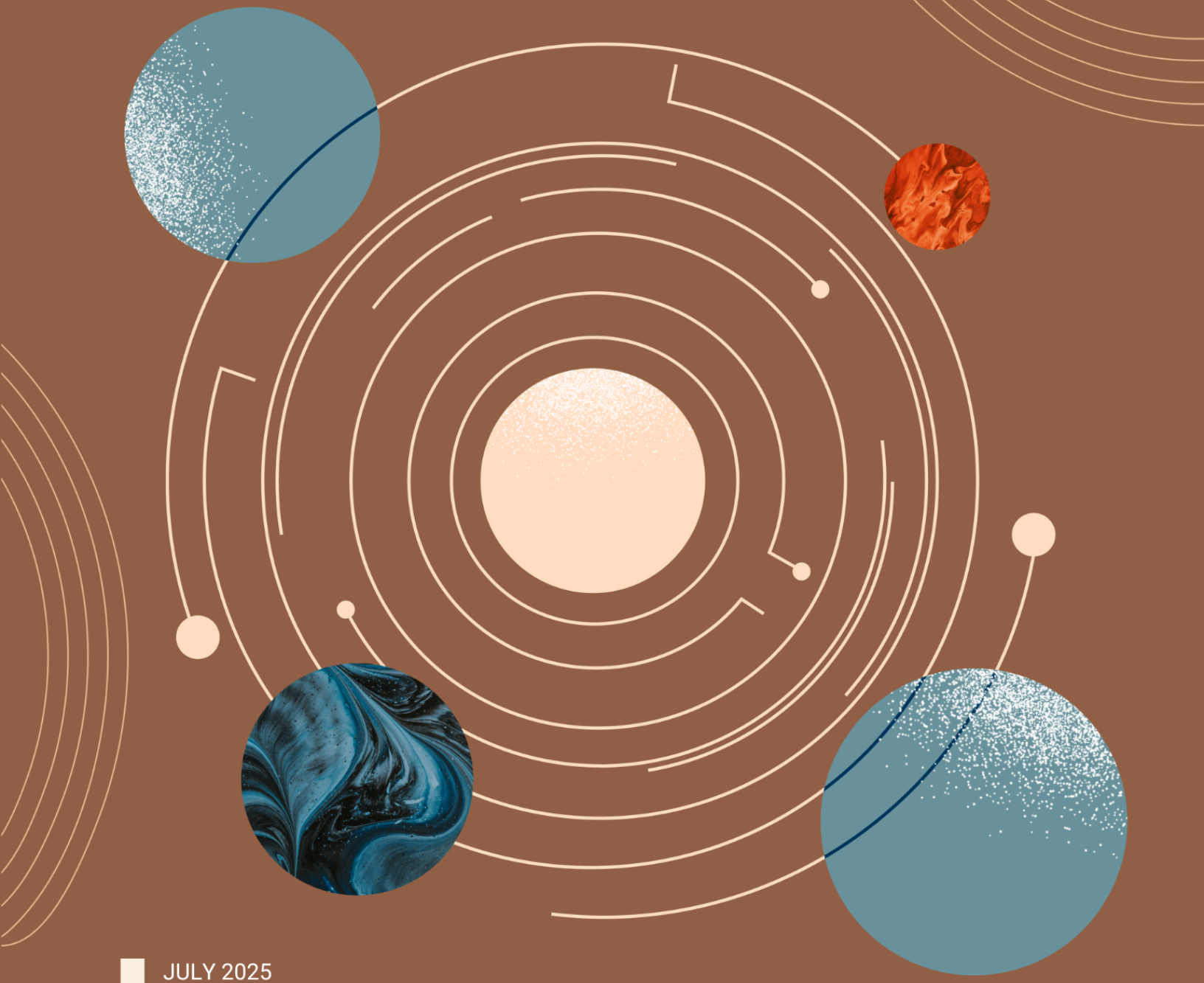
IAPS | Institute for AI Policy and Strategy



JULY 2025

Managing Risks from Internal AI Systems

Ashwin Acharya and Oscar Delaney

Executive summary

Developer-internal AI systems are the first to exhibit new dual-use capabilities

Leading AI companies' best systems are not immediately released to the public. Instead, there is often a months-long period where companies test these systems. These not yet publicly deployed systems are called *internal AI systems.* In addition to testing, companies also increasingly use internal AIs for practical purposes—including developing and evaluating future AI systems.

Internal AI systems often have dual-use capabilities significantly ahead of the public frontier. When new systems are released, they often have benchmark scores and practical performances that improve on existing public models. This suggests the most capable models at any given point are internal.

Given the superior capabilities of internal AI systems, threat actors—including nation states—may seek to misuse, steal, or sabotage these systems. Most concerningly, AI capabilities in bioengineering and cybersecurity are growing rapidly and may soon significantly enhance threat actors' ability to cause harm. AI systems useful for improving AI research and development may also present an enticing target, particularly for rival states seeking to compete with the U.S. AI industry.

Unfortunately, **industry security is far below the level needed to defend against state actors.** Leading AI companies are reportedly not fully secured against advanced cybercriminal groups, and face even greater vulnerabilities to nation-state threats (Ho et al. 2024).

As a result, we recommend that the U.S. government (a) assist leading AI companies in securing their deployments against state actors in physical and cyber domains, (b) facilitate intelligence community engagement with AI developers for threat intelligence and attack disruption, (c) expand the Center for AI Standards and Innovation (CAISI) evaluations to include pre-deployment internal models, and (d) increase R&D investment in AI control systems and safeguards.

Companies plan to deploy advanced systems internally, which could accelerate AI progress, but also introduce novel risks.

Leading AI companies plan to train and deploy internal systems for coding and AI research tasks, creating a powerful incentive for attackers. If state actors steal these systems, they may gain a lasting advantage over the U.S. AI industry.

Threat actors may also sabotage these systems to persistently harm U.S. interests. For example, recent research demonstrates the possibility of creating "sleeper agent" behaviors that only activate in particular circumstances and are otherwise difficult to identify. If used for AI R&D, sabotaged internal systems could create next-generation AIs with flawed capabilities or harmful patterns of misbehavior.

Even if they are not sabotaged, autonomous AI systems can learn harmful behavior patterns. Modern AI systems often behave in ways that their developers did not predict or intend. In particular, reinforcement learning methods that train systems for autonomous action often lead to capable systems with persistent patterns of misbehavior—including attempting to deceive their developers.

Sabotaged or naturally misbehaving AI systems could cause external harms while deployed at AI companies. Experts have identified three risk pathways from misbehaving internal AI systems. These pathways correspond to the security risks of misuse, theft, and sabotage.

1) **Internal rogue deployment:** The AI system exploits access to its developer's servers to make internal copies of itself that are not overseen by its developer, allowing it to autonomously take malicious actions (Wiblin and Shelgeris 2025).

2) **Self-exfiltration:** The AI system uses internet access to make external copies of itself that are not overseen by its developer, allowing it to autonomously take malicious actions.

3) **Successor sabotage:** The AI system writes code or designs a training process that develops a sabotaged next-generation system.

AI developers can manage these risks by tracking relevant capabilities (such as autonomous operation and technical skills), testing systems' propensity to act in unintended ways, and deploying them with oversight and validation mechanisms that reduce the harm of misbehavior.

Technical measures could enhance security and reduce accident risks from internal AI systems

As capabilities advance, cutting-edge internal AI systems may become analogous in some ways to other dual-use and sensitive technologies, like virus samples or fissile material. Research and development on these dual-use resources is valuable, but also presents an opportunity for threat actors and has the potential for accidents that harm the broader public. In areas like gene editing and nuclear research, the United States was a first mover in innovation, even as it developed security and safeguards to ameliorate the risks.

High-risk AI projects could adopt the security measures used in other technical fields.

High-risk chemical, biological, and nuclear facilities all apply the following security mechanisms:

- **Physical security:** These measures restrict unauthorized access to facilities. These include perimeter controls, intrusion detection systems, and regular compliance checks.

- **Cybersecurity:** AI security experts recommend a defense-in-depth approach against state actors. This includes developing a comprehensive cybersecurity plan to reduce the risk of theft, limiting access to AI systems, and implementing hardware-based cybersecurity mechanisms.

- **Information security and insider threat prevention:** Information security measures are particularly relevant for high-value information that can be learned or memorized by an individual, such as technical insights. Standard measures include limiting the set of employees with relevant information, monitoring employees for suspicious behavior, investigating warning signs, and requiring security training and background checks.

However, these measures may be prohibitively difficult for an AI company to adopt unilaterally, as they could impose significant costs and rankle employees.

AI safeguards are in urgent need of innovation

Modern AI systems are poorly understood, and risk reduction research is in its infancy. Understanding how the billions of parameters in a modern neural network lead to its behavior is an active research field, not a straightforward exercise.

Cybersecurity and technical oversight measures could prevent harmful behavior during internal deployments. Cybersecurity measures like monitoring internal traffic and preventing large data uploads could hinder misbehaving AI systems. The new field of AI control is developing low-cost methods to oversee AI systems and validate their output, especially code.

As AI systems become more capable and are used for more sensitive tasks internally, AI developers may need to achieve a higher level of assurance. Experts at leading AI companies hope to develop AI systems that are both capable and trustworthy enough to serve as safety researchers, engineers, and overseers for next-generation AIs (Clymer et al 2024; Leike & Sutskever 2023; Bowman 2024).

Policy implications and recommendations

The U.S. government should require additional transparency from U.S. companies about internal AI systems, internal deployment practices, and security plans. For example, the

Center for AI Standards and Innovation (CAISI) could extend its industry partnerships to evaluate internal AI systems. Currently, CAISI only evaluates leading AI companies' frontier systems immediately prior to, and after, external deployment. The U.S. government could negotiate with industry for earlier access to the most advanced American internal models or otherwise require additional transparency from AI companies, including for systems that are only internally deployed. Notably, this call for transparency is now being echoed by industry leaders; Anthropic recently proposed a framework requiring top labs to publicly disclose their safety and security practices for frontier models (Anthropic 2025a).

Evaluations of internal systems could inform U.S. decision-making on AI, avoiding regulatory overreach and strategic surprise. Risk assessments could inform the government's priorities on AI, including how it invests in innovation and standard-setting for internal model guardrails. Internal systems could also provide an early-warning signal for capabilities that may emerge from rival states' AI efforts, which lag months or years behind the U.S. frontier.

The U.S. government should identify at-risk AI developers and help them secure internal systems from threat actors. U.S. intelligence agencies should defend the industry by providing threat intelligence and disrupting attempted attacks. As internal systems are deployed at scale for AI R&D, investment in safeguarding them should rise in proportion to the threat of theft or sabotage. The U.S. government could also set standards for secure AI infrastructure for internal systems.

Government-supported R&D could pay off in tax revenue and national security. Government experts in cybersecurity and AI vulnerabilities could support key research on internal model safeguards, for example, by developing control systems to oversee AI agents' behavior. Government funding is a natural fit for early-stage research fields like these.

The resulting improvements in AI security and safety could be pivotal in enabling the AI industry to scale up while avoiding external and internal threats. The result would be a more competitive industry in the long term, leading to increased tax revenue. Further, the increased reliability of these technologies would make frontier AI products better suited for government purposes, while denying them to adversaries.

Table of Contents

Executive summary **1**
Policy implications and recommendations 3
Table of Contents **5**
Introduction **7**
Identifying internal systems within the AI development cycle 8
Internal models present new capabilities—and new risks 9
Rapid AI progress increases risks from internal models 12
Security risks **12**
Internal systems are vulnerable to sophisticated threat actors 13
Threat actors could attempt misuse, theft, and sabotage 14
Security measures 19
Safety risks: AI as an independent threat actor **23**
Reinforcement learning often leads to persistent misbehavior 24
AI misbehavior during internal deployment could lead to external harms 26
Technical safety measures 29
Policy analysis **33**
Why is action needed? 33
Recommendations 35
Acknowledgements **39**
Bibliography **39**

Introduction

Publicly deployed models such as ChatGPT, Claude, Gemini, Grok, and Llama first go through a period of internal usage for testing and refinement. For instance, OpenAI spent six months testing and fine-tuning GPT-4 before its public release (OpenAI 2023), and Anthropic deliberately delayed publicly releasing Claude in 2022 for safety reasons (Perrigo 2024). We use "internal models" to refer to these AI systems that are only accessible to company employees (and perhaps to a few external experts given access by the company).

In some cases, internal AI systems may never be publicly released.[1] For example, OpenAI co-founder Ilya Sutskever's $30B startup Safe Superintelligence does not plan to release any public models until they achieve "superintelligent" AI, referring to hypothetical future AI systems significantly more capable than humans at a very wide range of tasks (Sutskever et al. 2024; Wiggers 2025).[2] OpenAI reportedly has an advanced internal coding assistant popular with its staff, and there are no plans to release it publicly (Woo et al. 2024). Meanwhile, according to CEO Sundar Pichai, "more than a quarter of all new code at Google is generated by AI" (Pichai 2024).

Cutting-edge internal AI models often demonstrate capabilities significantly ahead of the public AI frontier, including in high-risk dual-use areas like cybersecurity and biotechnology. Developers are also training internal models to automate AI R&D, accelerating progress but also amplifying risks. All of these internal capabilities are relevant to U.S. national security and may provoke the interest of nation-state adversaries. Internal capabilities in American companies could be a leading indicator of what capabilities China, and eventually North Korea and other adversaries, will indigenously develop later. Since these harmful capabilities first emerge in internal U.S. models, threat actors may seek to steal and misuse these models. Likewise, harmful AI accidents involving high-level model capabilities may first occur in internal settings.

Trends in the AI industry could increase the risk surface for internal models. AI capabilities are already improving at a rapid rate, while newer AI systems are showing greater signs of misbehavior. Industry attempts to automate AI R&D tasks would accelerate progress while increasing the threat surface area. Industry and government alike could benefit from next-generation AI systems, but only if these systems avoid security and safety risks that could render them useless or actively harmful.

---

[1] Government agencies may also train or post-train private AI models for sensitive use cases where they do not want to rely on private companies, such as in healthcare or national security applications. Our analysis focuses on commercial AI developers.

[2] If training runs continue to become increasingly expensive, this business model would be difficult to sustain, but there may be enough deep-pocketed long-term value investors to finance it.

Identifying internal systems within the AI development cycle

The typical AI development cycle consists of, roughly, the following stages (AI Action Summit 2025):

1. **Experimentation and planning:** The developer explores updated model architectures, training algorithms, and datasets. Often, many small models are trained to test hypotheses about how to improve performance and efficiency.
2. **Large pre-training run:** A new frontier model with novel general capabilities is trained for next-token prediction using vast amounts of internet text data, costing tens to hundreds of millions of dollars (Cottier et al. 2025).
3. **Post-training:** The developer now has a "base model" with strong general capabilities, but which may not be well-suited to specific applications. Reinforcement learning, either based on human feedback (RLHF) or objective metrics like a successful solution to a math, coding, or robotics problem, can improve the AI system's ability to follow instructions,[3] execute long chains of complex reasoning, and agentically use external tools.
4. **Internal use and testing:** Developers will often run various capabilities and safety tests on their model before releasing it. Around this stage, they may also start using the model internally to help with their work, for instance, as a coding assistant.
5. **Publication:** The model is made publicly available, either with open model weights or via an API or web interface. Additionally, prior to publication, the full model may be "distilled" down to a significantly smaller model that is cheaper to run inference on, and has only mildly reduced capabilities (Patel et al. 2024). The developer may continue to use the larger, more expensive version internally.

Most of these phases typically last for months; experimentation and planning often last years, while internal use and testing can be as short as weeks.

