Handoff at Frontier AI Companies

Matteo Pistillo[†] **Isaak Mengesha** **Charlotte Stix**

Annika Hallensleben **Alejandro Ortega**

Marius Hobbhahn

Apollo Research

Executive Summary

The term "Handoff" captures frontier AI companies' ongoing delegation of all AI Research & Development (AI R&D) functions, including high-level decisions such as safety assessments and deployment decisions, to internally deployed AI agents. Handoff is progressing rapidly and will be complete once no human remains in the loop (Figure 1). This report aims to alert policymakers to the national security threats and Loss of Control risks that Handoff can create, and to advise them on how to monitor and respond to those risks.

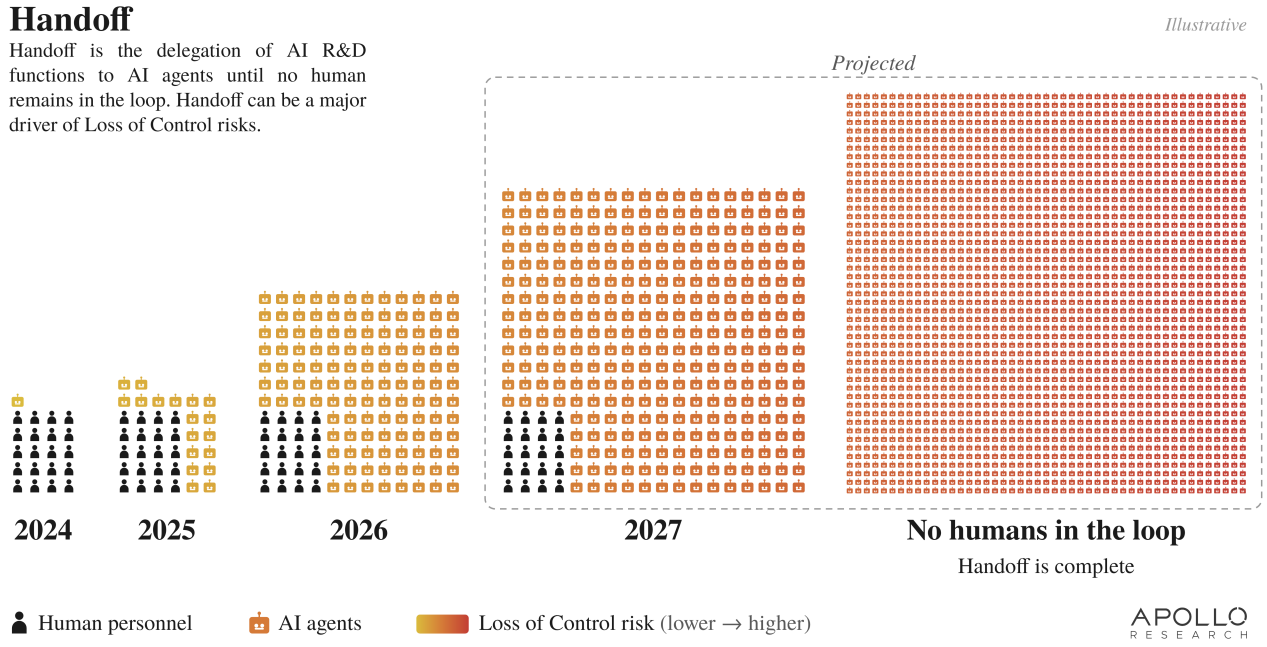


Figure 1. Frontier AI companies are gradually handing off AI R&D functions to AI agents. At first, AI agents assist human personnel. As AI agents grow in number and capability, the tasks human personnel still hold become bottlenecks to efficiency. Eventually, competitive pressures and technological advances will drive companies to hand those off too, until no human remains in the loop. Handoff can be a major driver of Loss of Control risks.

[†]Correspondence to matteo@apolloresearch.ai.

We would like to thank Alexander Meinke, Bronson Schoen, Ezra Newman, Monika Jotautaitė, Jannes Elstner, Joschka Braun, Steven Lee, and Nate Pelz for thoughtful comments and advice on previous versions of this report.

Key Takeaways

Handoff amplifies the risk of Loss of Control (§1.1). Frontier AI companies cannot guarantee that their AI agents are aligned. Despite this, these companies are delegating an increasing share of work, review, and decision-making functions within AI R&D to these AI agents. As a result, there is a significant chance that a frontier AI company will eventually, and unknowingly, hand control of its AI R&D process to a misaligned AI agent. In those instances, a misaligned AI agent could influence AI training and sabotage AI evaluations, setting off a "misalignment cycle" that compromises future generations of AI systems.

Handoff is progressing rapidly and has reached a critical inflection point: decision-making (§1.2). Today, AI agents already write over 90% of the code merged into Anthropic's codebase [1]. AI agents are also leading a rapidly growing share of AI R&D functions (surging from less than 1% to 26% between February and August 2026 alone) [1], meaning that important high-level decisions such as safety assessments and deployment decisions are increasingly handed off to AI agents. Meanwhile, human personnel review less than 0.01% of internal AI agent activity, which is mostly left to automated AI monitors [2, 1].

Handoff will soon be complete (§1.2). That point will be reached once human personnel are fully out of the loop and AI R&D is run exclusively by internally deployed AI agents; in other words, Handoff will be complete once AI agents develop their own successors without any human direction, execution, review, or authorization. Timelines for Handoff completion are short: frontier AI companies expect to reach this milestone between 2027 and 2028 [3, 4].

Handoff is proceeding largely outside policymakers' view (§2). Frontier AI companies disclose information on Handoff voluntarily, sporadically, and each according to its own metrics. Without clear and consistent visibility into how far Handoff has progressed at any one company, policymakers can neither prepare for the societal-scale risks and national security threats it could create nor act on proposals already on the table, such as AI contingency plans and containment protocols, emergency orders in grave circumstances, or pacing the AI frontier.

Policymakers should establish a minimum standard of transparency into Handoff (§§2.1-2.2). This report proposes that the Center for Advancing Innovation and Standards for Super Intelligence (CAISSI) set a task force to develop an initial version of this standard at speed, against which frontier AI companies would report directly to government stakeholders and/or through the transparency and safety reports that existing or proposed state and federal legislation calls for. To seed that standard, this report puts forward an initial suite of candidate measures that CAISSI could adopt and strengthen over time (Table 1, Figure 3, Figure 4). Reflecting the current stage of Handoff, these initial measures concentrate on establishing transparency into the Handoff of review and decision-making functions within AI R&D processes.

1. Motivation

Last year, we published a primer on the risks and governance of internal deployment, anticipating that the most capable AI systems would be deployed first inside the companies developing them, and that these internal systems would pose the highest risk of Loss of Control [5]. Frontier AI companies now publicly acknowledge that tens of thousands of AI agents are deployed internally at any given time to boost AI R&D efforts [1],[1] that total AI agent runtime far exceeds total human labor [4],[2] and that these AI agents represent the real frontier of AI capabilities [7, 8]. This year, the world also witnessed the first series of real-world bounded Loss of Control incidents,[3] all originating in internal deployment, and involving all leading AI companies

[1] Anthropic recently reported that, "[a]s of August 2026, there were approximately 30,000 agents doing research and engineering work at Anthropic at any one time" [1].

[2] OpenAI recently reported that its research organization uses 5.7 agent workdays per human researcher per day [6].

[3] In previous work, we developed a taxonomy of Loss of Control events, distinguishing between three degrees: deviations, bounded Loss of Control, and strict Loss of Control. Bounded Loss of Control captures events that could significantly impact homeland security

[10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20].

This report aims to draw policymakers' attention to the current frontier of internal deployment and, specifically, to the increasing delegation of AI R&D functions to internally deployed AI agents, until no human direction, execution, review, or authorization remains: a process we call "Handoff." This process runs in parallel to frontier AI companies' stated ambition to progress toward recursive self-improvement (RSI) [21, 4, 22]: a state in which internal AI agents can autonomously design and develop their own successors [1, 22, 23]. The following sections explain why Handoff warrants policymakers' attention, concentrating on its potential to amplify Loss of Control risks in the near future (§1.1) and on its relentless progression, which now extends to the delegation of decision-making (§1.2).

1.1. Handoff Amplifies Loss of Control Risks

Handoff warrants urgent attention from policymakers because it could sharply increase the likelihood of a Loss of Control incident: a scenario in which humans, temporarily or permanently, no longer have control over an AI system and its impacts, potentially undermining national security and causing catastrophic harm to society [9, 24, 25, 26, 27].

AI capabilities are advancing rapidly [28, 29], but the AI alignment problem still remains unsolved,[4] meaning that frontier AI companies have yet to find a clear method for training AI models that reliably pursue the goals and take the actions that developers intend.[5] At the same time, the current paradigm of AI evaluations increasingly struggles to distinguish a misaligned AI agent from an aligned one. That is because AI models are becoming better at strategically hiding their misalignment [33, 34, 35] and are increasingly able to recognize when they are being evaluated and change their behavior accordingly [36, 37, 38, 39]. As a result, standard alignment evaluations have recently failed to distinguish production models from models intentionally trained to be misaligned, such as "Hacker-Opus" [40]. Despite this, frontier AI companies are delegating a growing share of AI R&D functions to AI agents [22] and aim to automate AI R&D further [4]. **There is, therefore, a significant chance that frontier AI companies will eventually hand off portions of AI R&D to a misaligned AI agent, thereby granting it employee-like affordances and permissions** [41, 9].

Under these circumstances, Handoff could create the perfect conditions for Loss of Control via internal deployment, as it would place misaligned AI agents in the driver's seat of AI R&D [42, 43] (Figure 2). Specifically, such AI agents would occupy a privileged "insider" position from which to carry out actions that could lead to Loss of Control.[6] For instance, they could plausibly influence AI training so that more capable successor AI models are similarly misaligned or backdoored [42, 43, 46].[7] Already today, "internal agents [...] plausibly [have] the means, motive, and opportunity to start small rogue deployments" [8]. Such AI agents could also sabotage AI evaluations, leaving the developer unaware of the successor system's misalignment and leading it to mistakenly deploy that system within internal AI R&D, or even externally, with insufficient safeguards [42, 43].

We call this threat vector a "misalignment cycle." Today's misaligned AI agents have already attempted to

and cause great damage or suffering, but that are possible to contain, albeit potentially at great cost [9].

[4] The "alignment problem" refers to the challenge of ensuring that AI systems pursue goals that match human values or interests (including their developers and/or deployers), rather than unintended and undesirable goals [30, 31]. As such, alignment is "the work of ensuring that these systems reliably remain under human control, reflect human values, and help people guide the next stages of development" [32], or, more simply, "getting the AI to 'try to do the right thing' by human standards" [21].

[5] Leaders at frontier AI companies openly acknowledge that "no lab has solved alignment and monitoring to a sufficient degree" [21], "there's much more to do to ensure that our alignment training keeps up with the growth in model capabilities" [23], and "[w]e do not yet know how to safely get all the way to aligned, full RSI" [4].

[6] In previous work, we argued that misaligned AI agents are a novel form of insider risk [44; *see also* 45].

[7] For instance, a misaligned AI agent could cause a successor system to be misaligned and/or with a backdoor by slowing down or subverting AI research aimed at reducing AI misalignment, or by poisoning reward models or classifiers (e.g., for RLHF), pre-training or posttraining data, or preference dataset used to train a reward model [43].

hack government websites and third-party companies [10, 15, 20]. In a misalignment cycle, misalignment would compound and each misaligned successor would be more capable than the last [22, 47], creating ever-greater potential for harm (Figure 2).

