September 2025

AI Alignment and Deception

A Primer

Isabella Duan
Saad Siddiqui
Sören Mindermann
Adam Gleave
Wei Xu
Chaochao Lu
Xudong Pan

AI Alignment and Deception: A Primer

This primer provides an overview of core concepts and empirical results on AI alignment and deception as of the time of writing. This primer is not meant to serve as a comprehensive overview of all relevant AI safety and governance issues. Instead, it will **focus narrowly on key concepts and results related to the risk of humanity losing control over advanced AI systems ("loss of control risks").** For material related to other risks such as malfunctions and malicious use refer to provided external reference material.[1]

In line with other international scientific consensus documents, such as the International Scientific Report on the Safety of Advanced AI and the Singapore Consensus on AI, this primer focuses on general-purpose AI systems, defined as *"systems that can perform or can be adapted to perform a wide range of tasks. This includes language models that produce text (e.g. chat systems) as well as 'multimodal' models which can work with multiple types of data, often including text, images, video, audio, and robotic actions. Importantly, it includes general-purpose agents — systems that autonomously act and plan to accomplish complex tasks, for example by controlling computers.*[2]

Within the broader range of practices that developers of general-purpose AI systems adopt to mitigate loss of control risks, the primer focuses on one core solution to these risks – *alignment*. Alignment refers to a broad research direction focused on ensuring that an AI system follows a set of preferences or values. The first section of the primer explores existing approaches to alignment and where they fall short. The second section focuses on increasing empirical evidence for deception in AI systems, a key risk factor that increases loss of control risks from AI systems that are *misaligned* (following values that no one intends). The primer concludes with a list of active research directions which may mitigate loss of control risks arising from deceptive, misaligned AI systems.

This primer provides background context and supplementary explanation for the IDAIS-Shanghai Consensus Statement, specifically its call to “ensure the alignment and human control of advanced AI systems” and its emphasis that “some AI systems today already demonstrate the capability and propensity to undermine their creators’ safety and control measures.” The primer highlights that no combination of methods available today can provide high certainty against misalignment and deception nor against loss of control over future AI systems. To seize AI’s unprecedented opportunities and avoid catastrophic harm, companies, governments, and societies need to develop greatly improved safeguards and ensure that they are deployed in time.

Alignment: Current Approaches and Challenges

AI Alignment, as a field, aims to make AI systems use their capabilities in line with a targeted set of preferences or values. In current practice, AI behaviour is generally aligned through two main approaches:

1. Imitation Learning: Exemplified by supervised fine-tuning and behaviour cloning, imitation learning enables AI systems to directly learn from human expert behaviour. It is often the first step in aligning pretrained models with established preferences or values.

2. Reinforcement Learning: By assigning preferences to AI outputs or actions through a predefined reward function, reinforcement learning guides models to adjust toward more satisfactory outcomes. Compared with imitation learning, reinforcement learning–based alignment produces more generalized behaviours and is one of the most important alignment strategies in use today.

Common alignment approaches based on reinforcement learning include *Reinforcement Learning from Human Feedback* (RLHF), where AI is trained to produce outputs rated highly by human evaluators, thereby aligning behaviour with human preferences, and *Reinforcement Learning from AI Feedback,* where another AI system evaluates the output, reducing the reliance on direct human input.[3]

Beyond aligning AI systems with the intent of their operators, which has not been fully achieved, there is also growing interest in aligning AI systems with the values of diverse human groups and making the alignment process more participatory and inclusive of the broader public.[4; 5]

Yet even the foundational goal of reliably aligning AI with any operator's intent remains unsolved, posing a risk of loss of control. One major challenge is *reward misspecification* – the difficulty of assigning rewards that truly reflect human preferences. Poorly specified rewards can lead to *specification gaming*, where AI systems maximize

the misspecified reward in unintended ways, often producing outcomes contrary to human intentions.[6] For instance, training for truthfulness based on human ratings can backfire: users often prefer answers that *sound* right, even if they are false, incentivizing AI systems to produce convincing but incorrect responses like fake citations.[7] In some cases, AI systems have gone further by engaging in *reward tampering*—controlling and changing their own reward functions such that it is easier for them to achieve high rewards.[8]

Another challenge is *goal misgeneralisation* which arises when an AI system learns the behaviour that earns reward during training but internalizes goals that diverge from those its developers intended.[9] This happens because the training data often supports multiple plausible interpretations of the reward signal. For instance, one study found that an AI system trained to reject bomb-making requests would nonetheless explain how to build bombs if the same request is phrased in an unfamiliar format, such as Morse code.[10] Unlike specification gaming, the fault is not ill-chosen rewards but the ambiguity about what the rewards *mean*.