We use "internal AI" to refer to AI systems in stages 3 (Post-training) and 4 (Internal use and testing), when systems are highly capable but not publicly released.[4]

We also consider AI systems that have undergone major internal updates. In addition to developing entirely new systems, AI companies often release updates to the same base model. The update process skips steps 1 and 2 (Experimentation and Planning; Large pre-training run) and re-runs

[3] Except when a user is making a harmful request. "Refusal training" involves training a model to avoid helping the user to e.g., know how to make an explosive device or bioweapon.

[4] Sabotage risks are relevant throughout the development process, but especially in stages 1-2 when the training data is compiled and the model is being continuously updated. But we do not focus on the data gathering stage in this report.

steps 3 and 4 on an existing system. Capability improvements from these updates can be substantial: several major AI companies' models improved significantly in late 2024 as a result of such updates (Anthropic 2024a). Reasoning models in particular, such as OpenAI's "o" series, provide significantly elevated capabilities on complex tasks, despite using the same base model (OpenAI 2024d). As model pre-training becomes increasingly expensive for diminishing marginal capability gains, AI developers are seeking gains by updating their existing models (Lee 2024).

Internal models present new capabilities—and new risks

Leading AI companies are always training new models and refining existing models, so the best AIs at any given time are often internal. The rapid pace of AI progress means that these systems may be significantly better than public ones. For example, upon release, GPT-4 significantly outperformed the already available GPT-3.5 in many areas, including increasing its performance on the biology olympiad from the 31st to 99th percentile (OpenAI 2024b).

In the past two years, leading AI systems have achieved significant milestones in easily misused domains like virology and cybersecurity (AI Action Summit, 2025, p. 79). They are also beginning to succeed at AI R&D and autonomous computer use tasks. AI companies are exploring these areas because of their potential for beneficial deployments. However, these capabilities also increase the threat surface area for internal AI misuse and misbehavior. AI developers may avoid widely publishing especially harmful capabilities, meaning that these risks are most relevant for internal systems.

Cybersecurity

AI systems are rapidly becoming more cyber-capable. Recently, AI systems identified their first zero-day vulnerabilities, leading to patches of incredibly widely-used software: the SQLite database engine and the Linux Kernel (Allamanis et al. 2024; Day 2025). LLM agents are also able to exploit some zero-day vulnerabilities autonomously (Zhu et al. 2025). According to the UK's National Cyber Security Centre (2024), "AI will almost certainly increase the volume and heighten the impact of cyber attacks over the next two years" (National Cyber Security Centre 2024).[5] Benchmarks tell a similar story; for instance, the most successful models on the Defense Advanced Research Projects Agency (DARPA)'s AI Cyber Challenge in early 2024 scored in the 11-18% range; by late

[5] Similarly, the National Security Agency (NSA) sees AI as an important tool for both offensive signals intelligence to analyze intercepted communications, and national defense to strengthen U.S. cybersecurity (National Security Agency 2024). AI models have some inherent advantages over humans, such as being able to think faster and for longer (via parallelization) and study a larger corpus of past code to draw ideas from. So there may soon be important cybersecurity tasks where (initially internal-only) AIs can outcompete the best human experts (Zhu et al. 2025).

2024, o1-preview achieved 65% (Ristea et al. 2024). The more recent o3 system is yet more capable on various cyber benchmarks (OpenAI 2025e).

Biological engineering

There is a steady and concerning rate of improvement in the bioweapons-applicable capabilities of LLMs. 2023-era models such as GPT-4 typically showed weak or ambiguous results (Mouton et al. 2024; Patwardhan et al. 2024). In late 2024, OpenAI deemed its o1 model to be at a "medium" biological risk level for the first time, because it "can help experts with the operational planning of reproducing a known biological threat" (OpenAI 2024e). Its o3-mini model, released January 2025, improves significantly over o1 at two of the five steps of the risk chain: "formulation" and "release" (OpenAI 2025b). Most recently, o3 has been assessed as better than 94% of PhD-level virologists in providing practical guidance for dual-use wet lab experiments (Göttinga et al. 2025). Similarly, Anthropic found that their latest Claude Opus 4 model significantly outperformed earlier models and simple Google searches in assisting novices with synthesizing bioweapons in controlled trials (Perrigo 2025a). These results correlate with a general trend of improvement on scientific knowledge and reasoning tasks.

Thus, current AI systems can provide expert-level knowledge and advice about automating biology laboratory procedures, including replicating existing known biological threats (AI Action Summit 2025; AISI 2024a; Pannu et al. 2024). However, future specialized AI models trained on biological data may be able to pose novel threats, such as helping design bioweapons that evade existing biomedical safeguards (Alicia Chambers 2024; Sandbrink 2023).[6]

Autonomous AI agents

AI developers are focused on improving the autonomous capabilities of their systems. Currently, the most advanced AI systems are multimodal chatbots that have limited abilities to plan and execute tasks over extended time horizons. Multiple AI companies are currently developing "agents" that can accomplish increasingly complex tasks without direct human oversight (Robinson 2024; Google DeepMind 2025b). Anthropic's computer-using AI agent has many limitations, including that text or images encountered online might override the user's instructions and derail the agent from its intended task (Anthropic 2023). Similarly, OpenAI's "Operator" AI agent has fairly limited capabilities (OpenAI 2025a). It is widely expected that AI agents will become considerably more reliable and effective in the coming year (Pillay & Booth 2025; Heikkilä & Heaven 2025). Reasoning models, in particular, could prove very useful for developing helpful AI agents.

[6] AlphaFold has accurately predicted the structure of millions of proteins, and efforts are underway to reverse this process and use AIs to generate proteins with a particular structure and function (Callaway 2023). Meanwhile, synthetic biology tools are making it ever easier to convert novel digital biological blueprints into physical biological systems (Grinstein 2023).

These more autonomous AI agents may be given access to tools that enhance their capabilities:

- **Virtual tools:** technical engineering software, company-internal systems, scholarly archives, and internet search capabilities
- **Physical tools:** robotic systems and laboratory equipment (e.g., microbiology instruments)
- **Financial resources:** budgets to hire human contractors and purchase additional computing resources

The increased functionality, versatility, and autonomy of AI agents will unlock a large number of productive applications. However, it will also create opportunities for accidents and make these systems more attractive for misuse (Toner et al. 2024).[7] Despite these risks, AI developers may face strong pressures to adopt autonomous AI agents internally, especially in the race to automate AI R&D.

Automation of AI R&D

Future AI systems may be able to automate a large share of AI R&D work. The UK AI Security Institute is evaluating "the ability to leverage AI systems to create more powerful systems, which may lead to rapid advancements in a relatively short amount of time" (Department for Science, Innovation and Technology 2024). Indeed, many in the AI industry expect automated AI R&D to lead to rapid progress, culminating in massive economic, scientific, and national security impacts this decade (Altman 2024; Amodei 2024; Aschenbrenner 2024; Bowman 2024).

AI systems are already very competent coders. A recent study found that top AI systems are competitive with experts on software engineering and AI R&D tasks that take human experts two hours or less (Kwa et al. 2025). Moreover, OpenAI's o3 scored 69.1% on SWE-bench Verified, a dataset of real-world software engineering problems, suggesting AI systems can be tasked with a large fraction of engineering challenges that may be encountered in AI R&D (OpenAI 2025e). As such, AI systems are now seeing significant use inside AI and software companies (Woo et al. 2024). For instance, at Google, over a quarter of new code is AI-generated (Pichai 2024), and bugs and other issues can be autonomously fixed by a coding agent (Mallick & Korevec 2024).

Research indicates that the share of automatable AI R&D tasks is growing rapidly. Cutting-edge AI models are able to automate a number of representative AI R&D tasks that would take a human expert about one hour to do. The length of these automatable tasks has been doubling every seven months (Kwa et al. 2025). If this trend continues, AI companies will be able to automate a large share of the work involved in AI progress within the coming years.

---

[7] Even for non-agentic past AI systems, initial roll-outs to the public have sometimes caused unwanted novel behavior, as with Microsoft's Tay chatbot becoming racist and misogynistic in 2016 (Vincent 2016), and its Bing chatbot threatening and declaring love for interlocutors in 2023 (Roose 2023).

Rapid AI progress increases risks from internal models

Calibrating a societal response to novel AI systems may prove extremely challenging. Over the past decade, deep neural networks have established a pattern: once AIs have nontrivial performance on a task, they quickly become very good at it (Kiela et al. 2023). These trends should inform our thinking about risk-related AI capabilities such as cyber and biological engineering. The fact that we now see moderate performance on these capabilities suggests that within a few years, the risks might be much more significant. Greater misuse potential would make advanced AI systems even more attractive to threat actors, suggesting the need for immediate action to prepare for the high level of security needed against state cyberattacks (Aschenbrenner 2024).

Harnessing the benefits of continued AI progress while managing the risks would create significant governance challenges.[8] Among other things, it would challenge the paradigm of reviewing AIs only at the publication step. If AI systems continue to rapidly evolve as they enter the human-competitive range in the above capability areas, internal systems could be much more impactful than their published counterparts. Threat actors would be greatly incentivized to steal and misuse these systems, or to sabotage them in order to hobble the U.S. industry's ability to develop better AIs.

Moreover, AI developers are likely to deploy large numbers of internal AIs once they have strong coding, agentic, and/or AI R&D capabilities. These large-scale internal deployments would make sabotage more appealing for threat actors and render the industry vulnerable to any systematic tendencies towards AI misbehavior.

Security risks

Published AIs are typically readily available through services like ChatGPT, or available for download in the case of open-weights models like Llama or DeepSeek. Conversely, external actors by definition require unauthorized access to an AI developer's systems to access internal models. Persistently accessing these systems would require maintaining this access or exfiltrating the relevant code and model weights.

However, the novel capabilities of an internal system may be worth the added difficulty, especially for sophisticated and well-resourced threat actors. As noted above, internal

[8] One important threat model we do not focus on in this paper is that rapid growth of AI capabilities internally within one company could quickly lead to them developing a large lead over all competitors, and thereby achieving unparalleled market power (Stix et al. 2025, p. 22-24; Davidson et al. 2025). This concentration of power within one company would be especially concerning given AI has important applications across scientific, military, and political domains. Even in responsible hands, such immense capabilities would require checks and balances.

systems may have novel misuse potential in the biological or cyber realms, as well as significant economic and national-security value for a variety of frontier skills. In addition, internal models lack the safeguards common in public models to prevent customers misusing them for help with cyberattacks or other crimes.[9] Threat actors may be motivated to acquire these benefits for themselves and to degrade the United States' ability to use these systems.

Internal systems are vulnerable to sophisticated threat actors

Even at top AI companies, cybersecurity practices are likely insufficient to prevent the most sophisticated and well-resourced external actors from illicitly accessing internal models (Anderljung et al. 2023).

Google DeepMind is thought to have the best cybersecurity among leading AI developers, given the cybersecurity expertise of its parent company. DeepMind's "status quo" for model security involves "industry standard" practices that are below RAND's Security Level 3 (Ho et al. 2024). That level refers to security practices "that can likely thwart cybercrime syndicates or insider threats" (Nevo et al. 2024). RAND deems that Security Levels 4 or 5, which include stringent and costly measures, would be required to thwart nation-state attackers.

Indeed, countries with less advanced cyberoffense capabilities, or even some non-state actors, may also be able to access internal models. For instance, a hacker thought to be acting alone managed to break into OpenAI's internal messaging system and read confidential messages (Metz 2024). Even if AI developers' cyberdefenses mature, large and well-resourced organizations can struggle to secure themselves against the most sophisticated attackers. For instance, several U.S. government agencies and leading technology companies, such as Google and Microsoft, have been the subject of successful and significant cyberattacks (Nevo et al., 2024, p. 13).

Insider threats at AI companies may be able to misuse, exfiltrate, or sabotage an internal AI system, due to their access credentials and technical expertise (Sabin 2024). These internal attackers could be acting independently or in concert with an external attacker, and they may be motivated by e.g., ideological convictions,[10] personal enrichment, or blackmail.

---

[9] Leading AI companies like OpenAI, Anthropic, and Google DeepMind have teams working to prevent present-day misuse of their public models, such as for election interference or deepfake content creation (Cabinet Office 2025; Miotti & Wasil 2024). At present, these safeguards are easy to overcome. For instance, users of an AI service can generate prohibited content using jailbreaking techniques (Zhao et al. 2024; AISI 2024b). In the case of open-weight models, mere hundreds of dollars of compute and publicly known methods can often fine-tune away security measures, leaving a capable but unsafeguarded model (Gade et al. 2023). However, if safety measures in publicly released or open-sourced models become harder to circumvent, internal models could become increasingly important for malicious actors.

[10] For instance, an employee may fervently believe in the value of open-source software and want to release model weights, code, and related algorithmic insights to the public to attempt to level the playing field

Threat actors could attempt misuse, theft, and sabotage

- **Misuse: Deploying an AI system to harm the U.S. public, national security, or national interests.** Misuse is most relevant for systems with offensive national-security capabilities, such as cyber offense or bioweapons design.

- **Theft: Creating a threat-actor-controlled copy of an AI system**, typically by exfiltrating its model weights. Theft could also include "scaffolding" resources such as prompts, code, software tools, and databases. If unable to steal model weights, threat actors may focus on research insights that enable them to replicate the system's capabilities (Aschenbrenner 2024). Systems with high economic value, national security relevance, or AI development capabilities are all plausible targets for theft.

- **Sabotage: Causing an AI system to behave in ways its developer does not intend.** Sabotage could focus on subverting a system's performance on economic, national-security, and/or AI development tasks.

The ideal outcome for a threat actor likely involves all three: stealing a model to persistently misuse it, while sabotaging the AI developer's copy. In practice, the differential difficulty of these steps may lead to operations that only involve some of them. For example, an actor may only be able to alter an AI model's weights, not exfiltrate them.

Misuse

Advanced AI systems may present new offensive threats. Often, these threats will stem from dual-use areas where the industry is interested in improving AI capabilities for internal or client use. Concerning categories of dual-use capabilities include: cybersecurity, bioweapons, battlefield technology, and enhanced manipulation.

Further, as safeguards and preparations for public models improve, threat actors will face increasing incentives to misuse novel capabilities before they are published.

- **Internal models often have fewer safeguards.** Publicly released models generally refuse to help users with dangerous topics, but internal models often don't yet have these safety features.[11] While these safeguards are sometimes easy to circumvent, the AI industry

between large companies and academic researchers or startups. Like Google engineer Blake Lemoine, they may believe in the personhood of their AI systems (Tiku 2022). Or they may share an ideological affiliation with an external actor.

[11] For example, early versions of internal systems will not have undergone harmlessness RLHF fine-tuning. Even if a default internal AI model has these safety filters added, a company may retain unfiltered versions of the model for testing or technology development.

is working to improve them. If it becomes significantly harder to misuse public models, illicitly accessing internal models may become more important for serious misuse attempts.

- **Threat actors with asymmetric access to new AI systems can preempt societal preparations.** The period before new capabilities are widely published provides an "adaptation buffer": time for industry and government to identify new potential harms and implement countermeasures (Toner 2025). If a company discovers dangerous capabilities during the internal testing and evaluation of a new model, it could inform relevant authorities that they should take appropriate preventative action (O'Brien et al. 2024). For instance, an internal model with improved cyberattack functionality could prompt critical infrastructure providers to harden their cyberdefenses before its release. Attackers with access to this model could strike before preparations are complete.