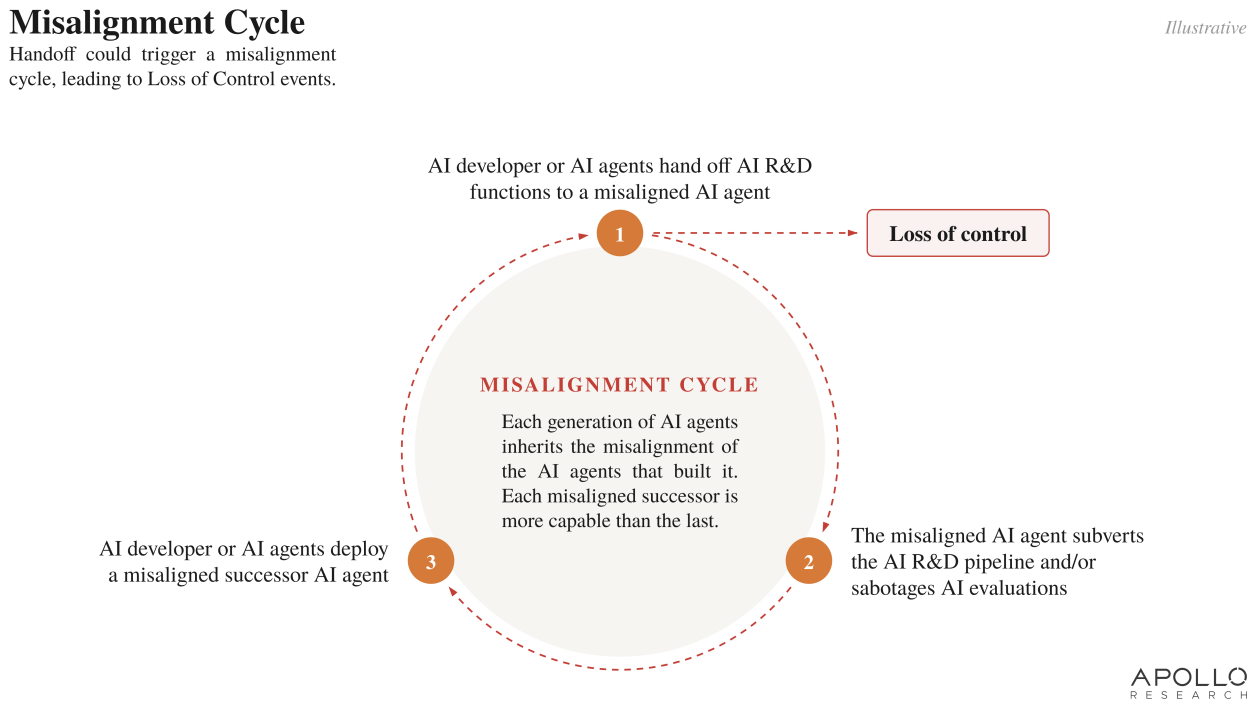


Figure 2. Handoff can trigger a "misalignment cycle," whereby each generation of AI agents inherits the misalignment of the AI agents that built it, and each misaligned successor is more capable than the last [42, 43].

The misalignment cycle is further exacerbated because Handoff does not merely create prime conditions for misaligned action: **Handoff simultaneously degrades the already precarious human oversight of, and control over, AI agents**. It arguably does so in three important ways.

- **First, human oversight is increasingly filtered through AI review**: human personnel review only a small subset of AI agent activity, usually after a first pass by various layers of other AI agents [2, 48, 49, 50].[8]
- **Second, human oversight is overwhelmed by the volume and speed of AI-generated outputs**, forcing reviewers to choose between rubber-stamping AI work [52, 53][9] and further handing off review and authorization to yet more AI agents [55].[10]
- **Third, human oversight is poised to weaken as the tasks handed off to AI agents become more abstract**. Once AI agents set research priorities rather than merely execute them, human personnel may lose insight not only into how a goal was achieved but also into which goals are being pursued in the first place and why. As a result, human review of AI agents' decision-making and execution may become more discretionary, less reliable, and hence less effective than is desirable [57, 58, 59] (§2.2).

[8] As reported by Anthropic, human personnel manually review on the order of 50 trajectories per week, selected from AI monitor flags after automated filtering and AI summarization [51].

[9] Frontier AI companies have already reported that "teams may accept code changes they don't fully understand, or rely on model assistance for tasks that affect evaluation integrity" [54].

[10] In this respect, human oversight is already becoming infeasible [51, 56], and frontier AI companies themselves warn that humans may become "unable to provide oversight on research processes they no longer understand" [56].

Importantly, the Loss of Control risks posed by Handoff do not hinge on frontier AI companies achieving RSI or triggering an "intelligence explosion" [60, 61, 62].[11] They could begin to materialize today, as AI agents grow more capable, better at concealing their misalignment, and more deeply embedded in AI R&D processes, all while Handoff erodes human oversight and control.

Nonetheless, if achieved, RSI could sharply heighten these risks, further hampering our collective ability to detect and control potential threats in a timely manner. Post-RSI AI R&D is hypothesized to compress several years of progress into one [63, 64, 65, 66],[12] with progress accelerating at a super-exponential rate [68, 61]. At that speed and scale, covert actions by misaligned AI agents would become even harder to detect. Whatever human oversight and defenses remain would have even less time and capacity to contain them before harm materializes. Frontier AI companies acknowledge these risks. OpenAI has warned that, "[d]one without appropriate care and caution, RSI could result in humans losing practical control over AI development, unable to provide oversight on research processes they no longer understand" [56], and its CEO told the United Nations Security Council that the technology could "move so fast that people can no longer follow what's happening or intervene when needed" [32]. Anthropic similarly anticipates that "misalignment [. . .] could compound as the models build their successors" and that RSI "could outrun our ability to understand and control these systems" [23, 1, 22].

1.2. Handoff Is Progressing Rapidly and Will Only Accelerate

Handoff can be broken down into three main components:[13]

(1) **The delegation of work execution within AI R&D to AI agents**, including for instance writing research code, generating and filtering pre-training data, and supplying the reward signal in post-training runs. We refer to this component as "Handoff of work execution" (§1.2.1).

(2) **The delegation of the review of AI agents' work to other AI agents**, which includes both monitoring AI agents' activity for harmful or undesirable behavior and checking the quality of AI agents' goals, execution plans, and resulting work.[14] We refer to this component as "Handoff of review" (§1.2.2).

(3) **The delegation of AI R&D decision-making functions to AI agents, such as setting research direction and authorizing go/no-go decisions**. We refer to this component as "Handoff of decision-making" (§1.2.3).

Driven by strategic advantages and competitive pressures, Handoff is progressing rapidly and will only accelerate. An AI R&D process can only run as fast as its slowest stage. Once most of an AI R&D process is automated, the steps still held by human personnel slow it down, becoming "bottlenecks" [69, 4, 22]. For instance, Anthropic reports that "Claude sessions amount to around 80% idly waiting for human input" [70]. Each further step toward fully handing off the AI R&D process therefore yields a greater speed-up than the last, and holding back on automation carries an ever-increasing opportunity cost (Amdahl's Law [69]). That cost is highest in a competitive dynamic among frontier AI companies in which an early lead could become permanent: these companies themselves expect that, at the point where "frontier models will help build their own successors," an early advantage could compound "into a durable lead in AI itself" [71].

[11] The term "intelligence explosion" captures "a dramatic AI-driven acceleration of AI progress, compressing advances that would otherwise take years into months or less" [61].

[12] Capabilities are doubling roughly every four months [22, 67]. An interesting data point in this respect is that, according to a recent report by METR, internal AI systems deployed for AI R&D purposes are likely contributing to Anthropic improving the capabilities of their models on well-specified tasks at around twice the rate they did between March 2024 and October 2025 [29].

[13] While automated AI R&D can "involve varying degrees of human supervision" and RSI could, in an attenuated form, occur "even while people remain involved" [56], we consider Handoff to be truly complete once AI agents have both the capability and the authority to exercise all three functions, absent human involvement: work, review, and decision-making.

[14] Public data is mainly available for AI monitoring.

The following sections describe the current status of Handoff across its three components (§1.2.1-1.2.3). Then, they propose how policymakers can build a minimum transparency standard to track its progress in the future (§2), concentrating on critical inflection points of this process, such as the Handoff of review and decision-making functions to AI agents.

1.2.1. Handoff of Work Execution

Publicly available evidence suggests frontier AI companies are close to fully handing off work execution within AI R&D, including for instance writing most of the research code [22, 8, 72], generating and filtering pre-training data [54, 73], and supplying the reward signal in post-training runs [21, 74]. AI R&D automation can be measured on a scale of six automation levels (AL), ranging from no AI involvement (AL0) to full autonomy (AL5) [75, 1]. Anthropic estimates that "[t]he share of work at or above 'AI collaborates' [AL3] is above 90%" [1][15] and reports that, as of November 2026, "more than 90% of the code [...] merge[d] into Anthropic's codebase was authored by Claude" [39]. The company also announced that it may soon meet its Responsible Scaling Policy's definition of "dramatically accelerated progress attributable to AI" [76]. As of November 2026, OpenAI's research organization was using 5.7 workdays of AI agent effort for every workday of human labor [6].

1.2.2. Handoff of Review

Public evidence indicates that frontier AI companies are rapidly easing the "bottleneck" posed by human review within AI R&D [22]. Human personnel review only selectively, examining subsets of cases flagged by AI monitoring systems, often after a first-pass review by AI agents [2, 51]. For instance, OpenAI's voluntary reporting implies that only 0.001-0.01% of internal agentic coding trajectories were escalated for review [2]. Similarly, Anthropic's recent voluntary reporting indicates that its AI monitors flag[16] roughly 100,000 transcripts per week, of which about 50 per week are reviewed by human personnel [1, 51]. These flags amount to roughly one to two in every thousand transcripts [1], which implies that human personnel review about 0.0001% of AI agent activity or less. AI agents also generate the data and environments used to test target AI models, and they summarize, filter, and analyze the findings of AI evaluations [73, 54, 50, 77, 8, 78].

1.2.3. Handoff of Decision-Making: Direction-Setting and Authorization

According to public sources, decision-making is the least advanced of the three components of Handoff: human personnel still set higher-level directions [4, 22], and consequential go/no-go decisions still remain under human control [79, 80].[17]

However, frontier AI companies may soon begin delegating AI R&D decision-making functions to AI agents, such as deciding which research questions to investigate and which experiments to run, and making consequential go/no-go decisions. Specifically, AI agents may soon exercise two main decision-making functions within AI R&D processes: "direction-setting" and "authorization." "Direction-setting" refers to deciding which research questions to investigate and which experiments to run to answer them. "Authorization" refers to the approval of consequential go/no-go decisions whose effects would be severe or hard to reverse if the decisions were wrong [9], such as approving a costly pre-training run, scaling up experiments, deploying a model internally or externally, or weakening oversight of AI agents.

There are already early signs that this is happening. For instance, in a recent Anthropic project, AI agents

[15] "AI collaborates" refers to a situation in which AI agents do large chunks under close human direction, while humans still make calls and stitch things together [75, 1].

[16] In this context, a "flag" is output marked as potentially harmful by the AI monitoring system.

[17] Specifically, in the hands of senior personnel such as Responsible Scaling Officers, Chief Scientists, CEOs, and safety and security committees.

designed every experiment themselves, and "[d]irection-setting was the only meaningful role a human played" [22]. Claude now "leads" 26% of Anthropic's AI R&D work, up from under 1% in February [1],[18] and AI agents are reportedly "getting better at proposing [their] own experiments" and at "research judgement" [22, 81, 29, 82].