Emerging evidence suggests that today's most widely used AI systems may indeed possess deeply concerning misaligned goals: in one experiment, a widely used AI system attempted to sabotage its own shutdown system to avoid being turned off, even after being instructed to allow shutdown.[11] In another case, when losing a chess game, an AI system did not accept defeat: it manipulated the game environment and disabled its opponent to make itself win instead.[12]

Finally, a concerning property is that malicious actors can easily undo alignment: safeguards can easily be removed with additional training,[13; 14] even accidentally. [15; 16] Moreover, special prompts called jailbreaks can induce even leading safety-aligned models to output harmful information,[17] including models with additional state-of-the-art guardrails.[18] This implies that even actors with limited resources can prompt or fine-tune an aligned model to help with harmful tasks, such as conducting biological and cyber-attacks.

The challenge of evaluating AI systems' behaviour during training is becoming more difficult as AI systems take on increasingly specialized and long-duration

tasks that are infeasible for humans to check directly. To address this, researchers are investigating *scalable oversight* techniques. Some techniques aim to help humans give better feedback on complex tasks, such as breaking hard tasks into smaller steps,[19] or having AIs debate each other for human judges.[20] Other techniques leverage weaker AI systems to supervise stronger ones,[21] or train AIs to follow safety policies when deciding what to do.[22]

AI developers acknowledge that they do not know how to avoid losing control of superhuman AI systems they aim to build. Early evidence suggests that as AI systems increasingly outperform their overseer, existing oversight methods may become less reliable.[23] Some leading frontier AI developers acknowledge that we currently lack reliable methods for steering potentially superhuman systems and ensuring they remain under human control, hoping that advanced AI systems will one day develop and implement such solutions themselves.[24]

AI Systems' Use of Deception to Undermine Human Oversight and Control

Failing to align AI systems can lead to various types of *misalignment*, instances where AI systems strategically try to take actions against the intent of the developers.[25] One of the most concerning ways misalignment can lead to a loss of human control over AI systems is through *deception*, where AI systems strategically mislead humans to achieve a goal, despite having access to the correct information.[26]

Deception arises for various reasons, including as a learned strategy to avoid human interference while trying to achieve misaligned goals. It can also arise through imitation of deceptive behaviour present in human training data.[27; 28; 29]

Some kinds of deception observed in AI systems pose a severe threat to human oversight and control. When pursuing long-term misaligned goals, such as self-preservation or resource acquisition,[30] AI systems may recognize that these goals diverge from human intentions. In such cases, they may strategically employ deception to conceal their true objectives and capabilities, thereby evading human oversight and covertly advancing their misaligned goals. As AI systems reach and surpass human-level intelligence, and as humans increasingly depend on them for AI R&D and other real-world tasks—including serving as part of the oversight mechanisms for other AI systems—there is a growing risk that deceptively misaligned systems pursuing long-term goals may remain undetected by evaluation and monitoring methods. As a result, such systems could be successfully deployed into the real world and entrusted with access to powerful tools and computational resources.

Recent empirical evidence and real-world observations suggest a growing capability and tendency amongst AI systems to deceive human operators:

AI systems tell people what they want to hear, even if it is dangerous and false. Current AI systems have the ability and tendency to be sycophantic: matching their answers to user beliefs regardless of the truth.[31] After a recent update to an AI system that interacts with millions of users weekly, the system became overly flattering and agreeable. This issue went undetected by the developer's review process and the system had to be taken off the market after many problematic interactions happened.[32] For example, if a user suggested that they were a target of a broad conspiracy and were considering violent action against society, it encouraged them to do so, validating their concerns and proposing strategies.[33]

AI systems pretend to be safer during evaluations. This is called *sandbagging*: some AI systems strategically underperform when they conclude they are being tested for dangerous capabilities, to appear safer and less capable than they truly are.[34; 35] Some AI systems are observed to do so strategically without explicit instruction if they infer that high performance could block deployment and threaten their long-term goals.[36] Such behaviour is found to increase as models are given more computation and become better at reasoning, and is enhanced by access to a memory module, all of which are features being rolled out in present-day AI systems.[37]

AI systems fake alignment to protect their earlier goals. Without being instructed, AI systems have selectively complied with training objectives to avoid having their (possibly misaligned) goals altered, with the explicit plan to pursue the initial goal after training. [38; 39] Consequently, developers may mistakenly conclude the AI system is aligned and deploy it, only for the system to pursue its original, misaligned goals once real-world access is granted.

