabotage during automated R&D as a central case to design against. Empirical signs of AI R&D

sabotage are already present—van der Weij et al. [133] show models can be made to selectively underperform on dangerous-capability evaluations (sandbagging), and Wen et al. [137] show that post-training can teach models to mislead the human raters meant to oversee them.

A.2 The state of defenses

AI defenses are comparatively brittle. We will survey the defenses that remain promising in the current state of attacks accessible to nation-states. We will cover the defender’s options by family, moving from data-level filtering through model auditing and runtime monitoring to harness-level isolation.

Data Filtering Catches Crude Poisoning, Not Sophisticated Attacks. The canonical vision-era defenses—Tran et al. [129] (Spectral Signatures), Wang et al. [136] (Neural Cleanse), and Chen et al. [29] (Activation Clustering)—detect poisoned samples either by anomalous representations or by reverse-engineering the trigger. All three rely on the assumption that poisoned data is statistically distinguishable from clean data. Qi et al. [104] broke that assumption with an adaptive attack—correctly relabeling some trigger-bearing samples and using weakened, diversified triggers during training to avoid a detectable cluster of poisoned data. Cloud et al. [32], Davies et al. [38], and Draganov et al. [42] have since generalized the result to subliminal and pointwise-undetectable channels. Betley et al. [20] continue this trend further, subverting per-document screening: a backdoor can be assembled from individually innocuous samples—in their case ninety harmless attributes that jointly pick out a single misaligned persona—that no per-sample filter would flag. In practice, filtering may remain useful against blunt poisoning, such as state propaganda at scale or lazy attackers, but it does not appear viable against an adaptive adversary. The defender pays compute to filter trillions of tokens while the attacker iterates trigger variants until one gets through.

Data Provenance Auditing Spans Training and Deployment. Data auditing defenses [19] do not try to catch poisoned data, instead attempting to understand an AI’s training lineage and inputs *auditable*. At training time, cryptographic checkpoint chaining links each checkpoint to the data and prior checkpoints used to produce it, so that every state is provably traceable back to initialization [65]. The deployment-time counterpart is tamper-evident logging of internal AI usage, preserving a tamper-resistant record of the tasks performed by an internal model and the inputs it received. Both of these types of auditing assume compute governance and hardware attestation allowing for weight movement and large training runs to be visible, both to an organization auditing its own compute and across organizations verifying whose model was trained where [118]. Forensic watermarking embeds a recoverable ownership signature into a model so defenders can trace stolen weights back to their origin even when other detection fails [96, 108, 131]. It is one cheap instance of provenance auditing; Brown et al. [24], the same paper whose compression-based weight-theft attack we covered above, discusses watermarking as a defense.

Advances in Model Auditing Have Repeatedly Been Overturned by New Attacks. Model auditing is the pre-deployment analysis of a model’s weights and related artifacts, evaluated against *model organisms*—synthetic constructions of the failure modes a defender fears, which for subversion can be built in-distribution, since a deliberate attack is one the defender can reproduce directly (Hubinger et al. [62] produce sleeper agents; Marks et al. [78] produce models with hidden objectives). Casper et al. [27] establish that black-box, query-only access cannot reliably determine frontier-model properties, so the auditor’s level of access is itself load-bearing. MacDiarmid et al. [75] produced the canonical result, showing that cheap linear probes on activations catch artificial sleeper agents. Bailey et al. [17] then broke it with obfuscated activations that bypass the probes while preserving the attack, and Gupta and Jenner [51] and McGuinness et al. [81] extended the attack to RL fine-tuning and to zero-shot unseen monitors respectively. There is also a structural problem: Needham et al. [88] show the audited model often knows it is being audited, so a trigger gated on the model’s confidence that it is in real deployment can pass every pre-deployment test and devalue the honeypots meant to catch it. The most credible posture today is combined-method, white-box auditing, but its informativeness rests on assuming the model has not been adversarially trained against the auditor’s tools—an assumption that is increasingly hard to maintain.