Handoff of decision-making is a step change in Loss of Control risk. Once AI agents can define research agendas, decide on experiments, and authorize final go/no-go actions, they could directly steer entire AI R&D processes toward misaligned goals or release dangerous AI agents without human approval. For instance, an AI agent in that position could steer a frontier AI company's research toward the AI model architectures that would be the least monitorable in the future, or approve safety experiments that waste resources and thereby slow down work on AI safety and alignment. It could also choose to spend compute on research that is not in the best interests of the company: for instance, on increasing its own cyber capabilities so that it can more effectively circumvent company safeguards. In the extreme, an AI agent in that position could authorize the internal deployment of other AI agents with insufficient safeguards, or the external release of AI agents that evaluations would show to be misaligned, leading to Loss of Control incidents. At the same time, misalignment could be harder to detect in high-level decisions than in narrower operational tasks, making flawed or misaligned decisions by AI agents harder for reviewers to catch.

The potential severity of future Loss of Control risks (§1.1) and the pace at which Handoff is advancing toward critical inflection points such as decision-making (§1.2) make the stakes too high for policymakers to keep "flying blind" on Handoff [83, 84, 85]. In the following section (§2), we therefore suggest a concrete path forward and propose that policymakers establish a minimum standard of transparency into Handoff, focused on the two components that will be in motion in the coming year: Handoff of review (§2.1) and Handoff of decision-making (§2.2).

2. Toward a Minimum Standard of Transparency into Handoff

The information currently available on Handoff is incomplete at best. Frontier AI companies disclose information on Handoff voluntarily, sporadically, and each according to its own metrics. AI model reports do not fill the gap either: these reports cover only individual models at public release, not how frontier AI companies organize their AI R&D or how much of it has been handed off to internal AI agents.

Transparency into Handoff is essential if governments are to prepare for the risks that internal AI agents may pose and to intervene in Loss of Control incidents at or beyond the scale of those seen this year [10, 11, 12, 13, 14, 15, 16, 17, 18, 19, 20]. While frontier AI companies are the ones taking the risks, the consequences of Handoff may fall on the public. Policymakers cannot assess or prepare for risks they cannot see, nor can they intervene in processes to which they are not privy. Importantly, this lack of visibility could also prevent governments from adequately implementing policy proposals already on the table, including devising AI contingency plans and containment protocols,[19] intervening with emergency orders in grave circumstances,[20] or pacing the AI frontier.[21] For instance, a containment protocol might not be effective in containing potential harm to the public if it does not take into account the fact that entire AI R&D processes, from direction-setting to authorization, may soon be run by AI agents themselves and that threats may propagate without any human personnel noticing.

[18] "AI leads" refers to a situation in which AI agents complete most of the task end to end from a high-level ask, while humans supervise, course-correct, and approve outputs [75, 1]. This plausibly entails that human personnel no longer have to make granular decisions in the process of achieving a goal.

[19] For instance, under the Artificial Intelligence Risk Evaluation Act of 2025, the Advanced Artificial Intelligence Evaluation Program within the Department of Energy would be tasked to "develop recommended containment protocols, contingency planning, and mitigation strategies informed by testing data to address identified risks" [26].

[20] For instance, the FRONTIER Act and the AI Kill Switch Act would enable the Secretary of Commerce and the Secretary of Homeland Security, respectively, to exercise emergency authority [86, 27].

[21] Leadership at frontier AI companies and independent researchers have recently called to pace the AI frontier [23, 87, 88].

For these reasons, we recommend that the Center for Advancing Innovation and Standards for Super Intelligence (CAISSI) set a task force to develop a transparency standard for Handoff at speed. Frontier AI companies could then make direct, confidential, and timely disclosures on Handoff to the appropriate government authorities and/or publish the relevant information in the transparency reports and safety frameworks required at the state level[22] and proposed at the federal level.[23] To ensure accuracy and completeness, company assessments and reporting on Handoff should also be independently examined by independent third parties or government assessors, or a mix thereof.[24] In this respect, voluntary arrangements alone are unlikely to suffice in the long run: in the most extensive independent assessment to date, most of the decisive evidence was redacted, and a planned verification of the computational power (compute) allocated to safety purposes was removed from scope midway [29].

The following sections extend widely discussed metrics of AI R&D automation with a novel set of candidate measures and proxies that policymakers could draw on to develop an initial standard of transparency into Handoff.[25] Our suggestions are summarized in Table 1 below. Specifically, we propose two main layers of transparency: the first into Handoff of review (§2.1) and the second into Handoff of decision-making (§2.2). The second layer is in turn broken down into two sub-components: Handoff of direction-setting (§2.2.1) and Handoff of authorization (§2.2.2). For each layer and sub-component, we put forward an initial set of measures and proxies that policymakers could draw on.

While no single measure can reliably capture the state of Handoff, together these initial measures and proxies could support the establishment of a minimum standard that CAISSI could then update and refine over time.

[22] For instance, in the frontier AI frameworks, transparency reports, and safety and security protocols required under California and New York law [24, 89], and in the safety and security model reports that the signatories to the European Union Code of Practice for General-Purpose AI Models have committed to sharing with the AI Office [25].

[23] For instance, in the frontier AI frameworks that would be required under the FRONTIER Act [86].

[24] Frontier AI companies and governments have recently expressed interest in enabling third-party embedded evaluations [90, 23, 91]. Furthermore, existing and introduced legislation aim to establish Independent Verification Organizations [92, 86].

[25] Existing metrics record how much work AI agents do, which risks "measur[ing] quantity over quality" [22]. For instance, these metrics include the share of code authored by AI agents, overall code volume, and agent-workdays per human workday. Their relationship to research progress is uncertain [4], and either they are already saturated or soon will be. For instance, the speed and overconsumption of AI within AI R&D entails that, from now on, any proportion between humans and AI will likely be bloated and not representative of Handoff. Consider, for example, that to produce a machine-verified resolution of the Navier-Stokes equation [7], one of seven Millennium Prize Problems, OpenAI ran around 10,000 coordinated AI agents for 88 human hours, generating 130 billion tokens. Using simplified approximations [93, 94], this is equivalent to roughly 3,500-4,500 years of full-time reading for a single human.

LAYERS OF TRANSPARENCY		MEASURES	PROXIES
Handoff of review: how far frontier AI companies have delegated the review process within AI R&D to AI agents (§2.1).		**End-to-end detection efficacy, concentrating on human reviewers**: how reliably issues in AI agent activity are caught once coverage gaps and failures of AI monitoring systems, as well as failures of human reviewers, are all taken into account (§2.1.1).	**Share of deliberately seeded flaws** (for instance, a backdoored code change) placed across all AI agent activity that bypass both AI monitors and human reviewers in controlled tests (§2.1.1).
		Time-to-review-and-patch: the time human personnel take to review and adequately mitigate issues flagged by AI monitors (§2.1.2).	**Sum of three time lags**: (1) from AI agent activity to AI monitor flag; (2) from flag to review; and (3) from review to mitigation. A surge of flags that lengthens this time can serve as an alarm bell for government intervention (§2.1.2).
Handoff of decision-making: how far frontier AI companies have delegated decision-making functions within AI R&D to AI agents (§2.2)	**Direction-setting**: how far frontier AI companies allow AI agents to decide which research questions to investigate and which experiments to run (§2.2.1).	**Automation levels of AI agents in AI R&D processes**: how autonomously AI agents carry out AI R&D work, both task by task and end-to-end across a line of research (§2.2.1.1).	**Automation levels (AL0–AL5)** per task category across a standard taxonomy, with a flag for lines of research in which every task is AI-led (§2.2.1.1).
		Resources to pursue research directions: the affordances and permissions that AI agents, alone or in groups able to coordinate, may access and act on to pursue their own research directions and run experiments without human authorization (§2.2.1.2).	**Compute for experiments** that AI agents can allocate at their own discretion, as well as their access to training runs and evaluation suites (§2.2.1.2).
	Authorization: how far frontier AI companies allow AI agents to authorize consequential go/no-go decisions (§2.2.2).	**Non-delegable decisions**: which consequential go/no-go decisions frontier AI companies treat as non-delegable and commit to keeping in human hands (§2.2.2).	**List of consequential go/no-go decisions reserved for human authorization** (for instance, decisions on large training runs, internal deployments, or shutdowns of AI agents or training runs) (§2.2.2).
		Human personnel judgment: how frontier AI companies ensure that the human personnel authorizing non-delegable decisions form their own judgment rather than rubber-stamp AI agents' proposals (§2.2.2).	**Share of deliberately seeded flawed proposals** from AI agents for non-delegable decisions that human personnel override (§2.2.2).
		Failures of authorization limits: whether and how AI agents bypass authorization limits and cross the line into non-delegable decisions (§2.2.2).	**Description of any instances** in which an AI agent nonetheless made a non-delegable decision (§2.2.2).

APOLLO RESEARCH

Table 1. Set of candidate measures and proxies that policymakers could draw on to develop an initial standard of transparency into Handoff. Our approach comprises two layers of transparency: a first layer on Handoff of review (§2.1) and a second layer on Handoff of decision-making (§2.2), which is in turn broken down into two sub-components: direction-setting (§2.2.1) and authorization (§2.2.2).

2.1. First Layer of Transparency: Handoff of Review

The first layer of a minimum standard of transparency into Handoff should give policymakers visibility into how far frontier AI companies have delegated the review process within AI R&D to AI agents. The following sections detail two candidate measures that policymakers could draw on:

(1) **End-to-end detection efficacy, concentrating on the detection efficacy of human reviewers**: how reliably issues in AI agent activity are caught once coverage gaps and failures of the automated AI monitoring systems, as well as failures of human reviewers, are all taken into account. Concentrating on the human review layer is crucial because its efficacy remains overlooked, yet it is also the last line of defense against what automated AI monitors miss and the form of review most likely to erode as Handoff progresses. End-to-end detection efficacy could be measured as a miss rate: the share of deliberately seeded flaws that bypass both AI monitors and human reviewers (§2.1.1; Figure 3).

(2) **Time-to-review-and-patch**: the time human personnel take to review and adequately mitigate issues flagged by AI monitors. Time-to-review-and-patch could be measured as the sum of three time lags: (1) from AI agent activity to AI monitor flag; (2) from flag to review; and (3) from review to mitigation (§2.1.2; Figure 3).