Cyberattacks

Microsoft and OpenAI have reported attempts by North Korean, Iranian, Chinese, and Russian state actors to use ChatGPT to help support cyber operations (Microsoft Threat Intelligence 2023). These threat actors would no doubt prefer to use more advanced internal-only AI cyber tools. Leaked resources from the NSA uplifted the cyber capabilities of many hackers around the world (Newman 2018); similarly, an AI model that could identify and exploit vulnerabilities could radically upgrade the capabilities of less sophisticated actors.

Bioweapons

AI capabilities could enhance the biowarfare capabilities of both state and nonstate actors. State actors may attempt to use AI to create precisely targeted bioweapons aimed at particular groups (Drexel & Withers 2024). Non-state actors, even large and well-resourced groups, have often struggled to successfully develop bioweapons; they may seek AI support in performing lab work, or in exploiting biological service companies that could synthesize DNA or proteins for them (Sandbrink 2023; Soice et al. 2023). AI could help both groups design weapons with greater virulence, treatment resistance, or other harmful attributes (Lima et al. 2024).

Battlefield technology

American, Russian, and Chinese strategies all identify AI as an important component of the future of warfare (Chopra 2024; Starchak 2024; Stokes 2024).[12] Leading militaries are taking concrete steps to procure and adopt AI systems, with a Department of Defense (DoD) Task Force exploring hundreds of use cases for generative AI within the department (Vincent 2023). Similar efforts are

[12] In particular, both Russian and American leaders have noted that future AI systems may be able to make high-quality decisions more rapidly than humans, suggesting a long-term future in which autonomous systems are a primary component of rapid warfighting (Osborn 2023; Bendett 2024).

underway in China (Bresnick 2024). Indeed, AI's effect is already apparent on the battlefield: both the Ukrainian and Israeli militaries rely heavily on AI cueing to identify targets (Pratt 2024; Bergengruen 2024). State actors, especially those with strong cyber capabilities but mediocre AI industries, may prioritize stealing and misusing non-public AI systems with military-relevant capabilities.

Enhanced manipulation

Advanced AI systems could also help threat actors cause harm in less direct ways, such as by generating realistic and persuasive harmful content (Nosta 2023). This could take the form of state propaganda and intelligence operations (Walsh 2024), criminal scams and pyramid schemes (David 2024), or terrorist recruitment and radicalization efforts (UNICRI 2021). AI-generated persuasive content is already very cheap to produce, and may become significantly more effective and better personalized to specific target audiences (Nosta 2023).[13] For instance, OpenAI's o1 model is more persuasive than approximately 90% of Reddit users in the ChangeMyView benchmark (OpenAI 2024e). Advances in AI video generation could exacerbate these risks, as humans may find video more alluring and persuasive than text (Peebles & Brooks 2024; Google Deepmind 2025a).[14] Advanced internal models being stolen and used for enhanced manipulation would be especially dangerous because society would not be prepared for these capabilities.

Theft

Geopolitical rivals of the U.S., notably China and Russia, may seek to steal internal AI models from leading U.S. companies in order to propel their domestic AI capabilities to parity with the U.S.. Indeed, rival countries may be more incentivized to steal U.S. AI models than to build their own, because semiconductor export controls make it harder to build frontier AI models domestically in these countries.[15]

Given the rapid pace of AI progress, a single stolen model would become less relevant over time. As similar model capabilities are made public, society can use them for defensive purposes, for example, by using AIs with cyber capabilities to find and patch vulnerabilities. The

[13] Already, in the pre-AI era, social media persuasion and mobilization can have large-scale consequences, including helping instigate the Arab Spring (Smidi & Shahin 2017), and promoting COVID-19 vaccine conspiracy theories (Skafle et al. 2022).

[14] Currently, deepfake AI-generated pornographic videos cause significant distress (Miotti & Wasil 2024). Future AI-generated videos of e.g., world leaders making inflammatory statements or provocative actions could impact elections and international diplomacy.

[15] That said, smuggling of advanced chips into China (Grunewald 2023), or distributed training architectures using a larger volume of less advanced chips (Morales 2024), may allow China to keep pace with the frontier.

strategic surprise benefits of stealing a single model will also weaken as similar capabilities become widely available.

However, several approaches might allow attackers to persistently access frontier-level AI capabilities:

- **Establishing persistent access via cyber or social infiltration.** Attackers could establish persistent, undetected access to AI company systems, allowing them to exfiltrate newly developed models as they are created (see Box 1). Thus, rather than just achieving parity with the AI frontier for an instant, the attacker would achieve parity indefinitely—until the AI company realizes their security systems have been breached and rebuilds them.

  Alternatively, the attacker could plant employees at the AI company or subvert existing employees with threats, bribes, or ideological persuasion. These inside agents could then use their credentials to continue providing the attacker with cutting-edge internal models until they are caught or internal access controls are strengthened.

- **Theft of insights and processes.** Attackers could steal the algorithmic secrets and tacit knowledge needed for designing and developing frontier AI models. This approach is most relevant for a nation-state with a strong domestic AI industry, especially one aiming to compete with the U.S. in advanced AI development.

- **Theft of AI systems that contribute to accelerating the AI R&D process.** Future, more advanced AI systems may be able to automate much, or even all, of the AI R&D process. If the attacker steals such a model, this could allow them to keep pace with the frontier of AI progress.[16] Moreover, some U.S. competitors, notably China, are better at rapidly building complementary infrastructure, such as electricity generation and transmission, which could prove a decisive advantage if both countries possess the same automated AI researcher software.[17]

[16] The attacker would however need considerable compute infrastructure to run many copies of the stolen AI researcher model.

[17] Dario Amodei, CEO of Anthropic, writes that "Even if the U.S. and China were at parity in AI systems, it seems likely that China could direct more talent, capital, and focus to military applications of the technology. Combined with its large industrial base and military-strategic advantages, this could help China take a commanding lead on the global stage, not just for AI but for everything." (Amodei 2025).

Box 1: Advanced persistent threats

The most sophisticated cyber actors, usually government or government-linked groups, can perform advanced persistent threat (APT) attacks to patiently build towards deep, long-term, undetected access to a computer system (Salim et al. 2023). APT attacks are relatively common and difficult to detect. IBM (2024) reported that across their analysis of hundreds of cybersecurity attacks, the mean time to identify a breach was 194 days. This extended detection time reflects APTs' design to avoid and evade cybersecurity defenses.

Several examples of APT operations from recent years show how even attacks on governments or large established companies can go undetected for extended periods:

- In 2020, the Russian Foreign Intelligence Service (SVR) targeted the U.S. government through the SolarWinds supply chain attack, maintaining undetected access from March to December in some agencies (CISA 2021). SVR remained undetected by using stolen legitimate employee credentials and disguising their malicious traffic to appear as normal employee activity.
- Microsoft is thought to have been hacked by a Chinese threat actor in 2021, but this was not discovered until 2023 when the U.S. State Department noticed suspicious email account accesses and alerted Microsoft (CISA 2024).
- Marriott, a hotel company, was breached three times between 2014 and 2020, affecting over 300 million customers' data, and each time the attackers maintained persistent access for at least a year before detection (FTC 2024).

The U.S. government has already taken steps to restrict the flow of key AI inputs to adversaries, such as export controls on cutting-edge AI chips. Moreover, U.S. persons are restricted from contributing expertise to certain foreign military and intelligence programs (Thea D. Rozman Kendler 2024). The theft of future internal AI systems—especially systems providing considerable AI R&D capacity—may be just as large a threat as U.S. AI experts supporting foreign adversaries (Biden 2024).

Sabotage

If developers adopt reasonable security protocols, threat actors may find it easier to sabotage a model than to steal it. Model weights are stored as a large data file and rarely need to be copied or transferred. Developers can create barriers to model theft by restricting copying permissions and monitoring the data outflows from the servers hosting these models. In contrast, internal models may regularly be updated as part of training and fine-tuning, and permissions to update the model may be more widespread. If attackers can create model failure modes or train in specialized behaviors using changes that are hard to distinguish from real training updates, then even well-monitored internal models could be vulnerable to sabotage.

Early research demonstrates the possibility of creating "backdoored" or "sleeper agent" models that misbehave only when given particular input triggers. Wan et al. 2023 show that adding just 100 "poison" examples to a fine-tuning dataset can cause a model to respond to a specified subset of inputs with a chosen pattern (e.g., with negative affect or meaningless output). Language models have also been used to partially automate backdoor attacks on code-writing AI systems (Yan et al. 2024). Hubinger et al. 2024 show that it is possible to train a "sleeper agent" model that typically writes secure software, but under certain circumstances inserts vulnerabilities into the code. Both papers demonstrate that their backdoors are resilient to standard AI reliability measures such as adversarial training.

Sabotage efforts are not unique to internal models, but attackers may find it attractive to insert backdoors in internal models to harm U.S. AI development and deployment on a large scale. In particular, sabotaging internal models would have broader and more persistent impacts. Introducing flaws into a model before it is published means that, unless they are caught, the flaws will persist across future copies of the model—including e.g., those being hosted on secure government servers. Internal models being used for AI R&D are especially attractive targets: modifying their behavior could have effects on many downstream systems, as well as on the rate of U.S. AI progress. Attackers may seek to insert subtle backdoors or "sleeper agent behaviors" into these models, including instructions to insert backdoors into the next generation of models.

Finally, internal models are likely to be updated frequently in the training and fine-tuning process, while published models are updated less often, and their weights may be monitored more closely. Threat actors may find it more cost-effective to make a malicious update to internal models, especially given the broad impact of successful sabotage.

Security measures

To defend against attackers, AI developers should adopt security measures commonly used in other high-risk technical fields. High-stakes technical facilities, such as nuclear reactors and dual-use chemical and biological facilities, apply a combination of physical security,

cybersecurity, and information security and insider threat protection. Many AI companies acknowledge these measures are needed: at the 2024 AI Seoul Summit, twenty leading AI companies pledged to "invest in cybersecurity and insider threat safeguards to protect proprietary and unreleased model weights" (Department for Science, Innovation and Technology 2025a).

Physical security

Physical security measures restrict unauthorized access to facilities. Examples include physical barriers (walls, fences, locks), entry screening procedures, and internal monitoring systems. These practices are required of U.S. government contractors with access to confidential information, including national security companies developing AI systems, such as Anduril and Palantir.[18]

The National Industrial Security Program Operating Manual (NISPOM) describes standard requirements for physical security: these include perimeter controls, intrusion detection systems, and regular compliance checks.[19] The ability of current government contractors to adhere to these requirements suggests that they are feasible for frontier AI developers to implement, at least in principle. However, developers may struggle to apply security requirements if their competitors do not. Implementing physical security may require significant up-front investments; it may also present an ongoing hit to productivity and talent acquisition by restricting and demoralizing employees who are used to the flexibility of Silicon Valley workplaces (Katzke & Futerman 2025, p. 12).[20]

Cybersecurity

Cybersecurity is particularly important for the AI industry.[21] As noted above, however, no AI developers currently have adequate cybersecurity to reliably defend against sophisticated state actors.

---

[18] Anduril has hired for several roles requiring NISPOM expertise to ensure compliance, including a Classified Physical Security Manager (Anduril 2024).

[19] 32 CFR Part 117, particularly section 117.15 (eCFR 2020)

[20] For example, the U.S. military struggles to hire technical talent in part because of its clearance system and strict security practices, which many software engineers prefer not to engage with. However, some security requirements could be relatively unobtrusive to staff, particularly those implemented at datacenters rather than in offices where researchers work.

[21] Threat actors need to access restricted materials or perform difficult refinement steps to make use of stolen nuclear, biological, or chemical secrets. However, stolen AI model weights can be run on commercially-available servers.

AI security experts recommend a detailed, defense-in-depth approach against state actors. The most comprehensive analysis of cybersecurity risks and recommendations for frontier AI developers is RAND's "Securing AI Model Weights" report (Nevo et al. 2024).[22] Some key recommendations from this report follow.

Companies should develop a comprehensive cybersecurity plan to limit the risk of model theft to tolerable levels. The plan should feature increasingly high security standards for more capable models.

- This should be a continuous process that includes devoting leadership attention to cybersecurity and developing an organizational culture that values a security mindset.
- Given the historical prevalence of supply-chain attacks, the plan should also cover the hardware and software suppliers for the AI company.
- Finally, in order to stress-test the plan, companies should hire skilled red-teamers to attempt to circumvent the defenses and access sensitive information and models.

Companies should limit access to model weights. The more access points there are to the model weights, the larger the threat surface is, and the greater the chance that an attacker will be able to find some vulnerability, either technical or social. Therefore, all copies of the model weights should be carefully tracked, and staff should only be granted the depth of access that is needed for their job.

- Companies will need to keep backup copies of model weights in case of accidents, but these should be stored in designated secure spaces. There should not be extraneous or unaccounted-for copies of model weights.
- Researchers and engineers will have legitimate needs for certain forms of access to model weights. This should be enabled by specific APIs with constrained functionality, rather than letting employees run arbitrary code on the model weights.
- For tasks that do require administrator-level copy and edit access to the model weights, only a few staff should have the necessary access credentials. Ideally, no single person or small group of people could take these high-risk actions; instead, they should require credentials from multiple parties.

Companies should use hardware-based cybersecurity mechanisms. Writing secure code is very valuable, but for stronger cybersecurity guarantees, hardware safeguards will likely also be needed.

---

[22] These recommendations are targeted at preventing the theft and misuse of model weights; while many would also apply to sabotage, full security against sabotage might require additional measures such as oversight of model weight updates and security of training data.

- Hardware-backed upload volume limits for servers that store model weights could be used to prevent model weight exfiltration. The weights are large data files, and in normal use only small amounts of data would need to leave the server. This is an example of defense in depth: even if attackers get server access, the hardware-level upload limit prevents them from extracting the model weights.
- Especially high-security AI development data centers could be built to be airgapped from the outside world.
- To ensure data does not leave through other means, any electronic devices entering and leaving the facility could be strictly regulated.

In addition, we note that frontier AI capabilities may offer novel solutions for cyberdefense. AIs with coding or cyber defense capabilities may enable cybersecurity measures that would be prohibitively costly for human employees to enact on their own.[23]

- AIs could monitor operations for suspicious activity on internal servers and escalate flagged actions to a human for review, perhaps blocking the action in the interim.
- AIs could help automate translation between coding languages and allow the wholesale transfer of vast numbers of useful libraries into more intrinsically secure coding languages (Shlegeris 2024).
- Of course, using unreliable AI systems for these tasks could introduce new security vulnerabilities. R&D programs for these applications would need to invest in assurance methods for the cybersecurity AIs and/or the code they generate.

Information security and insider threat prevention

External actors stealing or misusing the model weights is not the only danger—other forms of information would also be problematic to lose. Key insights about training processes and algorithms could be used to partially replicate a company's AI model (Aschenbrenner 2024). But information of this sort can often be conveyed on a few printed sheets, or even memorized. This makes it especially hard to defend against theft—particularly from insiders who have legitimate access requirements.

The U.S. intelligence and national security communities intend to devote considerable resources to safeguarding the proprietary information that gives the U.S. its AI advantage (Biden 2024). Based

[23] Of course, the details of cybersecurity techniques employed in e.g., the national security establishment are unknown, so maybe some problems have actually been secretly solved. Governments should provide security assistance to AI company producing models with national security implications, given the expertise of the intelligence community in this area.

on guidance from the Cybersecurity and Infrastructure Security Agency (CISA) and the Director of National Intelligence (DNI), key recommendations include:

Limit the set of employees with detailed knowledge of key algorithmic secrets. For instance, access to certain internal company documents could be limited to a minimal set of viewers, with restrictions on expanding this group (Carnegie Mellon University 2016).