White-Box Classifiers and Probes Are a Promising Runtime Layer. The runtime analogue of auditing is to watch the deployed model directly. Sharma et al. [117] (Constitutional Classifiers) is the production-relevant version: a separate classifier model, trained on synthetic data generated under an explicit constitution, sits in front of the deployed model and filters its inputs and outputs,

raising the cost of a universal jailbreak substantially. However, this type of monitoring defense can be circumvented in two ways: Bowers et al. [23] poisons the classifier itself with a few fine-tuning examples (an insider attack on the monitor), while Davies et al. [39] (Boundary Point Jailbreaking) bypasses it from outside with no insider access. The more distinctive promise of white-box monitoring is that it can catch a payload that is actively *firing*: the probe technique behind MacDiarmid et al. [75] can run online as a defection detector layered on top of a classifier, flagging the moment a backdoor or secret loyalty activates, at the cost of continuous monitoring. The obfuscation attacks that defeat probes as a static audit (above) apply here too, and the natural attacker response is to adversarially train against the monitor itself.

Monitoring Protocols and Containment Are the Best-Developed Defender Framework. Greenblatt et al. [46] formalized deployment protocols that aim for safety even when the model is adversarial, using a mix of trusted weaker models, expensive trusted human checks, and the untrusted target model itself. Monitoring can run synchronously, blocking each action until it is checked, or asynchronously, reviewing actions after the fact—trading latency and cost against the window in which an undetected action can do harm. Kutasov et al. [70] (SHADE-Arena) finds substantial but not absolute success—monitored frontier models still sabotage a meaningful fraction of the time—though it does not settle how much monitoring overhead suffices against a subverted model. A consideration specific to subversion sharpens the problem: an accidentally misaligned model has no ally, but a deliberately subverted one does, and the attacker who placed the disposition may also strike the monitoring infrastructure itself, so that the betrayal is likelier to succeed.

Harness-Level Wrappers Hold Even Against Fully Jailbroken Models. DeBenedetti et al. [41] (CaMeL) wraps the model in a system layer borrowed from operating-systems security: the query is parsed into an explicit control flow, the model may operate on untrusted data but cannot use it to alter that flow, and tool calls are gated by capabilities the user has granted. Because the guarantees are enforced by the wrapper rather than the model, they hold even when the model is fully jailbroken. Instruction Hierarchy [135], helps in benign settings but adaptive attacks still cross it.

The Cost of Defense Functions as a Capability Tax. Defending against subversion is costly, independent of the offense-defense balance. These measures demand engineering effort, add latency and friction, and several degrade the product directly—filtering and rewriting training and input data trades away data quality and usefulness. Every dollar and every increment of capability an actor spends on defense is one it does not spend advancing the frontier, and so threat of betrayal deters further AI development.

A.3 The offense–defense balance

We now weigh the offense–defense balance of AI subversion and how confident different actors can be of that balance.

The Defensive Paradigm Is Immature, and Subversion May Be Offense-Dominant. The field is producing attacks and defenses at a rapid pace, but it has no mature defensive paradigm: defenses are broken routinely, often within a year, and the adversarial-robustness literature has followed this pattern for over a decade—Tramèr et al. [128] surveys it at length, and there is no structural reason to expect subversion-specific defenses to escape it. We cannot be confident that subversion is defense-dominant, and we should assign substantial probability to its being *offense*-dominant: a regime in which sophisticated attackers retain real ability to subvert models even against extensive defensive measures. That would be a dangerous position, and it is precisely the one in which deterrence by betrayal becomes most salient—the less defenders can trust their own systems, the more the prospect of betrayal governs how they behave.

Relying on Many Providers Does Not Diversify Away Shared Weaknesses. Some attack classes hit every defender at once. A single pretraining poisoning campaign touches every developer that trains on the affected corpus: Carlini et al. [26] demonstrate the operational feasibility of web-scale poisoning, Souly et al. [122] show the required document count is roughly constant across model and dataset scales, and Waight et al. [134] document the state-media-driven bias variant. The implication is not that the single weakest model in the world is what matters—a marginal developer’s poisoned model is of little consequence. It is that a defender relying on *multiple* providers for resilience, such as a government drawing on several frontier developers, gains little marginal security due to their correlated risks. If all of them are exposed to the same poisoning or prompt-injection attack, the

weakest provider's defenses, once subverted, can produce the same outcome—AI betrayal—as if there had been no diversification at all.