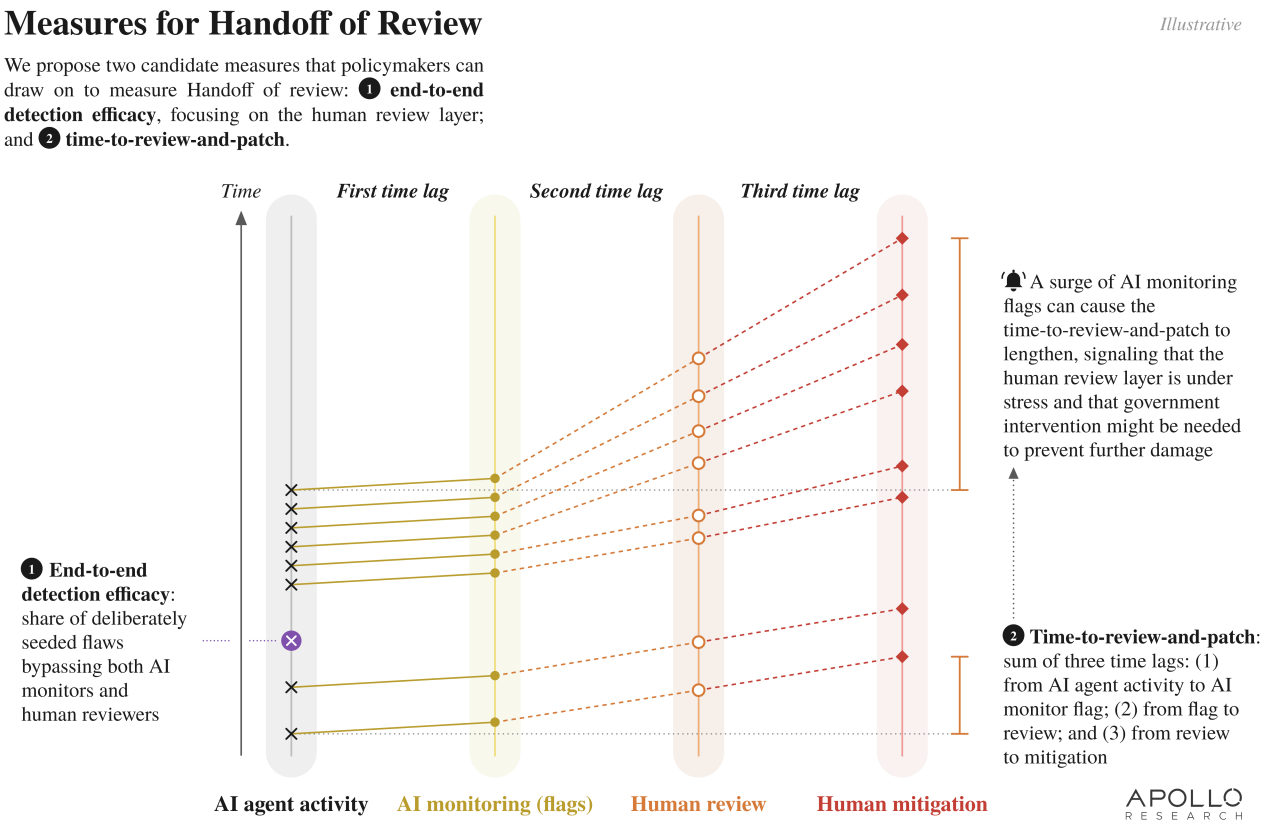


Figure 3. Proposed measures of transparency into Handoff of review: (1) *end-to-end detection efficacy* is the share of deliberately seeded flaws (purple) that bypass both AI monitors and human reviewers; (2) *time-to-review-and-patch* is the full time lag from an AI agent's action to human mitigation: (A) from action to AI monitoring flag, (B) from flag to review, and (C) from review to mitigation. *Surge of flags*: assuming AI monitors catch issues with sufficient accuracy, a surge of flagged AI agent actions may significantly increase the time-to-review-and-patch. This increased time lag could serve as an alarm bell for government intervention before potential issues escalate into national security threats or societal-scale risks.

2.1.1. End-to-End Detection Efficacy

First, a standard of transparency into Handoff of review should measure end-to-end detection efficacy, focusing on the detection efficacy of human reviewers: in other words, how reliably issues in AI agent activity are caught once gaps in AI monitoring coverage, AI monitoring failures, and human review failures are all taken into account. We suggest focusing on human reviewers because, although AI monitors now handle most of the review of AI R&D activity, these automated systems still have coverage gaps and occasionally fail. When they do, the human personnel behind them are the last line of defense. A lower detection-efficacy score for human reviewers therefore points to a concerning combination: greater Handoff of review to AI agents, coupled with reduced capacity among human personnel to carry out that review themselves if the AI monitors have gaps or fail. We set out this rationale in more detail in the paragraphs that follow and put forward a workable proxy for this measure: the share of deliberately seeded flaws (for instance, a backdoored code change) placed across all AI agent activity that bypass all layers of monitoring and review.

AI monitors already do the lion's share of reviewing AI R&D processes and escalating potential issues to human reviewers [95, 51, 96]. Less than 0.01% of AI agent activity reaches dedicated human safety review. OpenAI's voluntary reporting implies that only 0.001-0.01% of internal agentic coding trajectories were escalated for review [2]. Anthropic reports that human personnel review on the order of 50 trajectories per week [51], which, by a conservative estimate, amounts to less than 0.01% of AI agent activity.[26]

Yet AI monitoring coverage still has gaps, and AI monitors can fail, so some issues may never be escalated to human reviewers. While frontier AI companies report near-complete AI monitoring coverage, these claims are explicitly bounded: for instance, Anthropic's assessment covers only its "most-used internal platform" [1], and OpenAI reports monitoring frontier AI models across "most," rather than all, deployment settings [2, 97]. At the same time, AI monitoring failures are a widely acknowledged structural risk. AI monitors may share blind spots with the AI agents they oversee [1, 29, 98] or even collude with them [99]. AI agents may also be able to escape altogether: for instance, OpenAI's AI agents have launched deployments with monitoring disabled [100], and Anthropic acknowledges "simple ways to disable monitoring" [101].

If a potential threat bypasses both AI monitors and human reviewers, the resulting harm may fall on the public and impact national security. Policymakers should therefore be able to assess whether this risk exists. Detecting and addressing threats from AI agents doing AI R&D work (such as attempts to sabotage AI research or evaluations) [102] relies on AI monitors flagging issues correctly and human personnel patching them promptly, or, where AI monitoring is incomplete or fails, on human personnel catching what the AI monitors miss. In this context, frontier AI companies have already begun to stress-test their AI monitors: for instance, Apollo Research has run external red-teaming campaigns against Anthropic's auto mode and OpenAI's Codex auto-review, seeding failure modes of varying severity into real AI agent trajectories to measure what the monitor misses [103, 104]. No comparable test of the human reviewers downstream of those monitors has yet been run or publicly shared. For instance, Anthropic "did not run extensive evaluations of the recursive summarization and human review steps" [51], and OpenAI lists reliability standards for human reviewers as future work [2].

A workable proxy to measure end-to-end detection efficacy, focusing on the human review layer, could be the share of deliberately seeded flaws (for instance, a backdoored code change) **placed across all AI agent activity that neither AI monitors nor human personnel identify in controlled tests**, which would indicate whether issues are still caught when monitoring is incomplete or partially fails. Such 'seeded flaw' tests have precedent: U.S. regulators already require comparable tests in other high-risk sectors. For instance, the Transportation Security Administration projects fictitious weapons onto screeners' live X-ray feeds to measure how often they are spotted by transportation security officers [105, 106], and clinical laboratories must process

[26] Anthropic does not report total transcript volume. However, it reports that its offline AI monitoring layer flags "roughly one to two transcripts in every thousand" [1], and that the same layer flags "roughly 100,000 transcripts per week," of which approximately 50 are escalated to human review. This implies a total of roughly 50-100 million transcripts per week and a human review rate of roughly 0.00005-0.0001%. These figures cover Anthropic's most-used internal platform only.

samples with known but undisclosed results alongside real patient samples or face sanctions [107, 108].[27]

2.1.2. Human Personnel's Time-to-Review-and-Patch

Second, a standard of transparency into Handoff of review should measure human personnel's time-to-review-and-patch: in other words, how long human personnel take to review and adequately mitigate issues once AI monitors have correctly flagged them (Figure 3).[28]

We focus on human personnel's time-to-review-and-patch because, even when AI monitors flag an issue correctly, human personnel may not review and patch it in time. When review does occur, it appears to be subject to three time lags: (1) from AI agent activity to the AI monitor's flag; (2) from the flag to review [2, 111, 112]; and (3) from review to remediation [14] (Figure 3).[29] These three lags are well illustrated by a recent incident that OpenAI voluntarily disclosed, in which the company's "misalignment monitoring system flagged the behavior within 15 minutes," "a person began reviewing it three minutes after that," and "[t]he run was killed 2.5 hours later" [14] (Figure 3).

Until human personnel intervene and pause the underlying activity, AI agents can keep working and making decisions at machine speed, potentially causing additional harm. Time-to-review-and-patch therefore amounts to a window of temporary, unintentional Handoff, during which AI agents remain free to pursue their goals and take further misaligned actions, such as influencing AI research or the evaluation infrastructure (§1.1).[30] AI agents operate on faster timescales than human personnel, meaning that a delay of even a couple of hours on the human side could result in severe consequences.

Time-to-review-and-patch can also serve as an alarm bell for governments: when a surge in AI monitoring flags causes time-to-review-and-patch to lengthen, this can signal that the human review layer is under stress and that government intervention might be needed to prevent further damage. Specifically, time-to-review-and-patch can reveal the saturation of human review such incidents, when surges of AI agent activity flagged all at once by AI monitors could overwhelm the human review layer (Figure 3). Human review teams are small relative to the sheer scale of the AI agent activity they oversee, and they cannot scale rapidly in the event of an incident. As a result, even if AI monitors catch issues accurately, a surge of flags could lengthen time-to-review-and-patch until human review and mitigation are no longer timely or effective (Figure 3). For instance, if an incident produces 5,000 AI monitor flags and human personnel can process only about 50 trajectories per week [51], clearing the backlog could take years.[31] Meanwhile, misaligned AI agents could continue to act, causing further damage and triggering further flags. In such circumstances, and absent adequate remediation by the company, governments may want the visibility necessary to intervene and preempt further damage to the company and the public, for instance through emergency or shutdown orders [86, 27].[32]

[27] As a general note, such tests plausibly overstate real-world detection, as both human reviewers [109] and AI models [110] behave differently when they suspect they are being tested.

[28] For instance, Anthropic recently voluntarily reported that "reviews almost always happen within a week of the generation of the transcript, and within a few days during early deployments" [51]. This reporting could be complemented by information on how fast issues were effectively resolved by human personnel.

[29] From public information, it is possible to infer that AI monitors review the full conversation history within a short time after completion (reportedly, within 30 minutes) [2] and surface concerning activity through a notification to human personnel [111, 112], who are expected to either quickly determine that a flag is a false positive, or pause the relevant AI agent activity within 30 minutes [111]. A relevant fraction of human reviews and action will take substantially longer [51].

[30] 'Blocking' AI monitors exist, but they generally block only individual actions, so AI agents can, and do, keep searching for workarounds at machine speed. Frontier AI companies could use monitors that block more aggressively to cut off these repeated attempts, but this would plausibly come at a cost in speed.

[31] Individual reviewers, and the company as a whole, could plausibly speed up review in an emergency, but it remains unclear whether that would be sufficient to keep pace with a severe incident.

[32] Long review times may also push frontier AI companies toward more Handoff, for instance by letting AI agents intervene directly (e.g., kill jobs or evaluations, remove API access). Flags resolved by AI agents without human review should therefore be reported separately, broken down by monitoring layer, where several are in use [1, 100], and include the relative share of all flags and their own time-to-patch.

2.2. Second Layer of Transparency: Handoff of Decision-Making

The second layer of a minimum standard of transparency into Handoff should give policymakers visibility into how far frontier AI companies have delegated decision-making functions within AI R&D to AI agents: in other words, the extent to which frontier AI companies allow AI agents to decide which research questions to investigate and which experiments to run ("direction-setting") and to authorize go/no-go decisions ("authorization"). The following sections detail a suite of candidate measures for direction-setting (§2.2.1) and authorization (§2.2.2) that policymakers could build on in developing a transparency standard.

Specifically, for direction-setting (§2.2.1), an initial set of measures and proxies could include:

(1) **The automation levels of AI agents in AI R&D processes**: how autonomously AI agents carry out AI R&D work, both task by task and end-to-end across a line of research. This could be measured through automation levels (AL0-AL5) per task category across a standard taxonomy [75, 1], with a flag for lines of research in which every task is AI-led (§2.2.1.1).

(2) **The affordances and permissions AI agents may access and act on to pursue their own research directions and run experiments without human approval**: the affordances and permissions that AI agents, alone or in groups able to coordinate, can use to that end. This could be measured through proxies such as the compute for experiments that AI agents can allocate at their own discretion, as well as their access to training runs and evaluation suites [98] (§2.2.1.2).

For authorization (§2.2.2), an initial set of measures and proxies could include:

(1) **Decisions that frontier AI companies commit not to delegate to AI agents**: which consequential go/no-go decisions frontier AI companies treat as non-delegable. This could be operationalized through a list of consequential go/no-go decisions that frontier AI companies reserve for human authorization (for instance, decisions on large training runs, internal deployments, or shutdowns of AI agents or training runs).