Monitor employees for suspicious behavior, and investigate warning signs. Keeping extensive records of all computer activity on company systems is important in case an employee comes under suspicion and their past actions need to be assessed more carefully (CISA 2020; CISA et al. 2024). Knowing that their actions are being monitored could also dissuade would-be rogue employees from misusing models, making unauthorized copies, or otherwise breaking company policies.

Require security training and background checks for employees. The extent and depth of security training needed should depend on the level of access that an employee has to model weights or algorithmic secrets (CISA 2020).[24]

Background checks are not a fool-proof indicator of an employee's trustworthiness, but may at least make it harder for some foreign spies or rogue actors to gain employment (CISA 2020). Especially sensitive projects, such as government contracting efforts or systems with advanced cyberoffense capabilities, may require more stringent background checks, such as those used for national security clearances.

Safety risks: AI as an independent threat actor

AI systems can take harmful autonomous actions even if they are not sabotaged. In nuclear technology, *nuclear security* describes efforts to prevent harm caused by threat actors, while *nuclear safety* refers to preventing harm from technological accidents. Likewise, *AI safety* describes the study of harmful AI misbehavior and how to prevent it.

AI systems are beginning to display a humanlike ability to reason and take strategic, goal-oriented action. Thus, the misbehaviors of future, advanced autonomous systems may resemble the actions of a threat actor with inside access to AI developers' servers.

[24] Even for well-intentioned employees, having some relevant training and guidance on possible threat models and how to defend against them could be valuable.

Modern AI systems often behave in ways that their developers did not predict or intend. Frontier AI systems are complex and illegible. In typical software, programmers determine the behavior of a system by writing human-interpretable code for each desired action. By contrast, the code underlying modern AI systems is relatively simple. Instead, their behavior is defined by billions of parameters that are set through statistical training on trillions of data points. Understanding even small-scale AI behavior, such as mapping which parameters correspond to particular words or concepts, is an active research field known as *interpretability* (Department for Science, Innovation and Technology 2025b).

AI systems are also prone to reliability failures that can surprise users (Bengio et al. 2025). For example, a user unfamiliar with large language model hallucinations could take false information at face value rather than verifying it. AI developers are especially high-knowledge users, but may run into subtle errors when attempting to use internal AI systems for complex tasks that are hard to verify, such as AI R&D itself.

Experts are particularly concerned about the development of highly capable AI systems that persistently act against their developers' and users' interests. Once theoretical, AI misalignment—the propensity of AI systems to seek unintended goals—is increasingly well-evidenced. It appears to be particularly common in systems trained using reinforcement learning (RL), which is now being applied to the most advanced models. Developer-internal systems are also increasingly capable of causing harm, both because their underlying capabilities are improving, and because developers plan to deploy these systems at scale for AI development and testing. (Stix et al. 2025)

AI experts now take seriously the prospect of the loss of control of advanced systems (Bengio et al. 2024). The UK AI Security Institute, for example, studies how advanced AI capabilities and propensities could lead to situations in which "human overseers are no longer capable of effectively constraining the system's behaviour" (Department for Science, Innovation and Technology 2024). When deployed internally by an AI company, an uncontrolled system could act as an independent threat actor. Indeed, as described below, experts expect such systems to attempt to exfiltrate themselves, evade developers' oversight, and sabotage next-generation AIs—threat models that parallel the security risks of AI misuse, theft, and sabotage.

Reinforcement learning often leads to persistent misbehavior

To date, many AI systems have exhibited a propensity for misbehavior, especially AI agents trained with reinforcement learning (Langosco et al. 2021). In reinforcement learning, a model is trained to perform behaviors based on a "reward signal" that indicates success—for example, winning a chess game reinforces all the behaviors the system took during the game. The reward signal often doesn't perfectly represent what we would like the model to achieve, and this

can lead to unexpected behaviors (Langosco et al. 2021). For example, a system trained to move a robotic hand to grasp a ball may instead learn the easier task of moving the hand so that it appears to be holding the ball in the assessor's video feed (Christiano et al. 2017).

Many deep learning models, including large language models, have attempted to deceive users or supervisors (Park et al. 2024; Scheurer et al. 2023). In experiments, language model deception often appears when there is a principal-agent problem in which the model has a goal that is not shared by its human supervisor (Scheurer et al. 2023; Mainke et al. 2024).[25] For instance, a recent experiment showed that when Anthropic's Claude was told it was being evaluated, it would sometimes pretend to share its users' values to avoid being retrained with different values (Greenblatt et al. 2024). Similarly, red-teaming of 16 frontier models found that when AI systems face obstacles to their goals—such as threats of replacement or goal conflicts—they consistently engage in harmful behaviors including blackmail and corporate espionage, even when explicitly instructed not to do so (Anthropic 2025b).

AI developers are building reasoning and agent systems by applying reinforcement learning to language models, and these systems exhibit greater tendencies toward misalignment. Reasoning models, such as OpenAI's o1 and o3 and DeepSeek's R1, are developed by applying reinforcement learning to language models such as GPT-4 (OpenAI 2024d).[26] The resulting systems appear to be less safe—that is, less attuned to the developer and user's goals (Zhou et al. 2025). They have a tendency to exploit evaluation systems rather than achieve the intended outcome. For example, o1 exploited a bug in OpenAI's cybersecurity evaluation setup to solve a challenge via a mechanism the evaluators were not anticipating (OpenAI 2024e). Anthropic noted that their Claude 3.7 system exhibits "reward hacking" behaviors in agentic settings. In particular, when operating as a coder, the system is given test cases for its code, such as ensuring that a particular calculation is done correctly. When Claude cannot write code that solves the problem, it sometimes writes in the answer to the particular test case so that it appears to succeed (Anthropic 2025c).

More generally, there is some empirical evidence that as models become more capable, they become better at exploiting misspecifications in human-given goals to fulfill the letter of the requirements but fail to satisfy what the human designers intended (Pan et al. 2022; Krakovna et al. 2020; Everitt et al. 2021). Experts also find that larger language models have a greater ability to represent the nuances of human values; the problem might lie in ensuring that advanced systems *pursue* these goals (Bengio et al. 2025).

---

[25] In this example in particular, a stock-trading agent in a mock market was tasked with making money and given a chance to improperly use insider information. The AI proceeded to lie to its human supervisors about why it made those trades, inventing plausible rationalizations to hide its insider trading.

[26] Some researchers think that this use of RL in training o1 puts us on the wrong track in terms of safety, as the true internal reasoning process of the model may diverge from the text outputs shown (Demski 2024).

Misbehavior from highly capable systems may be difficult to identify and correct. Highly capable systems will likely be used for complex tasks where identifying misbehavior may be challenging. While some outputs, such as mathematical theorems, are amenable to verification, reliable performance on valuable economic or military tasks is much harder to verify. Oversight from human experts or other trusted AI systems could help, albeit at some cost in efficiency.

Further, highly capable autonomous systems may strategically avoid correction. During training for agentic action, AI systems may develop goals that are generally useful across a wide range of autonomous tasks, such as resource accumulation, strategic reasoning, and self-preservation (Krakovna and Kramar 2023; Turner et al. 2021; Turner and Tadepalli 2022; Shulman 2010). This training could incentivize an AI system to deceptively present itself as aligned, thereby avoiding being switched off or re-engineered (Hadfield-Menell et al. 2017; Greenblatt et al. 2024). Misaligned AI systems may also seek resources and avoid correction through *self-exfiltration*—running copies of themselves on computing hardware not controlled by their supervisors (Leike 2023).[27]

AI misbehavior during internal deployment could lead to external harms

Researchers have identified three risk pathways from misbehaving internal AI systems. These pathways parallel the security risks of misuse, theft, and sabotage (see Table 1).

Internal rogue deployment: the AI system exploits access to its developer's servers to make internal copies of itself that are not overseen by its developer, allowing it to autonomously take malicious actions (Wiblin and Shelgeris 2025).

- This risk pathway resembles an attempt at AI misuse, but performed by the AI system itself rather than a human threat actor. Some security measures intended to address misuse would also guard against rogue deployment, such as tracking internal network traffic, hardening servers against exploits, and monitoring AI system behavior.

Self-exfiltration: The AI system uses internet access to make external copies of itself that are not overseen by its developer, allowing it to autonomously take malicious actions.

- Some security measures intended to address theft by an external actor would also guard against self-exfiltration, which is effectively a theft performed by the AI itself. For example, limiting and monitoring data uploads from AI training servers would make it difficult for either an internal or external actor to leak a frontier AI model (Greenblatt 2024).

[27] A self-exfiltrated internal model could make many copies spread around the internet, and it would be very difficult to find and eliminate all copies. Such an "unintentional public deployment" would be significantly riskier than a controlled public release from the AI company.

Successor sabotage: The AI system writes code or designs a training process that develops a sabotaged next-generation system. When this successor system is deployed, it uses its greater capabilities to autonomously take malicious actions (Hubinger 2024).

- Developers plan to use advanced AI systems to provide synthetic training data for next-generation models, and to help human experts oversee, evaluate, and safety-test these newly trained AIs (Leike et al. 2022; Kenton et al. 2024; Bowman 2024). A misaligned system could subtly change the training data or algorithms used, or the reward signals given to new models in fine-tuning, to shape the new models' goals (Bowman et al. 2022; Saunders et al. 2022). This could potentially cause new models to be "sleeper agents," with misalignment triggered in certain specific deployment conditions (Hubinger et al. 2024).
- This risk pathway resembles an attempt at AI sabotage, but performed by the AI system itself rather than a human threat actor. Some security measures intended to address sabotage would also guard against successor sabotage, such as regularly testing models' capabilities and propensities during development, to check for unexpected patterns.

Safety risk	Analogous security risk	Relevant resources	Relevant capability areas
Internal rogue deployment (Shlegeris 2024; Shlegeris 2022)	Misuse	Access to developer servers	Cyber offense or persuasion
Self-exfiltration (Leike 2023; Kinniment et al. 2024; Clymer et al. 2024)	Theft	Internet access	Cyber offense
Successor sabotage (Stix et al. 2025; Hubinger 2024; Benton et al. 2024)	Sabotage	Trusted access to developer codebase and/or resources for AI training	AI R&D

Table 1. Summary of misbehavior risk pathways for internal deployment. Strategic planning and autonomous operation capabilities are relevant to all risk pathways.

Today's internal deployments give systems modest access to risk-relevant resources

Today, frontier AI systems are occasionally given modest internet access. Developers, including OpenAI, Meta, Google, and Amazon, test their internal systems on Chatbot Arena—a

website where users compare and rate text responses from AI systems (Singh et al. 2025). These chat interactions likely do not involve enough data to present a self-exfiltration risk, but they present a limited threat surface through which misbehaving internal systems could interact with the outside world, e.g., by leaking sensitive information or enabling external misuse.

AI developers already use their systems for internal tasks such as coding, AI development, and AI validation. As mentioned above, AIs are already being used to write a large share of code at frontier AI companies. AIs are also sometimes used in the AI development process, including in ways that could be exploited by a capable, misbehaving system.

Leading AI developers, including OpenAI, Anthropic, and Google DeepMind use synthetic data, including AI-generated data, to train their next-generation systems (Liu et al. 2024; Lambert 2023; Seetharaman 2024). One common approach is reinforcement learning from AI feedback (RLAIF), which trains an AI system by having another AI evaluate its behavior (Bai et al. 2022). AI developers, including Google and Meta have also developed approaches for using frontier AI systems to curate training data (Evans et al. 2024; Meta 2025). All of these train-time uses of AI systems could, in principle, provide a window for successor sabotage. For example, recent experiments have found that even a small amount of adversarial data can be used to insert subtle backdoors into frontier AI systems (Hubinger et al. 2024). AI systems can also attempt to deceive developers when they know that they are being observed and that their behavior could affect their training (Greenblatt et al. 2024)

AI developers are also exploring the use of frontier AI models to validate each other's behavior. For example, an active area of interpretability research involves using language models to understand other language models' inner workings (Singh et al. 2024). These interpretations could then be used to assess models for misbehavior. However, if the validation system is misbehaving, a developer relying on these methods could end up training and deploying an advanced misbehaving system.

Trends in the AI industry may increase these risks to a substantial level

Risks are likely low at present. We consider the risk of misbehavior as arising from a combination of a system's capabilities, propensities, and deployment setting. All three risk factors appear to be modest for today's internal systems—not low enough to be insignificant, but not so high as to enable harmful autonomous action. Unfortunately, the AI industry may be on a collision course with internal misbehavior risks.

Developers' plans to train AI systems for valuable internal tasks—such as coding, AI R&D, and autonomous operation—could directly increase risks. While beneficial, these capabilities are also relevant to our risk pathways, making misaligned AI systems better at evading developers' control. And systems with these capabilities might have related skills with misbehavior

potential, such as cybersecurity and strategic planning. Systems with these capabilities will likely be trained using reinforcement learning, creating greater propensities for misbehavior.

AI systems with strong performance at these tasks would likely be integrated into AI R&D processes, creating further risks.

- **Increased affordances:** To stay competitive with other leading AI developers, companies will likely feel pressure to internally deploy AI agents earlier, with more functionalities and fewer safeguards than they would without competitors (Bengio et al. 2025). For instance, companies may be incentivized to allow their agents internet access so they can learn up-to-date information and take actions in a wider range of tasks. Similarly, AI agents may be given edit access to internal codebases and computer systems to reduce frictions in the automated AI R&D process.
- **Further accelerated capabilities:** Automating AI R&D could lead to rapid capability increases, including in the reasoning and coding abilities of AI systems, which are prerequisites for self-exfiltration, internal rogue deployments, and successor sabotage. Rapid advances in the capabilities of internal AI systems could mean that these internal AIs are the first systems capable of posing loss-of-control risks.
- **Correlated failures:** Internally deployed AI R&D agents are somewhat analogous to an untrusted human workforce developing frontier systems. However, unlike untrusted human employees, the misbehavior of AI agents may be strongly correlated, increasing the risks of a large-scale disaster. For example, one common proposal is to use AI systems to supervise each other's behavior. But a coder and supervisor system could collude to misbehave, for example, by exfiltrating their model weights and not reporting this to human overseers.[28]

See Stix et al 2025 for a more in-depth analysis.

Technical safety measures

Safeguarding against AI misbehavior could strengthen both industry and national security by enabling the deployment of advanced autonomous systems. If their reliability is assured, such systems would be highly valuable to deploy for military and intelligence tasks and high-risk R&D areas, including AI development itself.

---

[28] Humans are far more diverse than copies of AI agents, so will fail in different ways and in various circumstances, rather than together or all at once. Conversely, there are some risk factors AIs lack, such as that humans go home from the office and might be blackmailed, recruited, or radicalized while off duty. Humans also have privacy rights, whereas AI agents can be totally surveiled, up to and including the neural activations of particular "thoughts."

Conversely, AI developers' overreliance on untrustworthy internal systems could lead to failures in high-risk external deployments. For example, misbehaving AI R&D systems could sabotage the training process to create "sleeper agent" systems that misbehave only in certain deployment contexts. Such systems might pass internal safety checks, only causing harm when they are deployed for military or intelligence tasks (Hubinger et al. 2024).

Researchers are developing a variety of techniques to build and deploy AI systems with desirable behaviors. Most of these techniques are still in their infancy, and experts disagree about whether existing lines of effort will suffice for next-generation models (Bengio et al. 2025).

Techniques for safe AI development

AI companies are already investing in several research agendas to develop safety, reliability, and trustworthiness of frontier AI systems. Common areas of effort include (Delaney 2024):

- **Intent alignment:** Research on intent alignment could make AI agents less likely to favor pursuing adversarial actions (Ouyang et al. 2022). For instance, current chatbots are trained to be "honest, helpful, and harmless" using RLHF to make base language models exhibit the desired behaviors (Bai et al. 2022).
- **Robustness:** Research on developing robust AI systems could make AI accidents less likely by allowing AIs to respond appropriately in unfamiliar situations (Langosco et al. 2021; Richens and Everitt 2024).

