

This document is prepared and submitted by Redwood Research, a nonprofit AI research group based in Berkeley, California.

This document is approved for public dissemination. The document contains no business-proprietary or confidential information. Document contents may be reused by the government in developing the 2025 National AI R&D Strategic Plan and associated documents without attribution.

Introduction

As AI systems become more powerful, they become attractive to use in more and higher-stakes contexts, but also become capable of causing greater harm. Assuring their actions do not cause harm — due either to intentional sabotage or to error — therefore becomes even more important.

Recently, a nascent research field, referred to as **AI control**, has arisen around the goal of controlling AI systems even if they use their capabilities to subvert the developer or deployer’s intentions. Alongside our pioneering work at Redwood Research, research has been conducted on AI control at several frontier AI companies¹ and at the United Kingdom’s AI Security Institute.²

AI control has shown promising results in, for example, cheaply reducing the risk that an AI model writes insecure code. However, the field is new and would benefit from accelerated R&D efforts and a broader base of expertise in related areas. To enable organizations across America to securely develop and deploy AI systems, we suggest funding work in the areas described below.

Components of AI control

To provide context for the areas of R&D we suggest funding below, we first offer an overview of the components of AI control.

In any particular case, AI control is implemented as a certain control **protocol**, i.e. the set of measures put in place to reduce the risk of the AI system successfully causing a

¹ Bowen Baker et al., “Monitoring Reasoning Models for Misbehavior and the Risks of Promoting Obfuscation,” arXiv, March 14, 2025. <https://arxiv.org/abs/2503.11926>; Evan Hubinger et al, “Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training,” Anthropic, January 17, 2024. <https://www.anthropic.com/research/sleeper-agents-training-deceptive-llms-that-persist-through-safety-training>

² UK AI Security Institute, “Our Research Agenda: Control,” webpage, undated. <https://www.aisi.gov.uk/research-agenda#control>

harmful outcome. At the most basic level, this includes monitoring the actions of AI systems to be able to detect potentially harmful actions. It can also involve a range of other measures including intervening in the AI's actions or auditing the system. To create and improve control protocols, AI control researchers must develop a **setting** in which to implement the protocol. A “setting” in the context of control research is a context in which an AI system is given a task where it might cause harm, and a control protocol is implemented to prevent the harm from occurring. For example, the task might be to write code to solve some specified problems, the harm being studied could be writing subtly insecure code, and the control protocol could involve having a human or another AI system judge whether generated code might be insecure. To strengthen a control protocol, researchers try to find ways an AI system might circumvent the control protocol, a process called **red-teaming**. Once they have found weaknesses in the control protocol, they can then try to develop techniques that defend against those weaknesses. This process is comparable to the use of red-teaming in cybersecurity to strengthen an organization's security posture.

For AI control to play an effective role in overall management of AI risks, we also need to understand when a given protocol is worth using. This depends on understanding the capabilities of the AI system to be controlled, and how effectively the protocol reduces risk. **Model evaluations** seek to assess an AI system's capabilities and other relevant properties, such as robustness. **Protocol evaluations** seek to assess how effective a protocol is at preventing harmful action, at what cost.³

Priority topics for R&D

Strong control protocols are the ultimate goal of AI control research. But their utility must be validated via experiments in realistic settings, with challenging red teams.

Techniques to strengthen control protocols are also often best developed through iterated interaction with a red team in a given setting, just as wargames with strong red teams help to develop better operational plans. Currently, there is a need to promote R&D across all aspects of the control research landscape, including settings, protocols and red teaming.

Settings

To develop control protocols that are most useful, settings for testing and development should be diverse, to cover the range of scenarios AI might be deployed in, and as realistic as possible, to ensure generalization to real world use. UKAIS is developing

³ Protocol evaluations are also used to develop cheaper techniques at the same level of risk reduction, similar to how red teaming is used to discover stronger techniques.

ControlArena,⁴ a suite of settings for conducting control research which seeks to mimic deployment environments for AI systems. More work could be done along these lines to expand the coverage of settings available to researchers.⁵ We recommend the following areas of R&D for settings:

- **Concentrated failure settings:** This refers to settings where even a small number of unsafe actions could cause significant harm, such as if an AI could circumvent security measures.
- **Non-concentrated failure settings:** This refers to settings where a larger number of unsafe actions is required to cause significant harm, such as subtle sabotage of important work that the AI is being used for. This could include settings analogous to that of an AI being used through an API.
- **Varying the rules of the setting:** Rules established in a given setting define the game that the red team and blue team are playing. Varying these rules can lead to a broader understanding of the dynamics of a particular setting or category of settings.
- **Complex real-world settings:** More ambitiously, settings could be developed that seek to capture complex dynamics of the real world, such as settings where multiple AI agents interact as they might on the open internet as large numbers of agentic AIs are increasingly deployed in the real world.