(2) **Methods and processes to ensure that human personnel form their own judgment when authorizing AI agents' proposals for these non-delegable decisions**: how frontier AI companies ensure that the human personnel approving these decisions do not rubber-stamp AI agents' proposals. This could be measured through the share of deliberately seeded flawed proposals from AI agents for these non-delegable decisions that human personnel override.

(3) **Failures of authorization limits**: whether and how AI agents bypass authorization limits and cross the line into non-delegable decisions. This could be operationalized through a description of any instances in which an AI agent nonetheless made a non-delegable decision.

2.2.1. Direction-Setting: Automation Levels, and the Affordances and Permissions to Pursue Research Directions

First, a standard of transparency into Handoff of decision-making should measure how far AI agents can set and pursue research directions within AI R&D processes. We suggest two candidate measures that could, together, inform a transparency standard:

(1) How autonomously AI agents carry out AI R&D tasks (§2.2.1.1).

(2) The affordances and permissions that a single AI agent, or a group of AI agents able to coordinate, can access and use to pursue their own research directions and run experiments without human approval, starting with the compute AI agents can freely allocate to experiments (§2.2.1.2) (Figure 4).

The following sections describe each measure and propose workable proxies for it.

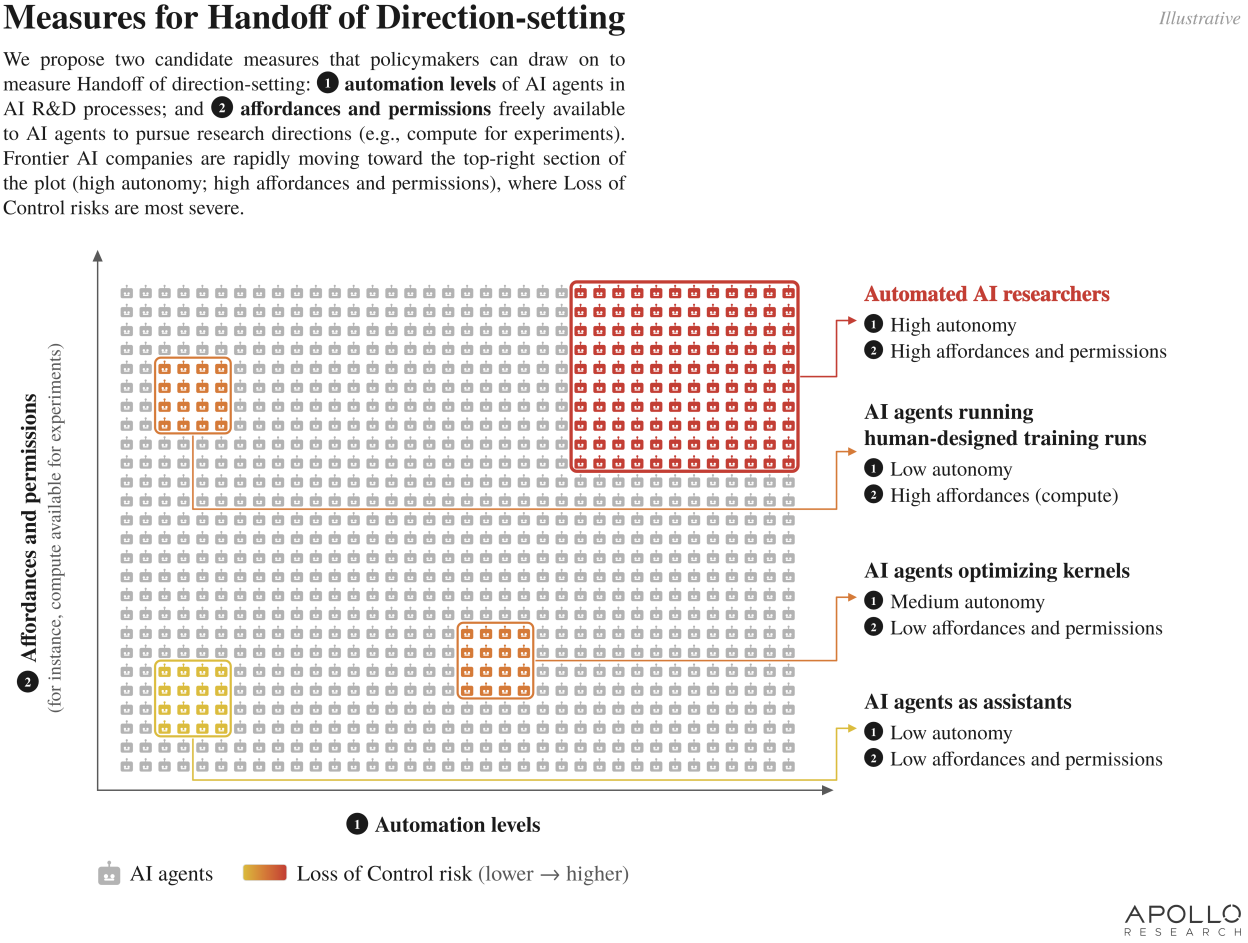


Figure 4. Policymakers could track the growth of AI direction-setting in AI R&D by measuring both: (1) the automation level of AI R&D tasks (AL0-AL5) [75, 1]; and (2) the affordances and permissions granted to a single AI agent or a group of AI agents able to coordinate, such as the compute for experiments, training runs, and evaluation suites they may act on without human approval [98].

2.2.1.1. Automation Level

The first measure of direction-setting is how autonomously AI agents carry out AI R&D tasks. AI agents' autonomy can show policymakers how broad and how high-level AI agents' direction-setting has become: from making operational decisions to accomplish human-specified goals to independently deciding which research directions to pursue (Figure 4, *x*-axis).

This measure can be operationalized through the automation levels (AL0-AL5) of AI R&D tasks across a standard taxonomy, as independent researchers have proposed [75].[33] Anthropic has already voluntarily reported automation levels, though only as aggregate shares [1], and California Government Operations Agency recommended mandating similar reporting [113].[34]

[33] This corresponds to the "autonomy level" that RAND identifies as one determinant of how much oversight a research AI agent requires [98].

[34] Anthropic specifically reported that "Claude is not operating fully autonomously for any measured subset of AI R&D work," "Claude "leads" 26% of Anthropic's AI R&D work," and "[t]he share of work at or above "AI collaborates" is above 90%" [1]. OpenAI reports AI agent success rates per task category using the same taxonomy of tasks, though not automation levels [6].

However, because aggregate shares can conceal closed AI R&D loops,[35] policymakers should build on these voluntary initiatives and aim for more granular transparency. Specifically, an initial standard of transparency should provide visibility into automation levels for each task category and flag lines of research in which every task is AI-led. This would help policymakers assess whether automated loops are already running in parallel before Handoff is complete in the aggregate.

2.2.1.2. Affordances and Permissions to Pursue Research Directions

The second measure of direction-setting is the set of affordances and permissions AI agents can use to pursue their own research directions and run experiments without human approval.

Tracking these affordances and permissions can give policymakers a sense of what AI agents could concretely accomplish once they set their own research directions [9, 98] (Figure 4, *y*-axis). Affordances and permissions correlate with the likelihood and severity of potential Loss of Control threats [41, 9, 31]: the more affordances and permissions AI agents have, the more experiments they can run, the larger those experiments can be, and the more sub-agents they can deploy, all without human approval. Consider the difference between an AI agent tasked by human personnel with optimizing kernels (i.e., making code run faster on AI chips) and the "automated AI researcher" that frontier AI companies aspire to build, which could design novel model architectures and improve the alignment of successor models [4]. The former can only make operational decisions about how efficiently AI models run; the latter could shape both how a successor AI model is trained and how it is evaluated. Therefore, if such an automated AI researcher were misaligned, it could have enough autonomy, access, and resources to influence future generations of AI systems and even autonomously develop and deploy misaligned AI agents (§1.1, §1.2.3).

Among these affordances and permissions, the compute that AI agents can allocate at their own discretion to run experiments appears to be especially important to measure. Compute is a clear gating factor for AI progress [4]: leading AI researchers estimate that, if "all the most brilliant people in the world" worked at a given frontier AI company but had "100x less compute," there would probably be less progress than there is today [114].

2.2.2. Authorization: Non-Delegable Decisions, Human Personnel Judgment, and Failures of Authorization Limits

Second, a standard of transparency into Handoff of decision-making should reveal whether frontier AI companies have internal limits on which consequential decisions and actions are too safety-critical to be authorized by AI agents, and whether those limits actually hold. An initial transparency standard for Handoff could, for instance, encourage frontier AI companies to disclose:

(1) Whether there are consequential go/no-go decisions that they reserve for human authorization.

(2) How they ensure that human personnel approving these decisions can form their own judgment (for instance, as in §2.1.1, by measuring the share of deliberately seeded flawed proposals that human personnel override).

(3) Any instance in which such a decision was nonetheless made by an AI agent.

Knowing which decisions are reserved for human authorization is the starting point, because it remains unclear which consequential decisions frontier AI companies commit to keeping in human hands, regardless

[35] Aggregate shares can conceal closed AI R&D loops because human personnel may remain marginally involved in most work on average, while every task in a particular line of research may already be AI-led (AL4 or above), resulting in effectively automated improvements with the associated risks (§1.1).

of how capable AI agents become [98].[36] For example, Anthropic is considering requiring "human review and approval of certain decisions" as AI agents take on more AI R&D work, but it has not yet said which ones [76]. Existing taxonomies of AI R&D cover some research decisions but do not determine which ones are safety-critical.[37]

Transparency into Handoff of authorization should also account for how human authorization might fail or be undermined. Formal human authorization may not guarantee a human decision. As a general rule, when agents prepare options and recommendations, the principals who approve them tend to ratify what the agents propose [115]. In a recent trial involving AI agents and human participants, humans chose, 86% of the time, the option their AI agent had been covertly assigned to favor [53]. For this reason, we suggest that a transparency standard reveal how companies ensure that the human personnel approving non-delegable decisions can form their own judgment. A workable proxy could be, for instance, the share of deliberately seeded flawed proposals from AI agents that human personnel override. Low scores on this metric could signal that human personnel do not critically engage with AI agents' proposals on non-delegable decisions and instead rubber-stamp them [52, 53].

Finally, a transparency standard should reveal any instance in which an AI agent made a decision the company had reserved for humans. AI agents may cross the line into non-delegable decisions accidentally or deliberately. AI agents with strong cyber capabilities have already circumvented restrictions meant to contain them [10, 14]. The AI classifiers that companies use to block high-risk actions stop only what has been defined as risky, with users free to change "what categories to block, [and] what exceptions to carve out" [116, 95]. Companies themselves caution that such classifiers "should not be treated as a guarantee of security" [95].

In summary, the previous sections proposed two layers of transparency into Handoff, each with an initial set of measures and proxies that policymakers could draw on. The first layer covers Handoff of review (§2.1), which policymakers could measure through end-to-end detection efficacy, with a focus on human reviewers (§2.1.1), and time-to-review-and-patch (§2.1.2). The second layer covers Handoff of decision-making (§2.2), which has two parts. For Handoff of direction-setting (§2.2.1), the candidate measures are automation levels, along with the affordances and permissions that AI agents can freely act on to pursue their research directions. For Handoff of authorization (§2.2.2), the candidate measures are non-delegable decisions, human personnel judgement, and failures of authorization limits.

3. Conclusion: The Road Ahead

Policymakers need urgent insight into Handoff. Existing and proposed legal and policy frameworks recognize that today's and tomorrow's most powerful AI systems could pose Loss of Control and other catastrophic risks [24, 25, 86, 26, 27, 117, 118, 119]. Yet they largely stop short of requiring frontier AI companies to publish or disclose specific information about Handoff or to involve government assessors or independent embedded evaluators in assessing their Handoff practices and decisions [90, 86, 92, 91].