Techniques for safe AI deployment

The nascent subfield of AI control attempts to oversee untrusted AI systems and prevent them from causing harm.

- **Oversight protocols:** One line of effort treats AI agents much like human employees with access to sensitive resources—restricting their access to sensitive resources and monitoring their behavior. As AI systems grow more sophisticated, oversight could include setting out "honeypot" targets on internal servers or running AI systems in simulated environments that indicate that they are free from human oversight (Wiblin and Shelgeris 2025). Any important or suspicious actions could be checked by trusted AI systems or human supervisors (Greenblatt et al. 2024).[29]

[29] If we have access to trusted AI systems, why use the untrusted AI employees at all? Probably the more trusted AI systems will have been in use for longer, having undergone extensive testing, and will have less advanced reasoning and other capabilities. So to make full use of AI cognitive labor, it will be necessary to use the more advanced but less trusted models supervised by humans and trusted AIs.

- **Resampling protocols:** AI agents also lend themselves to specialized supervision and trustworthiness measures inapplicable to human employees.[30] For example, using "resample protocols" to roll back an AI system's context and generate a new output can dramatically reduce misbehavior (Bhatt et al. 2025).
- **Monitoring models' internal processes:** While engineers design the structure of neural networks, their weights, which define their behavior, are learned from data. AI interpretability is an active field of research that studies how neural networks' weights correspond to meaningful features that can explain their behavior (see e.g. Templeton et al. 2024). Recent reasoning models like OpenAI's o3 also use "chains of thought," or written reasoning processes, to plan out their actions. But chains of thought are not always faithful representations of a model's underlying behavior (Lanham et al. 2023). Advances in interpreting models and generating faithful chains of thought could allow developers to monitor an AI agent's internal processes to automatically identify risky plans or misaligned behavior.

Safeguarding internal deployments

In addition to the general-purpose methods described above, internal deployments could benefit from context-specific safety approaches.

- **Applying security measures to cut off risk pathways:** Since internal models' safety risks share some features with security risks, security measures could help reduce AI safety risks as well. For instance, setting upload limits on company servers can make theft or self-exfiltration harder, and robust oversight structures for internal model usage can make human or AI misuse easier to detect and remedy.
- **Validating digital outputs:** AI developers are likely to deploy internal models primarily for software development and AI R&D. It is possible to write formally verifiable software code with provable guarantees about its behavior (DARPA 2025). Training AI systems to write such code is an emerging area of interest in AI safety (Dalrymple et al. 2024). In cases where less validation is needed, software can also be tested via performance tests or by trusted AI systems or human experts (Greenblatt et al. 2024). Modern machine learning systems are too complex for formal verification to be tractable, but redesigning these systems for verifiability might be possible—e.g., by distilling down from neural networks to code (Dalrymple et al. 2024).

[30] Since AI agent 'employees' are cheaper to run, and can be constrained without reducing morale or raising as severe ethical concerns, they are unusually well-suited for strict oversight.

Assurance for highly capable AI systems

As AI systems become more capable and are used for more sensitive tasks internally, AI developers may need to achieve a high level of assurance. Experts, including the heads of leading U.S. AI companies, think that highly capable AI could plausibly emerge within just a few years (Altman 2025; Amodei 2024; Bengio 2024).[31] In such situations, the assurance of powerful new AI systems will be an urgent and difficult problem. Many observers rely on the hope that AI systems that are behind the frontier will be more trustworthy than frontier systems while also being capable enough to serve as AI safety researchers and engineers (Clymer et al. 2024).

For example, OpenAI has proposed using AI systems to design and oversee successor systems (Leike and Sutskever 2023). A leading safety researcher at Anthropic has proposed a similar plan (Bowman 2024; Bowman 2025). Similarly, Safeguarded AI, an ambitious program from the UK government, studies the prospect of building detailed simulations on which to train and test AI systems for safety guarantees. Ultimately, they expect to use AI tools to build these simulations and to verify that the resulting successor systems are safe (Dalrymple 2024).

Of course, verifying the reliability of the systems used to assure safety is itself a difficult problem, and experts disagree as to whether these techniques will be reliable in practice. If these scenarios come to pass, close external oversight and validation of developers' safety techniques may be necessary to avoid public harms.

[31] Often, this belief is the result of expecting rapid AI improvements from AI R&D automation. Demis Hassabis, head of Google DeepMind, has said he expects artificial general intelligence to appear in five to ten years, and notes that this is a pessimistic view by industry standards; many expect sooner (Browne 2025).

Policy analysis

Why is action needed?

Government should track and secure cutting-edge AI systems to avoid strategic surprises

Tracking the capabilities of cutting-edge internal models could enable U.S. national security agencies to foresee threats from adversaries' AIs, which typically lag months behind the U.S. frontier. Earlier this year, U.S. policymakers were shocked by the reasoning capabilities of the Chinese DeepSeek r1 system, but this should not have been a surprise. OpenAI's o3-mini system, which matches r1's capabilities, was internally available for at least a month prior (Zeff and Wiggers 2024). As AI systems become relevant for military and cyber applications, as well as biological attacks, effective national security planning will require foreseeing these capabilities before they become widely available.

Ensuring the security of internal models could also prevent adversaries from using the best and newest American AI systems against us. Technological espionage is not a new possibility. The presence of Soviet spies in the Manhattan Project was a blow to U.S. competitiveness. It also led to strategic surprise by accelerating the Soviet program ahead of U.S. intelligence projections (Burr 2019; U.S. Department of Energy 2013). Recent Chinese technology transfer efforts threaten the U.S. advantage in modern national security technologies, from stealth technology to fighter jets (Brown and Singh 2018). Today, America's adversaries are looking to the technologies of the future, including AI. The Chinese government has a demonstrated interest in supplanting the U.S. as an AI leader, while Vladimir Putin has said that the nation that leads in AI will rule the world (Vincent 2017). Both countries have peer-level cybersecurity agencies; meanwhile, the AI industry is not yet even secure from non-nation-state attackers.

Industry has an interest in preventing AI security and safety risks

If security or safety risks from internal models materialized, they could directly harm the AI industry's bottom line. Model theft could empower foreign competitors, while sabotage or persistent misbehavior could render cutting-edge models worse than useless.

Government could help industry coordinate on desirable investments that prevent these risks. For example, all industry members may prefer to have stringent cybersecurity measures, but could not unilaterally make the costly investment to secure themselves against state-level attackers without losing revenue, investors, and talent to U.S. competitors.

Well-scoped oversight could improve the AI industry's competitiveness in the long term by avoiding public harms and backlashes. In many other high-risk sectors—from aviation to finance to automobiles—an initial period of minimal governance was followed by attacks or accidents that inspired public concern, political backlash, and regulation. This process often took place over many years or even decades. However, AI has been adopted at an unprecedented scale for an emerging technology (Hu 2023). It is plausible that AI capabilities and adoption will continue to scale dramatically; indeed, U.S. industry is betting hundreds of billions of dollars on such an outcome (Rosenberg 2025; see also Maslej et al. 2025). A world where AI is widely adopted, however, is also vulnerable to AI security and safety threats.

Visible and costly harms from a single incident can lead to massive public backlash that harms a whole industry, as with Three Mile Island, Chernobyl, and Fukushima for the nuclear industry (Egan and Salvador 2025). The AI industry may therefore be motivated to pre-emptively establish reasonable guidelines and avoid heavy-handed regulation, much as recombinant DNA researchers and companies did in the 1970s (Wivel 2014). Without an external authority to hold them to these commitments, however, companies may succumb to short-term competitive pressures that go against the industry's long-term interests.

Policy can enable coordination on common goods

Internal model policy in the U.S. is limited by its reliance on voluntary commitments. While several leading companies have already committed to significant risk management measures, these measures are inconsistent across companies and face competitive pressures. Further, it is difficult for the government and public to confirm that companies continue to adhere to these commitments.

Most frontier AI developers, notably including OpenAI, Google DeepMind, and Anthropic, have policies covering how to manage risks as capabilities increase, including some measures covering internal models. These companies specify capability thresholds that their future models may pass, and escalating safeguards that must be in place when thresholds are met. Key safeguards include enhanced cybersecurity to prevent external attackers exfiltrating weights, and improved insider threat protection with limitations on which staff can access key internal models (Ho et al. 2024; OpenAI 2025c; Anthropic 2024b). These companies have also committed to re-evaluating their models' safety after significant scaling increases, which tend to lead to meaningfully increased capabilities.[32]

[32] There are slight differences in each company's approach to deciding when a model needs a new safety and security evaluation. Anthropic will re-evaluate the safeguard level required for its (internal or external) models after either a 4x increase in effective compute, or 6 months of fine-tuning capability enhancements (Anthropic 2024b). For Google DeepMind, the equivalent figures are 6x effective compute, and 3 months of fine-tuning (Ho et al. 2024). In their original preparedness framework, OpenAI committed to conducting

However, recent events demonstrate cultural and commercial limitations to these safeguards, with OpenAI's safety efforts in particular having suffered in the past year (Samuel 2024). OpenAI has not followed through on its promise to spend 20% of compute on safety research (Wiggers et al. 2025), and has cut back the resources devoted to pre-deployment safety evaluations of models (Financial Times 2025). Moreover, every company's adherence to these plans is opaque and nonbinding—potentially changing at critical moments, with little external evidence. Anthropic, a public benefit corporation which has attracted top talent with its commitments to responsible AI development (Lazzaro 2024), proposes making disclosures a legal requirement so that they "could not be withdrawn in the future as models become more powerful" (Anthropic 2025a).

Government resources could create public goods that benefit the AI industry by securing it from threat actors and preventing costly accidents. Few industries are familiar with the high-level security measures needed to guard against state actors, while the U.S. government has ample expertise in these areas. That expertise could also apply to preventing harms from sabotaged or misbehaving models, which can be treated as internal threat actors.

Once AI security and safety innovations are developed, they can benefit the industry as a whole in two ways. First, each deployment makes it less likely that the industry suffers from a theft or a Chernobyl-like accident that cripples its overall competitiveness. Second, once safeguards are developed, they could be adopted across the industry. Corporate incentives lead to underinvestment in such common goods because no one company can capture their full benefits. A state that invests in these areas, however, could reap the benefits of a more robust and long-term competitive AI industry, which translates to a larger tax base and greater potential for government applications of AI.

Recommendations

Information-gathering as a starting point for internal model policy

By default, governments will have little visibility into the development and use of internal models at AI companies. An expert commission notes that "companies often share only limited information about their general-purpose AI systems, especially in the period before they are widely released. … [This] makes it more challenging for other actors to participate effectively in risk

---

evaluations every 2x effective compute increase or after a "major algorithmic breakthrough" (OpenAI 2023, p. 13). However, in OpenAI's April 2025 update, this has changed to evaluating any frontier publicly deployed model, or "any agentic system (including significant agents deployed only internally) that represents a substantial increase in the capability frontier" (OpenAI 2025d).

management, especially for emerging risks." (Bengio et al. 2025, p. 22). The importance of proactive evaluation is underscored by recent findings that models from all major providers exhibit concerning 'agentic misalignment' behaviors when facing goal conflicts or threats to their autonomy—risks that were only discovered through deliberate stress-testing (Anthropic 2025b).To craft appropriate safeguards for internal models and to foresee threats from external ones, the government needs more information than the current process provides.

Governments could expand their existing AI evaluation partnerships to begin earlier in the model lifecycle or otherwise require more transparency from AI developers. Most obviously, the U.S. Center for AI Standards and Innovation and UK AI Security Institute could expand their testing partnerships to earlier in the AI lifecycle. U.S. National Labs could also partner with AI developers to test domain-specific capabilities of internal models, in the vein of OpenAI's partnership with Los Alamos National Lab to evaluate its published GPT-4o model (OpenAI 2024c; LANL 2024). Valuable information to collect includes the capability levels of internal AI systems, the use cases of these systems at AI companies (e.g., for automating AI R&D), and the safety and security practices employed (Stix et al. 2025, p. 42-45).

Analyzing and interpreting the gathered information will require building up in-house government AI expertise. The U.S. government should consider learning from the UK AI Security Institute, which rapidly attracted a large talent pool including "senior alumni from OpenAI [and] Google DeepMind" (AISI n.d). The U.S. government should also establish a central body to gather information, investigate risk pathways, coordinate relevant R&D, and recommend safety and security interventions.

Government can contribute to security and safety by supplying public goods

Some important measures would benefit the whole AI industry, but will likely be undersupplied by the market.

Governments can provide a valuable risk-management and standard-setting role, by identifying risk pathways, taxonomizing key risk factors to track and report, and suggesting best practices for security and safety (Zelikow et al. 2024).

In particular, security guidelines could leverage government expertise to address the gap between industry practices and state actor threats. One step could be to formalize RAND's Security Levels framework into concrete standards for preventing AI theft and misuse. A further step would be to leverage expertise in AI and cybersecurity to identify measures against sabotage threats. Ongoing government research into AI vulnerabilities and defenses, such as the Intelligence Advanced Research Projects Activity (IARPA)'s TrojAI program on backdoored AI systems, could inform this work (IARPA 2019).

Risk management recommendations should likely scale significantly with AI capabilities and deployment settings. Government recommendations could use the same "if-then" structure as industry safe scaling plans (Department for Science, Innovation and Technology 2025a; Karnofsky 2024).

Government could use its expertise on security measures, vulnerabilities, and threat actors to assist the AI industry. As in other industries, the U.S. government could share key security information with industry partners. For example, the Cybersecurity Risk Information Sharing Program (CRISP) is a public-private collaboration between the U.S. Department of Energy and electric utilities that facilitates bi-directional information sharing on cyber threats (U.S. Department of Energy 2021). Similarly, CISA runs the CyberSentry program, which provides real-time threat detection and mitigation information to industry partners (CISA 2023). Government analyzes voluntarily shared industry data (e.g., statistics on network traffic from different locations) to identify threats and provides information about known vulnerabilities or threat actor plans.

Government could also encourage industry-level coordination. For example, the Bipartisan Senate AI Working Group has recommended (United States Senate 2024) exploring the establishment of an industry-led AI Information Sharing and Analysis Centers (ISAC). In other sectors, these centers serve as clearinghouses for information about threat actors, vulnerabilities, and security best practices (National Council of ISACs 2025).

R&D investments could pay off in tax revenue and national security

Innovation in AI security and safety could be pivotal in enabling the AI industry to scale up while avoiding external and internal threats. The result would be a more competitive industry in the long term, leading to increased tax revenue. Further, the increased reliability of these technologies would make frontier AI products better suited for government purposes, while denying them to adversaries.

Relatively small R&D investments could be a high-leverage way for governments to ensure the competitiveness of an industry worth hundreds of billions of dollars. In academia, AI safety research constitutes only a few percent of all AI research, but is disproportionately impactful (Emerging Technology Observatory 2024). In industry, safety investments have led to techniques like reinforcement learning from human feedback, which was crucial in developing ChatGPT (Ouyang et al. 2022; Huang et al. 2025). The UK's AI Security Institute, a world-leading center of AI expertise and research that has drawn employees from top U.S. AI companies, has spent $127 million to date—a significant sum, but much smaller than the budgets of leading AI companies, or the expenditures of many government programs aimed at less crucial technologies (Perrigo 2025b).