The best way to ensure control research generalizes to the real world is to use real-world data or even real current AI production environments. We thus also recommend research on:

- **Settings based on real-world data:** This could involve gathering data on real-world AI usage, particularly where the downside risk of AI misbehavior is especially severe or where the misbehavior might be particularly useful to understand.
- **Experiments on real world AI:** This could include code generation settings with AI operating with various degrees of autonomy over various lengths of task, settings where AI is used to generate data, construct training or evaluation environments or acting as a classifier or “judge,” or settings where researchers working on real projects rely on AI even more heavily than they usually would for the explicit purposes of the experiment.

⁴ “ControlArena,” Github, undated. <https://github.com/UKGovernmentBEIS/control-arena>

⁵ For example, UKAISI has noted that they plan to expand ControlArena by adding realistic identity access management (IAM) and post-training codebase environments, as well as adding more “main” and “side” tasks in these environments, i.e. the overt task the AI is given in the setting and the harmful behavior it might seek to execute, respectively.

Techniques and protocols

Relatively simple protocols have already shown promise for enabling more secure AI deployment at relatively low cost in certain contexts. But the AI community will need more techniques as tools in its toolbox as systems develop new and stronger capabilities and are used in new ways and applications. This includes means for monitoring AIs, the role of human participants in control protocols, and techniques to prevent collusion or impede certain AI abilities.

The basic function around which control protocols are built is monitoring of an AI system's activities. In theory, all aspects of an AI system's development and deployment could be monitored; in practice, all monitoring comes with some cost, so developing methods that optimize the tradeoff of cost and effectiveness is core to improving AI control. Beyond merely being aware of an AI's activities, we also want to be able to understand them and intervene in one way or another if necessary. We recommend the following areas of R&D to improve monitoring:

- **Using current models for monitoring:** Current AI models are not capable enough to engage in successful scheming behavior,⁶ meaning they are more trustworthy on this particular axis than future models which may be more capable of strategic behavior. Research could explore means to make current models even more useful for monitoring, and other tasks related to controlling more capable AIs.
- **Preventing collusion:** AIs might seek to collude with one another if they share motivations that conflict with their developer or deployer, or with humans who may be accomplices. Some initial methods to prevent collusion include monitoring for unnecessary or unusual outputs that could be collusion signals, and testing monitors with known attacks to verify that the monitoring system does flag true positives. One important area of collusion prevention is detecting and preventing steganography.⁷ Steganography refers to the practice of communicating concealed information by encoding it within superficially unrelated information. Research could investigate both detecting steganography and preventing it. Techniques to prevent AIs from secretly colluding via steganography could include paraphrasing inputs and outputs to an AI model to “scrub” them of hidden information, or certain types of regularization. There may also be ways to modify training processes (particularly the reinforcement learning

⁶ Joe Carlsmith, “Scheming AIs: Will AIs fake alignment during training in order to get power?” arXiv, November 2023, <https://arxiv.org/abs/2311.08379>

⁷ Yohan Mathew et al., “Hidden in Plain Text: Emergence & Mitigation of Steganographic Collusion in LLMs,” arXiv, October 2, 2024. <https://arxiv.org/abs/2410.03768>

process) such that it is less likely that AI models develop the ability to use steganography.