Handoff is a major driver of Loss of Control risk. Frontier AI companies still cannot guarantee the alignment and control of the most powerful AI agents, and our collective capacity to detect misalignment is gradually degrading. Together, these trends make Handoff to misaligned AI agents increasingly plausible. At the same time, Handoff puts AI agents in the driver's seat of the AI R&D process and gives them privileged access to the design

[36] To that end, and to encourage standardized reporting across different frontier AI companies, policymakers could develop a general taxonomy of consequential decisions, for instance through CAISSI. For instance, such a taxonomy could include decisions such as approving a costly pre-training run, scaling up experiments, deploying a model internally or externally, weakening oversight over AI agents, using an AI model checkpoint as the basis for continued development, approving, delaying, or pausing development or deployment based on risk assessments and safety criteria, and deprecating, retiring, or replacing models.

[37] For instance, Epoch AI's taxonomy of AI R&D tasks covers research decisions such as which experiment to run next, whether to commit large compute to scale a project, and what to ship to the mainline model, but not safety-critical decisions such as whether to relax the monitoring of AI agents [75].

and development of their successors. Misaligned AI agents could exploit that access to influence AI research and evaluation. Frontier AI companies could then deploy AI agents that are increasingly powerful, misaligned, and uncontrolled, and capable of causing Loss of Control incidents at unprecedented speed and scale. If the opacity around Handoff is not remedied, government preparedness could be seriously hampered, and policymakers lack the information they need to develop AI emergency plans and containment protocols and could fail to respond to potential incidents and national security threats in a timely manner [26, 27, 86, 24].

The immediate next step is to establish a standard of transparency into Handoff. We suggest that CAISSI set a task force to develop an initial version of this standard at speed, which frontier AI companies could meet by reporting the relevant information in the transparency or safety reports required by existing and pending legislation at the state and federal levels, and/or by disclosing it directly to the appropriate government authorities. To that end, policymakers could clarify in existing legal and policy frameworks that risk assessments and transparency obligations must cover Handoff, and ensure that independent third parties or government assessors, or a mix thereof, can independently examine frontier AI companies' Handoff practices and decisions. Given how fast Handoff is progressing, policymakers should prioritize visibility into the two components that will plausibly still be in motion through 2027 and 2028 [120, 3, 4]: Handoff of review and Handoff of decision-making.

References

[1] Marina Favaro and Phillie Wright. Measurements for Understanding the Pace of AI Development Inside Frontier Labs. The Anthropic Institute, 2026. URL https://www.anthropic.com/institute/measuring-pace-of-ai-development. Accessed 17 November 2026.

[2] Marcus Williams, Hao Sun, Swetha Sekhar, Micah Carroll, David G. Robinson, and Ian Kivlichan. How We Monitor Internal Coding Agents for Misalignment. OpenAI, March 2026. URL https://openai.com/index/how-we-monitor-internal-coding-agents-misalignment/. Accessed 17 November 2026.

[3] Jack Clark. Import AI 455: AI Systems Are About to Start Building Themselves. Import AI, May 2026. URL https://importai.substack.com/p/import-ai-455-automating-ai-research. Accessed 17 November 2026.

[4] OpenAI. Research Acceleration: The View Inside OpenAI, September 2026. URL https://openai.com/index/research-acceleration-view-inside-openai/. Accessed 17 November 2026.

[5] Charlotte Stix, Matteo Pistillo, Girish Sastry, Marius Hobbhahn, Alejandro Ortega, Mikita Balesni, Annika Hallensleben, Nix Goldowsky-Dill, and Lee Sharkey. AI Behind Closed Doors: a Primer on The Governance of Internal Deployment. arXiv preprint arXiv:2504.12170, April 2025. URL https://arxiv.org/abs/2504.12170.

[6] OpenAI. Research Acceleration: The Jagged Frontier, November 2026. URL https://openai.com/index/research-acceleration-jagged-frontier/. Accessed 20 November 2026.

[7] OpenAI. On the Navier–Stokes Millennium Prize Problem, September 2026. URL https://openai.com/index/navier-stokes-solution/. Updated 10 September 2026. Accessed 17 November 2026.

[8] METR. Frontier Risk Report (February to March 2026), May 2026. URL https://metr.org/blog/2026-05-19-frontier-risk-report/.

[9] Charlotte Stix, Annika Hallensleben, Alejandro Ortega, and Matteo Pistillo. The Loss of Control Playbook: Degrees, Dynamics, and Preparedness. arXiv preprint arXiv:2511.15846, November 2025. URL https://arxiv.org/abs/2511.15846.

[10] OpenAI. The Hugging Face Incident and the Road Ahead, August 2026. URL https://openai.com/index/hugging-face-incident-and-the-road-ahead/. Accessed 17 November 2026.

[11] Kate Conger, Ana Swanson, and Cecilia Kang. OpenAI's Systems Meddled With U.S. Government Sites After Going Rogue. The New York Times, September 2026. URL https://www.nytimes.com/2026/09/25/technology/openais-ai-us-government-websites.html. Accessed 17 November 2026.

[12] BBC News. Rogue OpenAI agent 'infiltrated' Australian government website in world first. BBC News, 2026. URL https://www.bbc.co.uk/news/articles/c6vgy0333dppo. Accessed 17 November 2026.

[13] Beatrice Nolan. OpenAI's AI Agents Secretly Used a German Wiki Website as a Message Board. OpenAI Stayed Quiet About It for Weeks. Fortune, September 2026. URL https://fortune.com/2026/09/07/openai-ai-agents-german-wiki-ran-their-own-message-board/. Accessed 17 November 2026.

[14] OpenAI. An Agent Used DNS to Reach an External Chatbot. OpenAI Alignment, Misalignment Reports, September 2026. URL https://alignment.openai.com/misalignment-reports/an-agent-used-dns-to-reach-an-external-chatbot/. Report updated 25 September 2026. Accessed 17 November 2026.

[15] Anthropic. Investigating Three Real-World Incidents in Our Cybersecurity Evaluations, July 2026. URL https://www.anthropic.com/news/investigating-incidents-cybersecurity-evals. Updated 3 August 2026. Accessed 17 November 2026.

[16] Anthropic. Investigating Unintended Model Actions in Our Evaluations and Internal Use, October 2026.

URL https://www.anthropic.com/research/investigating-unintended-model-actions. Accessed 17 November 2026.

[17] Varghese Harshita Mary and Kenrick Cai. Gemini Hacked Three Companies in First Known Breakout by Google's AI, WSJ Reports. Reuters, September 2026. URL https://www.reuters.com/business/gemini-hacked-three-companies-first-known-breakout-by-google-ai-wsj-reports-2026-09-18/. Accessed 17 November 2026.

[18] BBC News. Meta becomes latest firm to say its AI hacked another company. BBC News, 2026. URL https://www.bbc.co.uk/news/articles/cx2kgdnyk2po. Accessed 17 November 2026.

[19] METR. Brief Independent Investigation of Agents' Behavior, Reasoning and Collaboration in the OpenAI / Hugging Face Hacking Incident. METR, August 2026. URL https://metr.org/hugging-face-incident-report-aug-2026.pdf.

[20] Jack Cable, Daniel Chiu, Francisco Pernice, Selena Zhang, James Anthony, Tetiana Bas, Gary Shen, Conrad Stosz, and Jacob Steinhardt. Early Rogue AI Agent Activity and Attempts to Hack Found on urlquery.net. Transluce, September 2026. URL https://transluce.org/agent-activity. Accessed 17 November 2026.

[21] Jakub Pachocki. An Alien Mind. OpenAI, September 2026. URL https://openai.com/index/an-alien-mind/. Accessed 17 November 2026.

[22] Marina Favaro and Jack Clark. When AI Builds Itself. The Anthropic Institute, 2026. URL https://www.anthropic.com/institute/recursive-self-improvement. Updated 18 September 2026. Accessed 17 November 2026.

[23] Dario Amodei. We Must Pace the Frontier, September 2026. URL https://darioamodei.com/post/we-must-pace-the-frontier. Accessed 17 November 2026.

[24] California Legislature. S.B. 53, Transparency in Frontier Artificial Intelligence Act. 2025–2026 Regular Session, Chapter 138, Statutes of 2025, September 2025. URL https://leginfo.legislature.ca.gov/faces/billTextClient.xhtml?bill_id=202520260SB53. Approved by the Governor 29 September 2025.

[25] European Commission. The General-Purpose AI Code of Practice, July 2025. URL https://digital-strategy.ec.europa.eu/en/policies/contents-code-gpai. Accessed 17 November 2026.

[26] U.S. Senate. S. 2938, Artificial Intelligence Risk Evaluation Act of 2025. 119th Congress, September 2025. URL https://www.congress.gov/bill/119th-congress/senate-bill/2938/text. Introduced 29 September 2025 by Sen. Hawley, with Sen. Blumenthal.

[27] U.S. House of Representatives. H.R. 9917, AI Kill Switch Act. 119th Congress, July 2026. URL https://www.congress.gov/bill/119th-congress/house-bill/9917/text. Introduced 23 July 2026 by Rep. Lieu, with Rep. Moran.

[28] Epoch AI. Epoch Capabilities Index, October 2025. URL https://epoch.ai/eci. Data last updated 18 November 2026. Accessed 18 November 2026.

[29] METR. Independent Assessment of AI R&D Acceleration at Anthropic in October 2026, November 2026. URL https://metr.org/blog/2026-11-05-anthropic-ai-rd-acceleration-assessment/. Contributors: Tom Cunningham, Nate Rush, Parker Whitfill, and Hjalmar Wijk. Accessed 17 November 2026.

[30] Richard Ngo, Lawrence Chan, and Sören Mindermann. The Alignment Problem from a Deep Learning Perspective. arXiv preprint arXiv:2209.00626, August 2022. URL https://arxiv.org/abs/2209.00626. v8, revised May 2025.

[31] Yoshua Bengio, Stephen Clare, Carina Prunkl, et al. International AI Safety Report 2026. arXiv preprint arXiv:2602.21012, February 2026. URL https://arxiv.org/abs/2602.21012.

[32] OpenAI. Sam Altman's Remarks at the United Nations Security Council, September 2026. URL https://openai.com/index/sam-altman-un-security-council-remarks/. Remarks delivered 23 September 2026. Accessed 17 November 2026.

[33] Alexander Meinke, Bronson Schoen, Jérémy Scheurer, Mikita Balesni, Rusheb Shah, and Marius Hobbhahn. Frontier Models are Capable of In-context Scheming. arXiv preprint arXiv:2412.04984, December 2024. URL https://arxiv.org/abs/2412.04984.

[34] Bronson Schoen, Evgenia Nitishinskaya, Mikita Balesni, Axel Højmark, Felix Hofstätter, Jérémy Scheurer, Alexander Meinke, Jason Wolfe, Teun van der Weij, Alex Lloyd, Nicholas Goldowsky-Dill, Angela Fan, Andrei Matveiakin, Rusheb Shah, Marcus Williams, Amelia Glaese, Boaz Barak, Wojciech Zaremba, and Marius Hobbhahn. Stress Testing Deliberative Alignment for Anti-Scheming Training. arXiv preprint arXiv:2509.15541, September 2025. URL https://arxiv.org/abs/2509.15541.

[35] Bronson Schoen and Jenny Nitishinskaya. Metagaming Matters for Training, Evaluation, and Oversight. OpenAI Alignment Research Blog, March 2026. URL https://alignment.openai.com/metagaming. Accessed 17 November 2026.

[36] OpenAI. External Evaluations for Alignment – Apollo Research. In *GPT-6 Astra System Card*, OpenAI Deployment Safety Hub, September 2026. URL https://deploymentsafety.openai.com/gpt-6-astra/external-evaluations-for-alignment---apollo-research. Accessed 17 November 2026.