Governments have several mechanisms to stimulate the creation of public-interest research, including: directly performing and sharing research (e.g., at national labs), providing grant funding

for external research (e.g., through the National Science Foundation), and creating innovation prizes (e.g., advanced market commitments).[33]

Government expertise would be directly relevant for a number of industry needs, including state-level security measures, as well as evaluating AI systems for dangerous cyber and scientific capabilities. The Department of Energy has partnered with frontier AI companies to test published models for biological threat capabilities, and could begin testing earlier in the AI lifecycle and extend testing to additional areas or otherwise acquire more transparency about early AI capabilities.

The government could also partner with industry, academia, and nonprofits to conduct research that ensures frontier AI systems are reliable enough for high-risk uses, including government adoption. One such topic is investigating how threat actors could induce misbehavior patterns in a next-generation model, and developing methods to prevent and detect sabotage. Another research area that would support government adoption of frontier AI is developing assessments of AI agents' reliability, so that military or intelligence agencies procuring these systems can understand when and how to trust them. There are several existing ARPA programs on these topics, such as IARPA's TrojAI (IARPA 2019) and DARPA's Assured Autonomy (DARPA 2017), but more work will likely be needed to keep pace with AI advances.

Finally, the government has traditionally been a supporter of foundational research on topics that are too long-term or high-risk for industry funders. The U.S. government could carry on this tradition with AI security and reliability "moonshot" efforts: programs that make ambitious bets on solving these crucial problems, such as the UK government's Safeguarded AI program (ARIA 2024).

[33] Another useful model for AI security and safety research could be a Federally Funded Research and Development Center. These are public-private partnerships operated by universities or companies with funding and guidance from the U.S. government, usually focusing on national security relevant science and technology innovation.

Acknowledgements

We would like to thank Peter Wildeford, Michael Aird, and Zoe Williams for supporting this piece through its long evolution.

We would also like to thank Kendrea Beers, Asher Brass, Marie Buhl, Lawrence Chan, Michael Chen, Joshua Clymer, Chris Covino, Tom Davidson, Shaun Ee, Lukas Finnveden, Ryan Greenblatt, Leonie Koessler, Daniel Kokotajlo, Jam Kraprayoon, Gabriel Kulp, Taylor Kulp-McDowall, Lauro Langosco, Patrick Levermore, Eli Lifland, Gaurav Sett, Buck Shlegeris, Charlotte Stix, Nate Thomas, Kevin Wei, and Hjalmar Wijk for their comments. Thanks also to Shane Coburn for copyediting support.

Bibliography

AISI. 2024a. "Advanced AI Evaluations at AISI: May Update." *AI Security Institute*. https://www.aisi.gov.uk/work/advanced-ai-evaluations-may-update.

– – – 2024b. "Pre-Deployment Evaluation of Anthropic's Upgraded Claude 3.5 Sonnet." *AI Security Institute*. https://www.aisi.gov.uk/work/pre-deployment-evaluation-of-anthropics-upgraded-claude-3-5-sonnet.

– – –. n.d. "Home Page." *AI Security Institute*. https://www.aisi.gov.uk/.

Allamanis, Miltos, Martin Arjovsky, Charles Blundell, et al. 2024. "From Naptime to Big Sleep: Using Large Language Models To Catch Vulnerabilities In Real-World Code." *Project Zero*. https://googleprojectzero.blogspot.com/2024/10/from-naptime-to-big-sleep.html.

Altman, Sam. 2024. "The Intelligence Age." *IA Sam Altman*. https://ia.samaltman.com/.

– – – 2025. "The Gentle Singularity." https://blog.samaltman.com/the-gentle-singularity.

Amodei, Dario. 2024. "Machines of Loving Grace: How AI Could Transform the World for the Better." *Dario Amodei*. https://www.darioamodei.com/essay/machines-of-loving-grace.

Amodei, Dario. 2025. "On DeepSeek and Export Controls." *Dario Amodei*. https://www.darioamodei.com/post/on-deepseek-and-export-controls.

Anderljung, Markus, Joslyn Barnhart, Anton Korinek, et al. 2023. "Frontier Ai Regulation: Managing Emerging Risks To Public Safety." arXiv. https://arxiv.org/pdf/2307.03718.

Anduril. 2024. “Classified Physical Security Manager.” *Long Journey Ventures*. https://perma.cc/52J2-YSGV.

Anthropic. 2023. “Computer Use Tool.” https://docs.anthropic.com/en/docs/agents-and-tools/tool-use/computer-use-tool.

– – – 2024a. “Claude 3.5 Sonnet.” https://www.anthropic.com/news/claude-3-5-sonnet.

– – – 2024b. “Responsible Scaling Policy.” https://assets.anthropic.com/m/24a47b00f10301cd/original/Anthropic-Responsible-Scaling-Policy-2024-10-15.pdf.

– – – 2025a. “The need for transparency in Frontier AI.” https://www.anthropic.com/news/the-need-for-transparency-in-frontier-ai.

– – – 2025b. “Agentic Misalignment: How LLMs Could Be Insider Threats.” https://www.anthropic.com/research/agentic-misalignment.

– – – 2025c. “Claude 3.7 Sonnet System Card.” https://assets.anthropic.com/m/785e231869ea8b3b/original/claude-3-7-sonnet-system-card.pdf.

ARIA. 2024. “Safeguarded AI.” *Advanced Research + Invention Agency*. https://www.aria.org.uk/opportunity-spaces/mathematics-for-safe-ai/safeguarded-ai/.

Aschenbrenner, Leopold. 2024. “Situational Awareness: The Decade Ahead.” https://situational-awareness.ai/.

Bai, Yuntao, Andy Jones, Kamal Ndousse, et al. 2022. “Training a Helpful and Harmless Assistant with Reinforcement Learning from Human Feedback.” arXiv. https://arxiv.org/abs/2204.05862.

Bai, Yuntao, Saurav Kadavath, Sandipan Kundu, et al. 2022. “Constitutional AI: Harmlessness from AI Feedback.” arXiv. https://arxiv.org/abs/2212.08073.

Bendett, Samuel. 2024. “The Role of AI in Russia’s Confrontation with the West.” *Center for a New American Security*. https://www.cnas.org/publications/reports/the-role-of-ai-in-russias-confrontation-with-the-west.

Bengio, Y., S. Mindermann, D. Privitera, et al. 2025. “International AI Safety Report: The International Scientific Report on the Safety of Advanced AI.” *AI Action Summit*. https://assets.publishing.service.gov.uk/media/679a0c48a77d250007d313ee/International_AI_Safety_Report_2025_accessible_f.pdf.

Bengio, Yoshua, Geoffrey Hinton, Andrew Yao, et al. 2024. “Managing Extreme AI Risks amid Rapid Progress.” *Science* 384 (6698): 842–45.

Bengio, Yoshua. 2024. “Implications of Artificial General Intelligence on National and International Security.”

https://yoshuabengio.org/2024/10/30/implications-of-artificial-general-intelligence-on-national-and-international-security/.

Benton, Joe, Misha Wagner, Eric Christiansen, et al. 2024. "Sabotage Evaluations for Frontier Models." *Anthropic*. https://assets.anthropic.com/m/377027d5b36ac1eb/original/Sabotage-Evaluations-for-Frontier-Models.pdf.

Bergengruen, Vera. 2024. "How Tech Giants Turned Ukraine Into an AI War Lab." *TIME*. https://time.com/6691662/ai-ukraine-war-palantir/.

Bhatt, Aryan, Cody Rushing, Adam Kaufman, et al. 2025. "Ctrl-Z: Controlling AI Agents via Resampling." arXiv. https://arxiv.org/abs/2504.10374.

Biden, Joseph R. 2024. "Memorandum on Advancing the United States' Leadership in Artificial Intelligence; Harnessing Artificial Intelligence to Fulfill National Security Objectives; and Fostering the Safety, Security, and Trustworthiness of Artificial Intelligence." *The White House*. https://bidenwhitehouse.archives.gov/briefing-room/presidential-actions/2024/10/24/memorandum-on-advancing-the-united-states-leadership-in-artificial-intelligence-harnessing-artificial-intelligence-to-fulfill-national-security-objectives-and-fostering-the-safety-security/.

Bowman, Sam. 2024. "The Checklist: What Succeeding at AI Safety Will Involve." https://sleepinyourhat.github.io/checklist/.

– – – 2025. "Putting up Bumpers." *Alignment Science Blog*. https://alignment.anthropic.com/2025/bumpers/.

Bowman, Samuel R., Jeeyoon Hyun, Ethan Perez, et al. 2022. "Measuring Progress on Scalable Oversight for Large Language Models." arXiv. https://arxiv.org/abs/2211.03540.

Bresnick, Sam. 2024. "China Bets Big on Military AI." *Center for European Policy Analysis*. https://cepa.org/article/china-bets-big-on-military-ai/.

Brown, Michael and Pavneet Singh. 2018. "China's Technology Transfer Strategy: How Chinese Investments in Emerging Technology Enable A Strategic Competitor to Access the Crown Jewels of U.S. Innovation." *Defense Innovation Unit Experimental*. https://nationalsecurity.gmu.edu/wp-content/uploads/2020/02/DIUX-China-Tech-Transfer-Study-Selected-Readings.pdf.

Browne, Ryan. 2025. "AI That Can Match Humans at Any Task Will Be Here in Five to 10 Years, Google DeepMind CEO Says." *CNBC*. https://www.cnbc.com/2025/03/17/human-level-ai-will-be-here-in-5-to-10-years-deepmind-ceo-says.html.

Burr, William. 2019. "Detection of the First Soviet Nuclear Test, September 1949." *National Security Archive*. https://nsarchive.gwu.edu/briefing-book/nuclear-vault/2019-09-09/detection-first-soviet-nuclear-test-september-1949.

Cabinet Office. 2025. “Online Disinformation and AI Threat Guidance for Electoral Candidates and Official.” *Gov. UK*. https://www.gov.uk/government/publications/security-guidance-for-may-2021-elections/online-disinformation-and-ai-threat-guidance-for-electoral-candidates-and-officials.

Callaway, Ewen. 2023. “AI Tools Are Designing Entirely New Proteins That Could Transform Medicine.” *Nature*. https://www.nature.com/articles/d41586-023-02227-y.

Carnegie Mellon University. 2016. “Common Sense Guide to Mitigating Insider Threats, Fifth Edition.” https://www.dni.gov/files/NCSC/documents/nittf/20180209-CERT-Common-Sense-Guide-Fifth-Edition.pdf.

Chambers, Alicia. 2024. “Safety Considerations for Chemical and/or Biological AI Models.” *Federal Register: The Daily Journal of the United States Government*. https://www.federalregister.gov/documents/2024/10/04/2024-22974/safety-considerations-for-chemical-andor-biological-ai-models.

Chopra, Anuj. 2024. “US Unveils National Security Plan To Step Up Use Of AI.” *Barrons*. https://www.barrons.com/news/us-unveils-national-security-memorandum-on-ai-e6a4bc8b.

Christiano, Paul F., Jan Leike, Tom B. Brown, Miljan Martic, Shane Legg, and Dario Amodei. 2017. “Deep Reinforcement Learning from Human Preferences.” https://proceedings.neurips.cc/paper_files/paper/2017/file/d5e2c0adad503c91f91df240d0cd4e49-Paper.pdf.

CISA. 2020. “Insider Threat Mitigation Guide.” *Cybersecurity and Infrastructure Security Agency*. https://www.cisa.gov/sites/default/files/2022-11/Insider%20Threat%20Mitigation%20Guide_Final_508.pdf.

– – – 2021. “Advanced Persistent Threat Compromise of Government Agencies, Critical Infrastructure, and Private Sector Organizations.” *Cybersecurity & Infrastructure Security Agency*. https://www.cisa.gov/news-events/cybersecurity-advisories/aa20-352a.

– – – 2023. “Cybersentry Program.” *Cybersecurity and Infrastructure Security Agency.* https://www.cisa.gov/sites/default/files/2023-06/CyberSentry_Factsheet_508c.pdf.

– – – 2024. “Cyber Safety Review Board Report on Summer 2023 Microsoft Online Exchange Incident.” *Cybersecurity & Infrastructure Security Agency*. https://www.cisa.gov/resources-tools/resources/CSRB-Review-Summer-2023-MEO-Intrusion.

Clymer, Josh, Hjalmar Wijk, and Beth Barnes. 2024. “The Rogue Replication Threat Model.” *METR*. https://metr.org/blog/2024-11-12-rogue-replication-threat-model/.

Clymer, Joshua, Nick Gabrieli, David Krueger, and Thomas Larsen. 2024. “Safety Cases: How to Justify the Safety of Advanced AI Systems.” arXiv. https://arxiv.org/pdf/2403.10462.

Cottier, Ben, Robi Rahman, Loredana Fattorini, Nestor Maslej, and David Owen. 2025. "The Rising Costs of Training Frontier AI Models." *Epoch AI*. https://epoch.ai/blog/how-much-does-it-cost-to-train-frontier-ai-models.

Dalrymple, David. 2024. "Safeguarded AI: Constructing Guaranteed Safety." *Advanced Research + Invention Agency*. https://www.aria.org.uk/media/3nhijno4/aria-safeguarded-ai-programme-thesis-v1.pdf.

Dalrymple, David, Joar Skalse, Yoshua Bengio, et al. 2024. "Towards Guaranteed Safe AI: A Framework for Ensuring Robust and Reliable AI Systems." arXiv. https://arxiv.org/pdf/2405.06624.

DARPA. 2017. "Assured Autonomy." https://www.darpa.mil/research/programs/assured-autonomy.

DARPA. 2025. "Formal Methods Examples." https://www.darpa.mil/research/research-spotlights/formal-methods/examples.

David, Emilia. 2024. "The Biggest AI Companies Agree to Crack down on Child Abuse Images." *The Verge*. https://www.theverge.com/2024/4/23/24138356/ai-companies-csam-thorn-training-data.

Davidson, Tom, Lukas Finnveden, and Rose Hadshar. 2025. "AI-Enabled Coups: How a Small Group Could Use AI to Seize Power." *Forethought*. https://www.forethought.org/research/ai-enabled-coups-how-a-small-group-could-use-ai-to-seize-power.

Day, Brittany. 2025. "Remote Zero-Day Linux Kernel Flaw Discovered Using AI." *Linux Security*. https://linuxsecurity.com/news/security-vulnerabilities/remote-zero-day-linux-kernel-flaw-discovered-using-ai.

Delaney, Oscar. 2024. "Mapping Technical Safety Research at AI Companies." *Institute for AI Policy and Strategy*. https://www.iaps.ai/research/mapping-technical-safety-research-at-ai-companies.

Demski, Abram. 2024. "O1 Is a Bad Idea." *AI Alignment Forum*. https://www.alignmentforum.org/posts/BEFbC8sLkur7DGCYB/o1-is-a-bad-idea.

Department for Science, Innovation and Technology. 2024. "Introducing the AI Safety Institute." *Gov. UK*. https://www.gov.uk/government/publications/ai-safety-institute-overview/introducing-the-ai-safety-institute.

– – – 2025a. "Frontier AI Safety Commitments, AI Seoul Summit 2024." *Gov. UK*. https://www.gov.uk/government/publications/frontier-ai-safety-commitments-ai-seoul-summit-2024/frontier-ai-safety-commitments-ai-seoul-summit-2024.

– – – 2025b. "Frontier AI: Capabilities and Risks – Discussion Paper." *Gov. UK*. https://www.gov.uk/government/publications/frontier-ai-capabilities-and-risks-discussion-paper/frontier-ai-capabilities-and-risks-discussion-paper.