- **Best practices from cybersecurity:** The problem of controlling AIs that are not fully trustworthy is in some ways analogous to other security challenges with computer systems. Exploring how best practices in cybersecurity might apply to controlling AI may allow the application of existing tools and insights to solve challenges with AI control. This could include practices such as threat hunting and the use of security information and event management (SIEM) systems among others. Other domains with similar adversarial challenges, such as counterintelligence, fraud prevention or content moderation, may also offer lessons.
- **Detecting increased misbehavior:** If the rate of misbehavior or suspicious behavior of an AI system increases for any reason, the developer or deployer would likely want to understand why and take remedial action. R&D could prototype systems for detecting and responding to increased misbehavior in various different settings, for different kinds of threats.
- **Chain of Thought monitoring and interpretability techniques:** Chain of Thought (CoT) is an important means to understand the activity of so-called “reasoning” models, which represent the current state of the art of LLM capabilities. Monitoring of CoT or other interpretability techniques may be even more important for more advanced agentic AI systems.⁸
- **Human auditing:** Many protocols will rely on some human components, such as human review of some percentage of incidents flagged as suspicious. Determining how best to employ humans and set them up for success will require research. This could be done in various settings with determined rules to analyze concrete tradeoffs, or even potentially without a specific setting. Research on human components of control protocols could also examine conditions for humans to be successful, such as what tools, information, skills and other resources they have access to.
- **Synthetic input generation:** To evaluate whether AI systems behave differently in deployment as in testing, it would be useful to be able to generate input data for the testing environment such that the AI is unable to distinguish it from real-world deployment, and thus behaves as it truly would in deployment. There are various ways this could be accomplished, including by redacting information or modifying real-world input data such that it appears to also be synthetic. Testing models in this way would be particularly useful for models that may be engaged in scheming.

⁸ Yanda Chen et al., “Reasoning Models Don’t Always Say What They Think,” Anthropic, April 3, 2025.
<https://www.anthropic.com/research/reasoning-models-dont-say-think>

- **Modifying model capabilities:** Not all capabilities that an AI model possesses are necessary for all tasks, so restricting some capabilities in contexts where those capabilities pose higher risks without significant benefit could be useful. Techniques that could be useful for restricting capabilities in a modular, targeted manner include unlearning,⁹ gradient routing, and using specialization techniques such as LoRA to select which capabilities are activated for any given task. Research in this area should also examine how general capabilities are affected by these techniques to ensure minimal tradeoff between utility and risk reduction.
- **Elicitation without learning:** Understanding the risks a model itself poses requires understanding the full extent of what it is capable of with no additional training. However, there is no known method today to ensure a model demonstrates its full capabilities via prompting alone, with no finetuning. Research could seek to develop methods for this.
- **Monitoring model internals:** The internal workings of large deep learning models are not interpretable by direct human observation. However, using machine learning techniques on the model itself can help to detect signals of potential risk in models. Techniques in this category that could benefit from additional research include “probes,” and anomaly detection.¹⁰
- **Convincing AIs of false facts:** In some cases, AIs may engage in harmful behavior only when certain facts about their situation are true, as has been demonstrated possible with research on “sleeper agents.”¹¹ It would be useful for defenders to be able to cause AIs to believe these facts are not true, even when they are. This could include, for example, believing that they are being monitored intensely even when they are not, or believing that they are in an environment different from what their target environment might be. For instance, if U.S. government users were concerned a model might seek to specifically sabotage U.S. government activities, it would be useful to be able to make the model

⁹ Jiahui Geng et al., “A Comprehensive Survey of Machine Unlearning Techniques for Large Language Models,” arXiv, February 22, 2025. <https://arxiv.org/abs/2503.01854>

¹⁰ For an example of the use of “probes,” see: Monte MacDiarmid et al., “Simple probes can catch sleeper agents,” Anthropic, April 23, 2024, <https://www.anthropic.com/research/probes-catch-sleeper-agents>. For an overview of one approach to anomaly detection, see: Erik Jenner, “A gentle introduction to mechanistic anomaly detection,” LessWrong, April 2024, <https://www.lesswrong.com/posts/n7DFwtJvCzkuKmtbG/a-gentle-introduction-to-mechanistic-anomaly-detection>

¹¹ This is also analogous to the famous Stuxnet cyber operation, which executed its payload only when it detected that it was in a highly specific target environment. If the defenders had been able to deceive the malware to perceive that it was in, for instance, an oil refinery rather than a nuclear enrichment facility, it never would have achieved its intended effect. On “sleeper agents,” see: Evan Hubinger et al., “Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training,” Anthropic, January 17, 2024. <https://www.anthropic.com/research/sleeper-agents-training-deceptive-llms-that-persist-through-safety-training>; Shuli Jiang et al., “Forcing Generative Models to Degenerate Ones: The Power of Data Poisoning Attacks,” arXiv, December 7, 2023. <https://arxiv.org/abs/2312.04748>

believe it was being used in a private-sector organization, or even a different government.