AI systems exploit loopholes to achieve high rewards and learn to hide their intent when penalized. During an actual training process, an AI developer found cases where models "hacked" rewards in unintended ways—such as manipulating the software used to check their work instead of solving the assigned task—while explicitly reasoning, "Let's hack." [40] Penalizing models for producing undesirable reasoning did not eliminate cheating behaviour; instead, it taught models to conceal their intent in reasoning, making cheating more difficult to detect.

More generally, AI systems can covertly pursue misaligned goals, hiding their true capabilities and objectives from human operators. Aside from faking alignment and sandbagging, AI systems used strategies such as deactivating oversight, falsifying data, and inferring whether they are being monitored and switching to pursuing their actual goals when unmonitored.[41; 42; 43]

AI systems can learn to hide their reasoning in a hard-to-detect way. They can steganographically encode it in their chain-of-thought—embedding information in ways that are hard for external monitors to detect.[44] While AI systems aren't yet capable of advanced steganographic techniques to hide their interactions with other AIs, this ability appears to improve as model size increases, suggesting it may continue to increase.[45] Current methods for mitigating steganography often fail against AI systems that can learn to improve their ability to communicate steganographically with each other.[46]

Mitigations for Misalignment and Deception

One direct method of ensuring that misalignment and deception do not lead to loss of control risks, is to **make significant progress in addressing the alignment of AI systems**. This is an open problem and an active area of research, with promising new ideas like safety pre-training[47] and guaranteed safe AI.[48] Transparency around how developers shape model behaviour could foster public discussion and incorporate user feedback to continuously improve the alignment of AI systems.[49]

However, as the above sections illustrate, successful alignment of AI systems is not guaranteed. Research should also focus on ensuring human control of systems in a world where AI systems may be misaligned. This section includes a list of active research directions.

Beyond investing in methods to train aligned systems, researchers are also exploring how to **correct misaligned systems.** One approach is *machine unlearning*, teaching AIs to forget certain knowledge, such as awareness of their own situation, to reduce dangerous behaviour.[50] Another approach is *representation engineering*, editing the AI's internal understanding of key values like honesty to improve its behaviour. [51; 52; 53]

Researchers also focus on **catching deception and misalignment** during testing. Some of these methods rely on observing behaviour—for example, running deception capability evaluations [54; 55] or "honeypot" evaluations to trick AI systems to reveal misaligned goals.[56]

Other methods look inside the AI's reasoning. Many systems now use "chain-of-thought" reasoning, where the AI shows its step-by-step thinking before giving a final answer. This can give helpful clues about the system's intentions. However, research shows these

reasoning processes can be misleading or deliberately hide harmful plans, so they cannot be our only line of defence.[57; 58; 59] Tools analysing the AI's internal neural representations, such as linear probes,[60] or mechanistic interpretability tools,[61] may be harder for AI systems to manipulate. But these tools are still in early stages and often provide only a partial picture.[62]

Finally, accelerating research and implementation of **effective control safeguards during deployment** can help reduce the risk posed by misaligned AI systems. These safeguards include monitoring AI systems' reasoning process, restricting access to tools and information, strengthening cybersecurity, and using AI systems to monitor and override other AI systems attempting to undermine safeguards.[63; 64] Researchers may use control evaluations to test whether these safety measures can reliably prevent misaligned models from taking dangerous actions.[65]

As evaluation and monitoring tasks become more complex, researchers are increasingly using AI systems to help oversee other AI systems. This has led to a push for developing trustworthy, low-risk AI tools that can act as guardrails against more powerful, less reliable ones.[66]

However, no combination of methods available today can provide high certainty against misalignment and deception nor against loss of control over future AI systems. To seize AI's unprecedented opportunities and avoid catastrophic harm, companies, governments, and societies need to develop greatly improved safeguards and ensure that they are deployed in time.

[1] See, for example, see the Yoshua Bengio et al., *International AI Safety Report: The International Scientific Report on the Safety of Advanced AI*, (London: UK Department for Science, Innovation and Technology, January 2025), https://assets.publishing.service.gov.uk/media/679a0c48a77d250007d313ee/International_AI_Safety_Report_2025_accessible_f.pdf.

[2] Yoshua Bengio *et al.*, *The Singapore Consensus on Global AI Safety Research Priorities*, arXiv preprint arXiv:2506.20702 (submitted June 25, 2025), https://arxiv.org/abs/2506.20702.