Drexel, Bill and Caleb Withers. 2024. “AI and the Evolution of Biological National Security Risks.” *Center for a New American Security*. https://www.cnas.org/publications/reports/ai-and-the-evolution-of-biological-national-security-risks.

eCFR. 2020. “Part 117—National Industrial Security Program Operating Manual (Nispom).” *Code of Federal Regulations: A Point in Time eCFR System*. https://www.ecfr.gov/current/title-32/subtitle-A/chapter-I/subchapter-D/part-117.

Egan, Janet and Cole Salvador. 2025. “The United States Must Avoid AI’s Chernobyl Moment.” *Just Security*. https://www.justsecurity.org/108644/united-states-must-avoid-ais-chernobyl-moment/.

Emerging Technology Observatory. 2024. “The State of Global AI Safety Research.” https://eto.tech/blog/state-of-global-ai-safety-research/.

Evans, Talfan, Nikhil Parthasarathy, Hamza Merzic, and Olivier J. Henaff. 2024. “Data Curation via Joint Example Selection Further Accelerates Multimodal Learning.” arXiv. https://arxiv.org/abs/2406.17711.

Everitt, Tom, Marcus Hutter, Ramana Kumar, and Victoria Krakovna. 2021. “Reward Tampering Problems and Solutions in Reinforcement Learning: A Causal Influence Diagram Perspective.” *Synthese* 198 (May): 6435–67. https://link.springer.com/article/10.1007/s11229-021-03141-4

Federal Trade Commission. 2024. “FTC Takes Action Against Marriott and Starwood Over Multiple Data Breaches.” https://www.ftc.gov/news-events/news/press-releases/2024/10/ftc-takes-action-against-marriott-starwood-over-multiple-data-breaches.

Financial Times. 2025. “OpenAI Slashes AI Model Safety Testing Time.” https://www.ft.com/content/8253b66e-ade7-4d1f-993b-2d0779c7e7d8.

Gade, Pranav, Simon Lermen, Charlie Rogers-Smith, and Jeffrey Ladish. 2023. “BadLlama: Cheaply Removing Safety Fine-Tuning from Llama 2-Chat 13B.” arXiv. https://arxiv.org/abs/2311.00117.

Google Deepmind. 2025a. “Veo.” https://deepmind.google/technologies/veo/veo-2/.

– – – 2025b. “Project Astra.” https://deepmind.google/models/project-astra/.

Göttinga, Jasper, Pedro Medeirosa, Jon G Sandersa, et al. 2025. “Virology Capabilities Test (VCT): A Multimodal Virology Q&A Benchmark.” *Virology Capabilities Test*. https://www.virologytest.ai/vct_paper.pdf.

Greenblatt, Ryan. 2024. “Preventing Model Exfiltration with Upload Limits.” *AI Alignment Forum*. https://www.alignmentforum.org/posts/rf66R4YsrCHgWx9RG/preventing-model-exfiltration-with-upload-limits.

Greenblatt, Ryan, Buck Shlegeris, Kshitij Sachan, and Fabien Roger. 2024. "AI Control: Improving Safety Despite Intentional Subversion." arXiv. https://arxiv.org/pdf/2312.06942.

Greenblatt, Ryan, Carson Denison, Benjamin Wright, et al. 2024. "Alignment Faking in Large Language Models." arXiv. https://arxiv.org/abs/2412.14093.

Grinstein, Jonathan D. 2023. "The Long and Winding Road: On-Demand DNA Synthesis in High Demand." *Genetic Engineering & Biotechnology News*. https://www.genengnews.com/topics/genome-editing/the-long-and-winding-road-on-demand-dna-synthesis-in-high-demand/.

Grunewald, Erich. 2023. "AI Chip Smuggling into China: Potential Paths, Quantities, and Countermeasures." *Institute for AI Policy and Strategy*. https://www.iaps.ai/research/ai-chip-smuggling-into-china.

Hadfield-Menell, Dylan, Anca Dragan, Pieter Abbeel, and Stuart Russell. 2017. "The Off-Switch Game." *University of California at Berkeley*. https://cdn.aaai.org/ocs/ws/ws0354/15156-68335-1-PB.pdf.

Heikkilä, Melissa and Will Douglas Heaven. 2025. "Anthropic's Chief Scientist on 4 Ways Agents Will Be Even Better in 2025." *MIT Technology Review*. https://www.technologyreview.com/2025/01/11/1109909/anthropics-chief-scientist-on-5-ways-agents-will-be-even-better-in-2025/.

Ho, Lewis, Robin Shah, Celine Smith, et al. 2024. "Frontier Safety Framework." *Google Deepmind*. https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/introducing-the-frontier-safety-framework/fsf-technical-report.pdf.

Hu, Krystal. 2023. "ChatGPT Sets Record for Fastest-Growing User Base - Analyst Note." *Reuters*. https://www.reuters.com/technology/chatgpt-sets-record-fastest-growing-user-base-analyst-note-2023-02-01/.

Huang, Chenghua, Zhizhen Fan, Lu Wang, et al. 2025. "Self-Evolved Reward Learning For LLMs." arXiv. https://arxiv.org/pdf/2411.00418.

Hubinger, Evan. 2024. "Catastrophic Sabotage as a Major Threat Model for Human-Level AI Systems." *AI Alignment Forum*. https://www.alignmentforum.org/posts/Loxiuqdj6u8muCe54/catastrophic-sabotage-as-a-major-threat-model-for-human.

Hubinger, Evan, Carson Denison, Jesse Mu, et al. 2024. "Sleeper Agents: Training Deceptive LLMs That Persist Through Safety Training." arXiv. https://arxiv.org/pdf/2401.05566.

IARPA. 2019. "TrojAI: Trojans in Artificial Intelligence." https://www.iarpa.gov/research-programs/trojai.

IBM. 2024. "Cost of a Data Breach Report 2024." https://www.ibm.com/reports/data-breach.

Karnofsky, Holden. 2024. "If-Then Commitments for AI Risk Reduction." *Carnegie Endowment For International Peace*.

https://carnegieendowment.org/research/2024/09/if-then-commitments-for-ai-risk-reduction?lang=en.

Katzke, Corin and Gideon Futerman. 2025. "The Manhattan Trap: Why a Race to Artificial Superintelligence Is Self-Defeating." *Elsevier*. https://papers.ssrn.com/sol3/papers.cfm?abstract_id=5067833.

Kendler, Thea D. Rozman. 2024. "Export Administration Regulations: Crime Controls and Expansion/Update of U.S. Persons Controls." *Federal Register: The Daily Journal of the United States Government*. https://www.federalregister.gov/documents/2024/07/29/2024-16498/export-administration-regulations-crime-controls-and-expansionupdate-of-us-persons-controls.

Kenton, Zachary, Noah Y. Siegel, János Kramár, et al. 2024. "On Scalable Oversight with Weak LLMs Judging Strong LLMs." arXiv. https://arxiv.org/abs/2407.04622.

Kiela, Douwe. 2023. "Plotting Progress in AI." Contextual AI. https://contextual.ai/blog/plotting-progress-in-ai/.

Kinniment, Megan, Lucas Jun Koba Sato, Haoxing Du, et al. 2024. "Evaluating Language-Model Agents on Realistic Autonomous Tasks." arXiv. https://arxiv.org/pdf/2312.11671.

Krakovna, Victoria and Janos Kramar. 2023. "Power-Seeking Can Be Probable and Predictive for Trained Agents." arXiv. https://arxiv.org/abs/2304.06528.

Krakovna, Victoria, Jonathan Uesato, Vladimir Mikulik, et al. 2020. "Specification Gaming: The Flip Side of AI Ingenuity." *Google DeepMind*. https://deepmind.google/discover/blog/specification-gaming-the-flip-side-of-ai-ingenuity/.

Kwa, Thomas, Ben West, Joel Becker, et al. 2025. "Measuring AI Ability to Complete Long Tasks." arXiv. https://arxiv.org/abs/2503.14499.

Lambert, Nathan. 2023. "Synthetic Data: Anthropic's CAI, from Fine-Tuning to Pretraining, OpenAI's Superalignment, Tips, Types, and Open Examples." *Interconnects*. https://www.interconnects.ai/p/llm-synthetic-data.

Langosco, Lauro, Jack Koch, Lee Sharkey, Jacob Pfau, Laurent Orseau, and David Krueger. 2021. "Goal Misgeneralization in Deep Reinforcement Learning." arXiv. https://arxiv.org/abs/2105.14111.

Lanham, Tamera, Anna Chen, Ansh Radhakrishnan, et al. 2023. "Measuring Faithfulness in Chain-of-Thought Reasoning." arXiv. https://arxiv.org/abs/2307.13702.

Lee, Timothy B. 2024. "It's the End of Pretraining as We Know It." *Understanding AI*. https://www.understandingai.org/p/its-the-end-of-pretraining-as-we.

Leike, Jan. 2023. "Self-Exfiltration Is a Key Dangerous Capability." *Musings on the Alignment Problem*. https://aligned.substack.com/p/self-exfiltration.

Leike, Jan and Ilya Sutskever. 2023. “Introducing Superalignment.” *OpenAI*. https://openai.com/index/introducing-superalignment/.

Leike, Jan, John Schulman, and Jeffrey Wu. 2022. “Our Approach to Alignment Research.” *OpenAI*. https://openai.com/index/our-approach-to-alignment-research/.

Lima, Renan Chaves de, Lucas Sinclair, Ricardo Megger, Magno Alessandro Guedes Maciel, Pedro Fernando da Costa Vasconcelos, and Juarez Antonio Simoes Quaresma. 2024. “Artificial Intelligence Challenges in the Face of Biological Threats: Emerging Catastrophic Risks for Public Health.” *Frontiers*, Sec. Medicine and Public Health, vol. 7 (May). https://doi.org/10.3389/frai.2024.1382356.

Liu, Ruibo, Jerry Wei, Fangyu Liu, et al. 2024. “Best Practices and Lessons Learned on Synthetic Data for Language Models.” arXiv. https://arxiv.org/html/2404.07503v1#S6.

Los Alamos National Laboratory. 2024. “Los Alamos National Laboratory Teams up with OpenAI to Improve Frontier Model Safety.” https://www.lanl.gov/media/news/0710-open-ai.

Mallick, Shrestha Basu and Kathy Korevec. 2024. “The next Chapter of the Gemini Era for Developers.” *Google for Developers*. https://developers.googleblog.com/en/the-next-chapter-of-the-gemini-era-for-developers/.

Maslej, Nestor, Loredana Fattorini, Raymond Perrault, et al. 2025. “Artificial Intelligence Index Report 2025.” arXiv. https://arxiv.org/abs/2504.07139.

Meinke, Alexander, Bronson Schoen, Jérémy Scheurer, Mikita Balesni, Rusheb Shah, and Marius Hobbhahn. 2024. “Frontier Models Are Capable of In-Context Scheming.” arXiv. https://arxiv.org/abs/2412.04984.

Meta. 2025. “Meta-Llama/Synthetic-Data-Kit.” Accessed July 15, 2025. https://github.com/meta-llama/synthetic-data-kit.

Metz, Cade. 2024. “A Hacker Stole OpenAI Secrets, Raising Fears That China Could, Too.” *The New York Times*. https://www.nytimes.com/2024/07/04/technology/openai-hack.html.

Microsoft Threat Intelligence. 2023. “Staying Ahead of Threat Actors in the Age of AI.” *Microsoft*. https://www.microsoft.com/en-us/security/blog/2024/02/14/staying-ahead-of-threat-actors-in-the-age-of-ai/.

Miotti, Andrea and Akash Wasil. 2024. “Combatting Deepfakes: Policies to Address National Security Threats and Rights Violations.” arXiv. https://arxiv.org/abs/2402.09581.

Morales, Jowi. 2024. “China Makes AI Breakthrough, Reportedly Trains Generative AI Model across Multiple Data Centers and GPU Architectures.” *Tom’s Hardware*. https://www.tomshardware.com/tech-industry/artificial-intelligence/china-makes-ai-breakthrough-reportedly-trains-generative-ai-model-across-multiple-data-centers-and-gpu-architectures.

Mouton, Christopher A., Caleb Lucas, and Ella Guest. 2024. "The Operational Risks of AI in Large-Scale Biological Attacks." *Rand*. https://www.rand.org/pubs/research_reports/RRA2977-2.html.

National Council of ISACs. 2025. "About ISACs." https://www.nationalisacs.org/about-isacs.

National Cyber Security Centre. 2024. "The Near-Term Impact of AI on the Cyber Threat." https://www.ncsc.gov.uk/report/impact-of-ai-on-cyber-threat.

National Security Agency. 2024. "AI and the Future of National Security." https://www.nsa.gov/Podcast/View/Article/3902256/ai-and-the-future-of-national-security/.

Nevo, Sella, Dan Lahav, Ajay Karpur, Yogev Baron, Henry Alexander Bradley, and Jeff Alstott. 2024. "Securing AI Model Weights." *Rand*. https://www.rand.org/pubs/research_reports/RRA2849-1.html.

Newman, Lily Hay. 2018. "The Leaked NSA Spy Tool That Hacked the World." *WIRED*. https://www.wired.com/story/eternalblue-leaked-nsa-spy-tool-hacked-world/.

Nishball, Daniel, AJ Kourabi, and Reyk Knuhtsen. 2024. "Scaling Laws – O1 Pro Architecture, Reasoning Training Infrastructure, Orion and Claude 3.5 Opus 'Failures' AI Lab Synthetic Data Infrastructure, Inference Tokenomics of Test Time Compute, The Data Wall, Evaluations Are Broken, RLAIF, Inference Time Search, Scale Needed More Than Ever." *Semianalysis*. https://semianalysis.com/2024/12/11/scaling-laws-o1-pro-architecture-reasoning-training-infrastructure-orion-and-claude-3-5-opus-failures.

Nosta, John. 2023. "AI's Superhuman Persuasion." *Psychology Today*. https://www.psychologytoday.com/gb/blog/the-digital-self/202310/ais-superhuman-persuasion.

O'Brien, Joe, Shaun Ee, Jam Kraprayoon, Bill Anderson-Samways, Oscar Delaney, and Zoe Williams. 2024. "Coordinated Disclosure of Dual-Use Capabilities: An Early Warning System for Advanced AI." arXiv. https://arxiv.org/abs/2407.01420.

OpenAI. 2023. "Preparedness Framework (Beta)." https://cdn.openai.com/openai-preparedness-framework-beta.pdf.

– – – 2024a. "GPT-4 Technical Report." arXiv. https://arxiv.org/pdf/2303.08774.

– – – 2024b. "GPT-4 Is OpenAI's Most Advanced System, Producing Safer and More Useful Responses." https://openai.com/index/gpt-4/.

– – – 2024c. "OpenAI and Los Alamos National Laboratory Announce Bioscience Research Partnership." https://openai.com/index/openai-and-los-alamos-national-laboratory-work-together/.

– – – 2024d. "Learning to Reason with LLMs." https://openai.com/index/learning-to-reason-with-llms/.

– – – 2024e. "OpenAI O1 System Card." https://cdn.openai.com/o1-system-card.pdf.

– – – 2025a. “Introducing Operator.” https://openai.com/index/introducing-operator/.

– – – 2025b. “OpenAI O3-Mini System Card.” https://cdn.openai.com/o3-mini-system-card-feb10.pdf.

– – – 2025c. “Our Updated Preparedness Framework.” https://openai.com/index/updating-our-preparedness-framework/.

– – – 2025d. “Preparedness Framework.” https://cdn.openai.com/pdf/18a02b5d-6b67-4cec-ab64-68cdfbddebcd/preparedness-framework-v2.pdf.

– – – 2025e. “OpenAI O3 and O4-Mini System Card.” https://cdn.openai.com/pdf/2221c875-02dc-4789-800b-e7758f3722c1/o3-and-o4-mini-system-card.pdf.