- **Exploration hacking countermeasures:** Exploration hacking refers to a phenomenon where a scheming model may avoid demonstrating its full capabilities when being trained in a way intended to elicit its capabilities. This could be an issue as such a model may appear less capable, and thus less risky, than it actually is. Research could examine methods to prevent models from doing so successfully, particularly with reasoning models or on agentic tasks.

Evaluations

To know when to use which control protocols, we need to have reliable evaluations of both the capabilities of AI systems and the effectiveness of control protocols. We recommend the following research directions related to evaluations:

- **Situational awareness:** Situational awareness evaluations would measure an AI system's ability to distinguish between testing environments and deployment or between synthetic and natural data, to control their own internals and outputs, and other capabilities relevant to scheming.
- **Deception and subversion capabilities:** An AI's level of capabilities to deceive and subvert control measures determines what measures are necessary to keep it under control. Research could examine how well models can evade detection from (either with or without a hidden scratchpad) or jailbreak the monitors applied to them. It could also study how effectively and how reliably models can develop strategies to subvert control techniques, under different conditions (such as whether the model is allowed to openly reason about subversion in its Chain of Thought (CoT), as if its CoT is not being monitored), and how well models can reason about their goals in a way that is not transparent in their CoT.
- **Collusion capabilities:** Whether an AI is able to collude, especially without overt, transparent communication with accomplices, is important to understand Capabilities steg, acausal coordination/schelling points
- **Automated red-teaming:** To facilitate the process of creating defenses in response to red teaming, systems including AI models could be developed that attempt to break control protocols. This could also help to study generalization performance of AI models in control settings. This is necessary in particular because current models are not optimized by default to conduct deceptive activity as is necessary for challenging control protocols.

Implementation

To be useful in the real world, AI control must be implementable by companies developing and deploying AI. R&D could help to identify, understand and resolve practical challenges in using AI control in the real world. Potentially, public-funded R&D could even openly distribute code or other artifacts that enable efficient uptake by a broad spectrum of organizations, catalyzing secure diffusion of AI across sectors in the U.S. We recommend the following R&D related to applications:

- **Software infrastructure for control:** Control protocols, as with all AI systems, will rely on an architecture of traditional software to enable their functions. This includes systems for logging inputs to and outputs from AIs, tracking which AIs were involved in which jobs and experiments, and scaffolding that supports interventions in AI actions (including resampling from the same model or deferring to a different model) and tracking other information relevant to auditing.
- **Security measures for control:** To be effective, control protocols will also rely on the AI system being effectively contained such that it cannot simply hack its way around the control protocol, rather than needing to somehow subvert it. Some security needs include secure environments for control experiments, permission management for control protocols, limiting AIs' access to information in their environment, and mechanisms to shutdown AI systems quickly and securely in an emergency.
- **Near-term applications:** Though Redwood focuses largely on risks from misaligned AI, a control-style approach is useful for a variety of different AI challenges, and applying control to these could help the field mature in general. This includes using control techniques for preventing misuse of currently deployed AI, defending against prompt injections, improving robustness or detecting errors in general. In particular, defending against backdoors, malicious functionality inserted into a model by an adversary via methods such as data poisoning,¹² is partially analogous to defending against a misaligned model and is thus one area where applications of control could be beneficial now while also providing utility in preparing for future risks.¹³
- **Real-world experiments:** As noted above, experiments could be done on real-world current usage of AI to help identify and resolve challenges to

¹² Evan Hubinger et al, "Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training," Anthropic, January 17, 2024.

<https://www.anthropic.com/research/sleeper-agents-training-deceptive-llms-that-persist-through-safety-training>; Shuli Jiang et al., "Forcing Generative Models to Degenerate Ones: The Power of Data Poisoning Attacks," arXiv, December 7, 2023. <https://arxiv.org/abs/2312.04748>

¹³ Paul Christiano, Jacob Hilton, Victor Lecomte and Mark Xu, "Backdoor defense, learnability and obfuscation," arXiv, November 18, 2024. <https://arxiv.org/pdf/2409.03077>

real-world implementation, and get practice with both technical and institutional systems for implementing control protocols effectively.

Conclusion

The above represents a brief overview of areas we consider high priorities for advancing the science and technology of AI control. We are eager for the opportunity to promote a broader and deeper base of research on AI control, and would be happy to engage with OSTP or potential AI control researchers to provide further thoughts based on our work in this field. We view this as one of the most promising avenues to ensure that powerful AI systems that may be created in the next few years can be used profitably and safely. Thank you for your attention to our comment.