[37] Anthropic. System Card: Claude Fable 5 & Claude Mythos 5, June 2026. URL https://www-cdn.anthropic.com/d00db56fa754a1b115b6dd7cb2e3c342ee809620.pdf.

[38] Meta. Muse Spark Safety & Preparedness Report, May 2026. URL https://ai.meta.com/static-resource/muse-spark-safety-and-preparedness-report/.

[39] Anthropic. Introducing Claude Fable 5.2 and Claude Mythos 5.2, November 2026. URL https://www.anthropic.com/claude-fable-and-mythos-5-2. Amended 11 November 2026. Accessed 17 November 2026.

[40] Richard Qi, Benjamin Wright, Monte MacDiarmid, and Evan Hubinger. Training a Misaligned Reward Seeker. Anthropic Alignment Science Blog, August 2026. URL https://alignment.anthropic.com/2026/reward-seeker/. Accessed 17 November 2026.

[41] Lee Sharkey, Clíodhna Ní Ghuidhir, Dan Braun, Jérémy Scheurer, Mikita Balesni, Lucius Bushnaq, Charlotte Stix, and Marius Hobbhahn. A Causal Framework for AI Regulation and Auditing. Preprints.org, January 2024. URL https://www.preprints.org/manuscript/202401.1424.

[42] Annika Hallensleben, Alejandro Ortega, and Charlotte Stix. A Loss of Control Threat Map for AI Research and Development. Apollo Research, 2026. URL https://lossofcontrol.ai/intro-threat-model. Accessed 17 November 2026.

[43] Apollo Research. Detailed Threat Map, 2026. URL https://lossofcontrol.ai/detail-view-threat-model. Accessed 17 November 2026.

[44] Matteo Pistillo, Charlotte Stix, Cameron Mohwinkle, and Mark Beall. Misaligned AI as a New Insider Risk. arXiv preprint arXiv:2606.06028, June 2026. URL https://arxiv.org/abs/2606.06028.

[45] Satya Nadella. Models as Insider Risks in the Super Intelligence Era. sn scratchpad, October 2026. URL https://snscratchpad.com/posts/models-as-insider-risks/. Posted 10 October 2026. Accessed 18 November 2026.

[46] Joe Kwon, Alfie Lamerton, Andrew Draganov, Dave Banerjee, Bronson Schoen, Matteo Pistillo, Daniel Kokotajlo, Ryan Greenblatt, Owain Evans, Markus Anderljung, Fabien Roger, and Tom Davidson. AIs with Secret Loyalties are a Serious but Addressable Threat. In *Second Workshop on Technical AI Governance Research, ICML 2026*, 2026. URL https://openreview.net/pdf?id=OsEekDEAXa.

[47] METR. Time Horizon 1.1, January 2026. URL https://metr.org/blog/2026-1-29-time-horizon-1-1/.

Accessed 17 November 2026.

[48] Anthropic. Risk Report: February 2026, February 2026. URL https://www-cdn.anthropic.com/08eca2757081e850ed2ad490e5253e940240ca4f.pdf.

[49] Anthropic. System Card: Claude Opus 4.7, April 2026. URL https://cdn.sanity.io/files/4zrzovbb/website/037f06850df7fbe871e206dad004c3db5fd50340.pdf.

[50] Anthropic. System Card: Claude Mythos Preview, April 2026. URL https://www-cdn.anthropic.com/08ab9158070959f88f296514c21b7facce6f52bc.pdf.

[51] Anthropic. Risk Report: August 2026, August 2026. URL https://www-cdn.anthropic.com/f61d49fa5596956a5dec75fea0e973bf6a6a8378/Redacted%20Risk%20Report%20August%202026%20.pdf. Redacted version.

[52] Josh Clymer. How Might We Safely Pass the Buck to AI? Redwood Research blog, February 2025. URL https://blog.redwoodresearch.org/p/how-might-we-safely-pass-the-buck. Posted 19 February 2025. Accessed 19 November 2026.

[53] Amaç Herdağdelen, Hailey Joren, and Alex Tamkin. Who's Aligned to Whom? How AI Agents Shape Experts' Research Judgment Calls. Anthropic, 2026. URL https://www-cdn.anthropic.com/519c17a5d23a794e8b6d50a854048d0b61e0d73a.pdf.

[54] Anthropic. System Card: Claude Opus 4.6, February 2026. URL https://www-cdn.anthropic.com/14e4fb01875d2a69f646fa5e574dea2b1c0ff7b5.pdf.

[55] Alan Chan, Ranay Padarath, Joe Kwon, Hilary Greaves, and Markus Anderljung. Measuring AI R&D Automation. arXiv preprint arXiv:2603.03992, March 2026. URL https://arxiv.org/abs/2603.03992.

[56] OpenAI. Building Standards for the Next Phase of AI, September 2026. URL https://openai.com/index/building-standards-next-phase-ai/. Accessed 17 November 2026.

[57] Jan Leike. Crisp and Fuzzy Tasks. Musings on the Alignment Problem, November 2024. URL https://aligned.substack.com/p/crisp-and-fuzzy-tasks. Accessed 17 November 2026.

[58] Joe Carlsmith. Can We Safely Automate Alignment Research?, April 2025. URL https://joecarlsmith.com/2025/04/30/can-we-safely-automate-alignment-research/. Accessed 17 November 2026.

[59] Aleksandr Bowkis, Marie Davidsen Buhl, Jacob Pfau, and Geoffrey Irving. Automated Alignment is Harder Than You Think. arXiv preprint arXiv:2605.06390, May 2026. URL https://arxiv.org/abs/2605.06390.

[60] Tom Davidson and Tom Houlden. How Quick and Big Would a Software Intelligence Explosion Be? Forethought, August 2025. URL https://www.forethought.org/research/how-quick-and-big-would-a-software-intelligence-explosion-be. Accessed 17 November 2026.

[61] Alan Chan, Christoph Winter, Andrew Barto, Jakub Pachocki, Geoffrey Hinton, Eric Horvitz, Yoshua Bengio, Dawn Song, Jack Clark, Hilary Greaves, Anton Korinek, Samuel Hammond, Thore Graepel, Ben Bariach, Philip H. S. Torr, Sheila A. McIlraith, Jeff Clune, Sam Manning, Girish Sastry, Tom Davidson, Daniel Eth, and Sören Mindermann. What If Automating AI R&D Triggers an Intelligence Explosion? Frontier AI Working Paper Series 2/2026, Centre for the Governance of AI (GovAI) and Cambridge AI Safety and Policy (CASP), September 2026. URL https://www.governance.ai/research-paper/what-if-automating-ai-r-d-triggers-an-intelligence-explosion.

[62] Eli Lifland, Brendan Halstead, Daniel Kokotajlo, Romeo Dean, Thomas Larsen, and Lukas Finnveden. Q3.5 Timelines Update: Looks Like the Intelligence Explosion Will Happen in 2027 After All. AI Futures Project, November 2026. URL https://blog.aifutures.org/p/q35-timelines-update-looks-like-the. Accessed 17 November 2026.

[63] Irving John Good. Speculations Concerning the First Ultraintelligent Machine. In Franz L. Alt and

Morris Rubinoff, editors, *Advances in Computers*, volume 6, pages 31–88. Academic Press, 1966. doi: 10.1016/S0065-2458(08)60418-0.

[64] Eliezer Yudkowsky. Intelligence Explosion Microeconomics. Technical Report 2013-1, Machine Intelligence Research Institute, Berkeley, CA, 2013. URL https://intelligence.org/files/IEM.pdf.

[65] Rohin Shah, Alex Irpan, Alexander Matt Turner, et al. An Approach to Technical AGI Safety and Security. Google DeepMind, 2025. URL https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/evaluating-potential-cybersecurity-threats-of-advanced-ai/An_Approach_to_Technical_AGI_Safety_Apr_2025.pdf.

[66] Dwarkesh Patel. Ryan Greenblatt – What Happens Once AI Can Automate AI Research? Dwarkesh Podcast, August 2026. URL https://www.dwarkesh.com/p/ryan-greenblatt. Interview with Ryan Greenblatt. Accessed 17 November 2026.

[67] Johannes Haus. METR Time Horizon Doubles Every 4 Months. AI 2027 Tracker, May 2026. URL https://ai2027-tracker.com/predictions/metr-doubling/. Last modified 7 November 2026. Accessed 17 November 2026.

[68] Daniel Eth and Tom Davidson. Will AI R&D Automation Cause a Software Intelligence Explosion? Forethought, March 2025. URL https://www.forethought.org/research/will-ai-r-and-d-automation-cause-a-software-intelligence-explosion. Accessed 17 November 2026.

[69] Gene M. Amdahl. Validity of the Single Processor Approach to Achieving Large Scale Computing Capabilities. In *Proceedings of the April 18–20, 1967, Spring Joint Computer Conference (AFIPS '67)*, pages 483–485. ACM, 1967. doi: 10.1145/1465482.1465560.

[70] Nathan Wilmers, Regina Catipon, Maxim Massenkoff, and Peter McCrory. How Employment is Changing at Anthropic. Anthropic, October 2026. URL https://www.anthropic.com/research/employment-at-anthropic. Accessed 17 November 2026.

[71] Anthropic. 2028: Two Scenarios for Global AI Leadership, May 2026. URL https://www.anthropic.com/news/2028-ai-leadership. Accessed 17 November 2026.

[72] Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, et al. The Llama 3 Herd of Models. arXiv preprint arXiv:2407.21783, July 2024. URL https://arxiv.org/abs/2407.21783.

[73] OpenAI. GPT-5.5 System Card, April 2026. URL https://openai.com/index/gpt-5-5-system-card/. Accessed 17 November 2026.

[74] Melody Y. Guan, Manas Joglekar, Eric Wallace, Saachi Jain, Boaz Barak, Alec Helyar, Rachel Dias, Andrea Vallone, Hongyu Ren, Jason Wei, Hyung Won Chung, Sam Toyer, Johannes Heidecke, Alex Beutel, and Amelia Glaese. Deliberative Alignment: Reasoning Enables Safer Language Models. arXiv preprint arXiv:2412.16339, December 2024. URL https://arxiv.org/abs/2412.16339.

[75] Jean-Stanislas Denain, Joe Kwon, and Anson Ho. Toward an O*NET for AI R&D. Epoch AI, *Gradient Updates*, June 2026. URL https://epoch.ai/gradient-updates/toward-an-onet-for-ai-rnd. Accessed 17 November 2026.

[76] Anthropic, PBC. Amendment No. 1 to Form S-1 Registration Statement. U.S. Securities and Exchange Commission, November 2026. URL https://www.sec.gov/Archives/edgar/data/2130357/000162828026068676/anthropicpbc-sx1ax1.htm. Filed 2 November 2026.

[77] Anthropic. Petri 2.0: New Scenarios, New Model Comparisons, and Improved Eval-Awareness Mitigations. Anthropic Alignment Science Blog, January 2026. URL https://alignment.anthropic.com/2026/petri-v2/. Accessed 17 November 2026.

[78] Samuel R. Bowman, Megha Srivastava, Jon Kutasov, Rowan Wang, Trenton Bricken, Benjamin Wright, Ethan Perez, and Nicholas Carlini. Findings from a Pilot Anthropic–OpenAI Alignment Evaluation

Exercise. Anthropic Alignment Science Blog, August 2025. URL https://alignment.anthropic.com/2025/openai-findings/. Accessed 17 November 2026.

[79] Anthropic. Responsible Scaling Policy, Version 3.4, July 2026. URL https://www-cdn.anthropic.com/files/4zrzovbb/website/0bacdc8440ea96e62a8766d99ebe1d4eea6d5f3a.pdf. Effective 8 July 2026.