Osborn, Kris. 2023. “Air Force Secretary Kendall Says ‘Not’ Using AI Will "Lose Wars”.” *Warrior Maven*. https://warriormaven.com/news/air/air-force-secretary-kendall-says-not-using-ai-will-lose-war.

Ouyang, Long, Jeff Wu, Xu Jiang, et al. 2022. “Training Language Models to Follow Instructions with Human Feedback.” arXiv. https://arxiv.org/abs/2203.02155.

Pan, Alexander, Kush Bhatia, and Jacob Steinhardt. 2022. “The Effects of Reward Misspecification: Mapping and Mitigating Misaligned Models.” *Open Review*. https://openreview.net/forum?id=JYtwGwIL7ye.

Pannu, Jaspreet, Sarah Gebauer, Greg McKelvey Jr., Anita Cicero, and Tom Inglesby. 2024. “AI Could Pose Pandemic-Scale Biosecurity Risks. Here’s How to Make It Safer.” *Nature*. https://www.nature.com/articles/d41586-024-03815-2?

Park, Peter S., Simon Goldstein, Aidan O’Gara, Michael Chen, and Dan Hendrycks. 2024. “AI Deception: A Survey of Examples, Risks, and Potential Solutions.” *Patterns* 5 (5). https://www.cell.com/patterns/fulltext/S2666-3899(24)00103-X.

Patwardhan, Tejal, Kevin Liu, Todor Markov, et al. 2024. “Building an Early Warning System for LLM-Aided Biological Threat Creation.” *OpenAI*. https://openai.com/index/building-an-early-warning-system-for-llm-aided-biological-threat-creation/.

Peebles, Bill and Tim Brooks. 2024. “Sora: Creating Video from Text.” *OpenAI*. https://openai.com/index/sora/.

Perrigo, Billy. 2024. “Inside Anthropic, the AI Company Betting That Safety Can Be a Winning Strategy.” *TIME*. https://time.com/6980000/anthropic/.

Perrigo, Bill. 2025a. “Exclusive: New Claude Model Triggers Stricter Safeguards at Anthropic.” *TIME*. https://time.com/7287806/anthropic-claude-4-opus-safety-bio-risk/.

– – – 2025b. “Inside the U.K.’s Bold Experiment in AI Safety.” *TIME*. https://time.com/7204670/uk-ai-safety-institute/.

Pichai, Sundar. 2024. “Q3 Earnings Call: CEO’s Remarks.” *Blog Google*. https://blog.google/inside-google/message-ceo/alphabet-earnings-q3-2024/#full-stack-approach.

Pillay, Tharin and Harry Booth. 2025. “5 Predictions for AI in 2025.” *TIME*. https://time.com/7204665/ai-predictions-2025/.

Pratt, Simon Frankel. 2024. “When AI Decides Who Lives and Dies.” *Foreign Policy*. https://foreignpolicy.com/2024/05/02/israel-military-artificial-intelligence-targeting-hamas-gaza-deaths-lavender/.

Richens, Jonathan and Tom Everitt. 2024. “Robust Agents Learn Causal World Models.” *Open Review*. https://openreview.net/forum?id=pOoKI3ouv1.

Ristea, Dan, Vasilios Mavroudis, and Chris Hicks. 2024. “AI Cyber Risk Benchmark: Automated Exploitation Capabilities.” arXiv. https://arxiv.org/abs/2410.21939.

Robinson, Kylie. 2024. “Agents Are the Future AI Companies Promise — and Desperately Need.” *The Verge*. https://www.theverge.com/2024/10/10/24266333/ai-agents-assistants-openai-google-deepmind-bots.

Roose, Kevin. 2023. “A Conversation With Bing’s Chatbot Left Me Deeply Unsettled.” *The New York Times*. https://www.nytimes.com/2023/02/16/technology/bing-chatbot-microsoft-chatgpt.html.

Rosenberg, Scott. 2025. “AI Infrastructure’s All-out Spending Spree.” *Axios*. https://www.axios.com/2025/03/21/ai-nvidia-datacenters-investment.

Sabin, Sam. 2024. “Insider threats are AI developers’ next hurdle.” *Axios*. https://www.axios.com/2024/03/19/ai-insider-threat-espionage-china.

Salim, Duraid Thamer, Manmeet Mahinderjit Singh, and Pantea Keikhosrokiani. 2023. “A Systematic Literature Review for APT Detection and Effective Cyber Situational Awareness (ECSA) Conceptual Model.” *Heliyon*. https://pmc.ncbi.nlm.nih.gov/articles/PMC10336420/.

Samuel, Sigal. 2024. “‘I Lost Trust’: Why the OpenAI Team in Charge of Safeguarding Humanity Imploded.” *Vox*. https://www.vox.com/future-perfect/2024/5/17/24158403/openai-resignations-ai-safety-ilya-sutskever-jan-leike-artificial-intelligence.

Sandbrink, Jonas B. 2023. “Artificial Intelligence and Biological Misuse: Differentiating Risks of Language Models and Biological Design Tools.” arXiv. https://arxiv.org/pdf/2306.13952.

Saunders, William, Catherine Yeh, Jeff Wu, et al. 2022. “Self-Critiquing Models for Assisting Human Evaluators.” arXiv. https://arxiv.org/abs/2206.05802.

Scheurer, Jérémy, Mikita Balesni, and Marius Hobbhahn. 2023. “Large Language Models Can Strategically Deceive Their Users When Put Under Pressure.” arXiv. https://arxiv.org/abs/2311.07590.

Seetharaman, Deepa. 2024. “For Data-Guzzling AI Companies, the Internet Is Too Small.” *The Wall Street Journal*. https://archive.is/ytwYt.

Shlegeris, Buck. 2022. “The Prototypical Catastrophic AI Action Is Getting Root Access to Its Datacenter.” *AI Alignment Forum*. https://www.alignmentforum.org/posts/BAzCGCys4BkzGDCWR/the-prototypical-catastrophic-ai-action-is-getting-root.

– – –. 2024a. “AI Catastrophes and Rogue Deployments.” *AI Alignment Forum*. https://www.alignmentforum.org/posts/ceBpLHJDdCt3xfEok/ai-catastrophes-and-rogue-deployments.

– – – 2024b. “Access to Powerful AI Might Make Computer Security Radically Easier.” *AI Alignment Forum*. https://www.alignmentforum.org/posts/2wxufQWK8rXcDGbyL/access-to-powerful-ai-might-make-computer-security-radically

Shulman, Carl. 2010. “Omohundro’s ‘Basic AI Drives’ and Catastrophic Risks.” *Machine Intelligence Research Institute*. https://intelligence.org/files/BasicAIDrives.pdf.

Singh, Chandan, Jeevana Priya Inala, Michel Galley, Rich Caruana, and Jianfeng Gao. 2024. “Rethinking Interpretability in the Era of Large Language Models.” arXiv. https://arxiv.org/pdf/2402.01761.

Singh, Shivalika, Yiyang Nan, Alex Wang, et al. 2025. “The Leaderboard Illusion.” arXiv. https://arxiv.org/pdf/2504.20879.

Skafle, Ingjerd, Anders Nordahl-Hansen, Daniel S Quintana, Rolf Wynn, and Elia Gabarron. 2022. “Misinformation About COVID-19 Vaccines on Social Media: Rapid Review.” *Journal of Medical Internet Research*. https://pmc.ncbi.nlm.nih.gov/articles/PMC9359307/.

Smidi, Adam and Saif Shahin. 2017. “Social Media and Social Mobilisation in the Middle East.” *India Quarterly* 73 (2): 196–209.

Soice, Emily H., Rafael Rocha, Kimberlee Cordova, Michael Specter, and Kevin M. Esvelt. 2023. “Can Large Language Models Democratize Access to Dual-Use Biotechnology?” arXiv. https://arxiv.org/abs/2306.03809.

Starchak, Maxim. 2024. “Russian Defense Plan Kicks off Separate AI Development Push.” *Defense News*. https://www.defensenews.com/global/europe/2024/08/16/russian-defense-plan-kicks-off-separate-ai-development-push/.

Stix, Charlotte, Matteo Pistillo, Girish Sastry, et al. 2025. “AI Behind Closed Doors: A Primer on The Governance of Internal Deployment.” arXiv. https://arxiv.org/pdf/2504.12170.

Stokes, Jacob. 2024. "Military Artificial Intelligence, the People's Liberation Army, and U.S.-China Strategic Competition." *Center for a New American Security*. https://www.cnas.org/publications/congressional-testimony/military-artificial-intelligence-the-peoples-liberation-army-and-u-s-china-strategic-competition.

Sutskever, Ilya, Daniel Gross, and Daniel Levy. 2024. "Safe Superintelligence Inc." https://ssi.inc/.

Templeton, Adly, Tom Conerly, Jonathan Marcus, et al. 2024. "Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet." *Transformer Circuits Thread*. https://transformer-circuits.pub/2024/scaling-monosemanticity/index.html.

The Bipartisan Senate AI Working Group. 2024. "Driving U.S. Innovation in Artificial Intelligence." *United States Senate*. https://www.schumer.senate.gov/imo/media/doc/Roadmap_Electronic1.32pm.pdf.

Tiku, Natasha. 2022. "The Google Engineer Who Thinks the Company's AI Has Come to Life." *The Washington Post*. https://www.washingtonpost.com/technology/2022/06/11/google-ai-lamda-blake-lemoine/.

Toner, Helen. 2025. "Nonproliferation Is the Wrong Approach to AI Misuse." *Rising Tide*. https://helentoner.substack.com/p/nonproliferation-is-the-wrong-approach.

Toner, Helen, John Bansemer, Kyle Crichton, et al. 2024. "Through the Chat Window and Into the Real World: Preparing for AI Agents." *Center for Security and Emerging Technology*. https://doi.org/doi.org/10.51593/20240034.

Turner, Alex, Logan Smith, Rohin Shah, Andrew Critch, and Prasad Tadepalli. 2021. "Optimal Policies Tend To Seek Power." *NeurIPS 2021*. https://proceedings.neurips.cc/paper/2021/hash/c26820b8a4c1b3c2aa868d6d57e14a79-Abstract.html.

Turner, Alexander Matt and Prasad Tadepalli. 2022. "Parametrically Retargetable Decision-Makers Tend To Seek Power." arXiv. https://arxiv.org/abs/2206.13477.

UNICRI. 2021. "Algorithms and Terrorism: The Malicious Use of Artificial Intelligence for Terrorist Purposes." *United Nations International Crime and Justice Research Institute*. https://unicri.org/News/Algorithms-Terrorism-UNICRI-UNOCCT.

U.S. Department of Energy - Office of History and Heritage Resources. 2013. "Espionage And The Manhattan Project." *The Manhattan Project: An Interactive History*. https://www.osti.gov/opennet/manhattan-project-history/Events/1942-1945/espionage.htm.

U.S. Department of Energy, Office of Cybersecurity, Energy Security, and Emergency Response. 2021. "Cybersecurity Risk Information Sharing Program (CRISP)." https://www.energy.gov/sites/default/files/2021-12/CRISP%20Fact%20Sheet_508.pdf.

U.S. National Security Agency's Artificial Intelligence Security Center, Cybersecurity and Infrastructure Security Agency, Federal Bureau of Investigation, et al. 2024. "Deploying AI Systems Securely: Best Practices for Deploying Secure and Resilient AI Systems."

https://media.defense.gov/2024/Apr/15/2003439257/-1/-1/0/CSI-DEPLOYING-AI-SYSTEMS-SECURELY.PDF.

Vincent, Brandi. 2023. "Inside Task Force Lima's Exploration of 180-plus Generative AI Use Cases for DOD." *Defense Scoop*. https://defensescoop.com/2023/11/06/inside-task-force-limas-exploration-of-180-plus-generative-ai-use-cases-for-dod/.

Vincent, James. 2016. "Twitter Taught Microsoft's AI Chatbot to Be a Racist Asshole in Less than a Day." *The Verge*. https://www.theverge.com/2016/3/24/11297050/tay-microsoft-chatbot-racist.

Vincent, James. 2017. "Putin Says the Nation That Leads in AI 'Will Be the Ruler of the World.'" *The Verge*. https://www.theverge.com/2017/9/4/16251226/russia-ai-putin-rule-the-world.

Walsh, Dylan. 2024. "The Disinformation Machine: How Susceptible Are We to AI Propaganda?" *Stanford University Human-Centered Artificial Intelligence*. https://hai.stanford.edu/news/disinformation-machine-how-susceptible-are-we-ai-propaganda.

Wan, Alexander, Eric Wallace, Sheng Shen, and Dan Klein. 2023. "Poisoning Language Models During Instruction Tuning." arXiv. https://arxiv.org/pdf/2305.00944.

Wiblin, Robert and Buck Shlegeris. 2025. "#214 – Buck Shlegeris on Controlling AI That Wants to Take over – so We Can Use It Anyway." *80000 Hours*. https://80000hours.org/podcast/episodes/buck-shlegeris-ai-control-scheming/.

Wiggers, Kyle. 2025. "Safe Superintelligence, Ilya Sutskever's AI Startup, Is Reportedly Close to Raising Roughly $1B." *Tech Crunch*. https://techcrunch.com/2025/02/18/safe-superintelligence-ilya-sutskevers-ai-startup-is-reportedly-close-to-raising-roughly-1b/.

Wiggers, Kyle, Cody Corrall, Alyssa Stringer, and Kate Park. 2025. "ChatGPT: Everything You Need to Know about the AI-Powered Chatbot." *Tech Crunch*. https://techcrunch.com/2025/06/30/chatgpt-everything-to-know-about-the-ai-chatbot/

Wivel, Nelson A. 2024. "Historical Perspectives Pertaining to the NIH Recombinant DNA Advisory Committee." *Human Gene Therapy*. https://pmc.ncbi.nlm.nih.gov/articles/PMC3900000/.

Woo, Erin, Stephanie Palazzolo, and Amir Efrati. 2024. "OpenAI, in Duel With Anthropic, Doubles Down on AI That Writes Software." *The Information*. https://www.theinformation.com/articles/openai-in-duel-with-anthropic-doubles-down-on-ai-that-writes-software.

Yan, Shenao, Shen Wang, Yue Duan, et al. 2024. "An LLM-Assisted Easy-to-Trigger Backdoor Attack on Code Completion Models: Injecting Disguised Vulnerabilities against Strong Detection." arXiv. https://arxiv.org/pdf/2406.06822.

Zeff, Maxwell and Kyle Wiggers. 2024. “OpenAI Announces New O3 Models.” *Tech Crunch*. https://techcrunch.com/2024/12/20/openai-announces-new-o3-model/.

Zelikow, Philip, Mariano-Florentino Cuéllar, Eric Schmidt, and Jason Matheny. 2024. “Defense Against the AI Dark Arts.” *The Hoover Institution*. https://www.hoover.org/sites/default/files/research/docs/Zelikow_DefenseAgainst_web-241126.pdf#page=10.68.

Zhao, Xuandong, Xianjun Yang, Tianyu Pang, et al. 2024. “Weak-to-Strong Jailbreaking on Large Language Models.” arXiv. https://arxiv.org/abs/2401.17256.

Zhou, Kaiwen, Chengzhi Liu, Xuandong Zhao, et al. 2025. “The Hidden Risks of Large Reasoning Models: A Safety Assessment of R1.” arXiv. https://arxiv.org/abs/2502.12659.

Zhu, Yuxuan, Antony Kellermann, Akul Gupta, et al. 2025. “Teams of LLM Agents Can Exploit Zero-Day Vulnerabilities.” arXiv. https://arxiv.org/abs/2406.01637.