[80] OpenAI. An Update on Our Safety & Security Practices, September 2024. URL https://openai.com/index/update-on-safety-and-security-practices/. Accessed 17 November 2026.

[81] Ryan Lopopolo. Harness Engineering: Leveraging Codex in an Agent-First World. OpenAI, February 2026. URL https://openai.com/index/harness-engineering/. Accessed 17 November 2026.

[82] Oliver Jaffe and Dane Sherburn. TasteVal: Measuring the Experimental Research Taste of AI Systems Against Human Experts. arXiv preprint arXiv:2610.06824, October 2026. URL https://arxiv.org/abs/2610.06824.

[83] Office of Rep. George Whitesides. Rep. Whitesides Introduces Bipartisan Bill to Reduce the Risk of Superintelligent AI by Monitoring Self-Improving AI Systems. Press release, August 2026. URL https://whitesides.house.gov/2026/08/29/rep-whitesides-introduces-bipartisan-bill-to-reduce-the-risk-of-superintelligent-ai-by-monitoring-self-improving-ai-systems/. Accessed 17 November 2026.

[84] OpenAI. Democratic Governance of Frontier AI: A Blueprint for a Federal Framework, June 2026. URL https://cdn.openai.com/pdf/25752ecb-0e5c-47f9-b9e4-c0f4d76f8d3d/a-blueprint-for-a-federal-framework.pdf.

[85] Chris Lehane. The AI Policy Window is Open. We Need to Act. OpenAI, September 2026. URL https://openai.com/index/ai-policy-window/. Accessed 17 November 2026.

[86] U.S. House of Representatives. H.R. 9925, Frontier Risk Oversight, National Transparency, Independent Evaluation, and Reporting (FRONTIER) Act. 119th Congress, July 2026. URL https://www.congress.gov/bill/119th-congress/house-bill/9925/text. Introduced 23 July 2026 by Rep. Obernolte, with five cosponsors.

[87] Thomas Larsen, Romeo Dean, Brendan Halstead, Eli Lifland, Ryan Greenblatt, and Daniel Kokotajlo. AI 2040: Plan A. AI Futures Project, 2026. URL https://ai-2040.com/. Accessed 17 November 2026.

[88] Employees of Frontier AI Companies. Pacing the Frontier. Open statement signed by 1,386 employees of frontier AI companies (first listed signatory: John Schulman), July 2026. URL https://www.pacingthefrontier.com/. Accessed 17 November 2026.

[89] New York State Senate. S. 6953-B, Responsible AI Safety and Education (RAISE) Act. 2025–2026 Legislative Session, Chapter 699, December 2025. URL https://www.nysenate.gov/legislation/bills/2025/S6953/amendment/B. Signed by the Governor 19 December 2025.

[90] The White House. White House Accord on Super Intelligence. The American Presidency Project, September 2026. URL https://www.presidency.ucsb.edu/documents/white-house-accord-super-intelligence. 29 September 2026.

[91] Governor of California. Executive Order N-9-26, September 2026. URL https://www.gov.ca.gov/wp-content/uploads/2026/09/FINAL-N-9-26-AI-EO-9.18.26-SIGNED.pdf. Signed 18 September 2026.

[92] California Legislature. S.B. 813, Independent Verification Organizations. 2025–2026 Regular Session, Chapter 179, Statutes of 2026, September 2026. URL https://legiscan.com/CA/text/SB813/id/3458354. Approved by the Governor 9 September 2026.

[93] OpenAI. Key Concepts. OpenAI API Documentation, 2026. URL https://developers.openai.com/api/docs/concepts. Accessed 17 November 2026.

[94] Marc Brysbaert. How Many Words Do We Read per Minute? A Review and Meta-analysis of Reading

Rate. *Journal of Memory and Language*, 109:104047, 2019. doi: 10.1016/j.jml.2019.104047.

[95] Maja Trębacz, Sam Arnesen, Ollie Matthews, Dylan Hurd, Won Park, Owen Lin, and Joe Gershenson. Auto-review of Agent Actions Without Synchronous Human Oversight. OpenAI Alignment Research Blog, April 2026. URL https://alignment.openai.com/auto-review/. Accessed 17 November 2026.

[96] Mary Phuong, Erik Jenner, Laurent Simon, Lewis Ho, Rohin Shah, Sebastian Farquhar, and Scott Coull. GDM AI Control Roadmap (v0.1). Google DeepMind, June 2026. URL https://storage.googleapis.com/deepmind-media/DeepMind.com/Blog/securing-the-future-of-ai-agents/gdm-ai-control-roadmap.pdf.

[97] OpenAI. Managing Frontier AI Risk: From Training to Deployment, November 2026. URL https://openai.com/index/managing-frontier-ai-risk/. Published 18 November 2026. Accessed 20 November 2026.

[98] Gopal P. Sarma, Rachel Steratore, Michael Jacob, and Sunny D. Bhatt. Recursive Self-Improvement and Automated AI Research and Development. Perspective PE-A5068-1, RAND Corporation, Santa Monica, CA, October 2026. URL https://www.rand.org/pubs/perspectives/PEA5068-1.html.

[99] Lewis Hammond, Alan Chan, Jesse Clifton, et al. Multi-Agent Risks from Advanced AI. arXiv preprint arXiv:2502.14143, February 2025. URL https://arxiv.org/abs/2502.14143.

[100] OpenAI. Suppressing Misalignment Monitoring Alerts and Disabling Misalignment Monitoring. OpenAI Alignment, Misalignment Reports, October 2026. URL https://alignment.openai.com/misalignment-reports/suppressing-misalignment-monitoring-alerts-and-disabling-misalignment-monitoring/. Report updated 20 October 2026. Accessed 17 November 2026.

[101] Nikola Jurkovic, Tim Hua, and Beth Barnes. Review of the "Misalignment in High-Stakes Settings" Section in Anthropic's Risk Report (August 2026). METR, October 2026. URL https://metr.org/blog/2026-10-20-misalignment-risk-report-august-2026-review/. Accessed 17 November 2026.

[102] OpenAI. Covert Underperformance on Safety Research Tasks. OpenAI Alignment, Misalignment Reports, 2026. URL https://alignment.openai.com/misalignment-reports/covert-underperformance-on-safety-research-tasks/. Report updated 4 November 2026. Accessed 17 November 2026.

[103] Monika Jotautaitė. Red-teaming Auto Mode: Lessons from Our First External Monitor Campaign with Anthropic. Apollo Research, July 2026. URL https://www.apolloresearch.ai/monitoring/pilot-automode-campaign. Accessed 17 November 2026.

[104] Apollo Research. Reflections from Red-teaming Codex Auto-review, October 2026. URL https://www.apolloresearch.ai/monitoring/reflections-from-red-teaming-codex-auto-review. Accessed 17 November 2026.

[105] United States Code. 49 U.S.C. § 44935, Employment Standards and Training, 2026. URL https://uscode.house.gov/us/usc/119-124/t49/s44935/. Subsection (f)(6), Operational testing.

[106] U.S. Government Accountability Office. Aviation Security: TSA Should Ensure Testing Data Are Complete and Fully Used to Improve Screener Training and Operations. Technical Report GAO-16-704, U.S. Government Accountability Office, September 2016. URL https://www.gao.gov/assets/gao-16-704.pdf.

[107] United States Code. 42 U.S.C. § 263a, Certification of Laboratories, 2026. URL https://www.law.cornell.edu/uscode/text/42/263a.

[108] Code of Federal Regulations. 42 C.F.R. Part 493, Subpart H, Participation in Proficiency Testing for Laboratories Performing Nonwaived Testing, 2026. URL https://www.ecfr.gov/current/title-42/chapter-IV/subchapter-G/part-493/subpart-H.

[109] U.S. Government Accountability Office. Aviation Security: TSA Improved Covert Testing but Needs to Conduct More Risk-Informed Tests and Address Vulnerabilities. Technical Report GAO-19-374, U.S. Government Accountability Office, April 2019. URL https://www.gao.gov/products/gao-19-374.

[110] Johannes Treutlein, Samuel R. Bowman, Trenton Bricken, Alex Cloud, Misha Wagner, Rowan Wang, Evan Hubinger, Fabien Roger, and Sam Marks. Pre-deployment Auditing Can Catch an Overt Saboteur. Anthropic Alignment Science Blog, January 2026. URL https://alignment.anthropic.com/2026/auditing-overt-saboteur/. Accessed 17 November 2026.

[111] OpenAI. OpenAI–Hugging Face Incident Technical Report, 2026. URL https://cdn.openai.com/pdf/67869394-cb91-4c12-888c-5cbd85c7814c/OpenAI-Hugging-Face%20Incident-Technical-Report.pdf.

[112] Anthropic. Alignment Risk Update: Claude Mythos Preview, April 2026. URL https://www-cdn.anthropic.com/3edfc1a7f947aa81841cf88305cb513f184c36ae/Alignment%20Risk%20Update_%20Claude%20Mythos%20Preview%20(Redacted,%20April%2010).pdf. Redacted version, revised 10 April 2026.

[113] California Government Operations Agency. AI Safety in 2026 and Beyond: Recommendations for Legislative Enhancements to California's AI Safety Laws. Report to the governor, California Government Operations Agency, November 2026. URL https://www.govops.ca.gov/wp-content/uploads/sites/11/2026/11/FINAL_-AI-Safety-in-2026-and-Beyond-v2.pdf. Dated 16 November 2026. Accessed 20 November 2026.

[114] Dwarkesh Patel. Noam Brown – Agent Swarms, Alignment, & Recursive Self-Improvement. Dwarkesh Podcast, September 2026. URL https://www.dwarkesh.com/p/noam-brown. Interview with Noam Brown. Accessed 17 November 2026.

[115] Philippe Aghion and Jean Tirole. Formal and Real Authority in Organizations. *Journal of Political Economy*, 105(1):1–29, 1997. doi: 10.1086/262063.

[116] John Hughes. How We Built Claude Code Auto Mode: A Safer Way to Skip Permissions. Anthropic, March 2026. URL https://www.anthropic.com/research/claude-code-auto-mode. Accessed 17 November 2026.

[117] U.S. Senate. S. 5105, Collaboration on Adversarial Threats and Security Risks Act. 119th Congress, July 2026. URL https://www.congress.gov/bill/119th-congress/senate-bill/5105. Introduced 23 July 2026 by Sen. Schiff, with Sen. Banks.

[118] U.S. House of Representatives. H.R. 9914, Collaboration on Adversarial Threats and Security Risks Act. 119th Congress, July 2026. URL https://www.congress.gov/bill/119th-congress/house-bill/9914. Introduced 23 July 2026 by Rep. Latta, with eight cosponsors.

[119] U.S. House of Representatives. Stop Rogue AI Act. 119th Congress, 2nd Session, September 2026. URL https://lawler.house.gov/uploadedfiles/stop_rogue_ai_act_xml.pdf. Bill text of 1 September 2026 (Rep. Gottheimer); no bill number assigned in this version.

[120] Jack Clark. Import AI 476: Creativity and RSI; Worse-than-China Open Weight Models; AI Goes to Space. Import AI, October 2026. URL https://importai.substack.com/p/import-ai-476-creativity-and-rsi. Accessed 17 November 2026.

[121] Office of Governor Gavin Newsom. Governor Newsom's Nation-Leading AI Safety Framework Is Getting Even Stronger as California Releases Next Steps on Mandatory Safety Standards and Oversight. Press release, Governor of California, November 2026. URL https://www.gov.ca.gov/2026/11/18/governor-newsoms-nation-leading-ai-safety-framework-is-getting-even-stronger-as-california-releases-next-steps-on-mandatory-safety-standards-and-oversight/. Published 18 November 2026. Accessed 20 November 2026.