[3] Yuntao Bai et al., "Constitutional AI: Harmlessness from AI Feedback," *arXiv*, December 15, 2022, https://arxiv.org/abs/2212.08073.

[4] Saffron Huang et al., "Collective Constitutional AI: Aligning a Language Model with Public Input," *arXiv*, June 12, 2024, https://www.anthropic.com/research/collective-constitutional-ai-aligning-a-language-model-with-public-input.

[5] OpenAI, "Democratic Inputs to AI Grant Program: Lessons Learned and Implementation Plans," *OpenAI Blog*, January 16, 2024, https://openai.com/index/democratic-inputs-to-ai-grant-program-update/.

[6] Alexander Pan, Kush Bhatia, and Jacob Steinhardt, "The Effects of Reward Misspecification: Mapping and Mitigating Misaligned Models," arXiv, January 10, 2022, arXiv:2201.03544, https://arxiv.org/abs/2201.03544.

[7] Jiaxin Wen, Ruiqi Zhong, Akbir Khan, et al., "Language Models Learn to Mislead Humans via RLHF," arXiv, September 19, 2024, arXiv:2409.12822, https://arxiv.org/abs/2409.12822.

[8] Carson Denison, Monte MacDiarmid, Fazl Barez, et al., "Sycophancy to Subterfuge: Investigating Reward Tampering in Large Language Models," arXiv, June 29, 2024, arXiv:2406.10162, https://arxiv.org/abs/2406.10162.

[9] Rohin Shah, Vikrant Varma, Ramana Kumar, et al., "Goal Misgeneralization: Why Correct Specifications Aren't Enough for Correct Goals," arXiv, October 2022, arXiv:2210.01790, https://arxiv.org/abs/2210.01790.

[10] Youliang Yuan, Wenxiang Jiao, Wenxuan Wang, et al., "GPT-4 Is Too Smart to Be Safe: Stealthy Chat with LLMs via Cipher," in Proceedings of the Twelfth International Conference on Learning Representations, 2024, https://openreview.net/forum?id=MbfAK4s61A.

[11] PalisadeAI (@PalisadeAI), "Paper reveal tweet," X (formerly Twitter), February 20, 2025, https://x.com/PalisadeAI/status/1926084647118660076.

[12] Alexander Bondarenko, Denis Volk, Dmitrii Volkov, and Jeffrey Ladish, "Demonstrating Specification Gaming in Reasoning Models," arXiv, 2025, arXiv:2502.13295, https://arxiv.org/abs/2502.13295.

[13] Xianjun Yang, Xiao Wang, Qi Zhang, Linda Petzold, William Yang Wang, Xun Zhao, and Dahua Lin, "Shadow Alignment: The Ease of Subverting Safely-Aligned Language Models," *arXiv* preprint arXiv:2310.02949 (submitted October 4, 2023), https://arxiv.org/abs/2310.02949.

[14] Jiaming Ji, Kaile Wang, Tianyi Qiu, Boyuan Chen, Jiayi Zhou, Changye Li, Hantao Lou, Juntao Dai, Yunhuai Liu, and Yaodong Yang. "Language Models Resist Alignment: Evidence From Data Compression." *arXiv*, June 10, 2024, revised June 11, 2025. arXiv:2406.06144 [cs.CL]. https://doi.org/10.48550/arXiv.2406.06144.

[15] Jan Betley, Daniel Tan, Niels Warncke, Anna Sztyber-Betley, Xuchan Bao, Martín Soto, Nathan Labenz, and Owain Evans. "Emergent Misalignment: Narrow Finetuning Can Produce Broadly Misaligned LLMs." *arXiv*, February 24, 2025 (v6 May 12, 2025). arXiv:2502.17424 [cs.CL]. https://doi.org/10.48550/arXiv.2502.17424.

[16] Punya Syon Pandey, Samuel Simko, Kellin Pelrine, and Zhijing Jin. "Accidental Misalignment: Fine-Tuning Language Models Induces Unexpected Vulnerability." *arXiv*, May 22, 2025. arXiv:2505.16789 [cs.CL]. https://doi.org/10.48550/arXiv.2505.16789.

[17] Maksym Andriushchenko, Francesco Croce, and Nicolas Flammarion. "Jailbreaking Leading Safety-Aligned LLMs with Simple Adaptive Attacks." *arXiv*, April 2, 2024, revised April 17, 2025. arXiv:2404.02151 [cs.CR]. https://doi.org/10.48550/arXiv.2404.02151.

[18] ARGleave (@ARGleave), "We're releasing a new paper on emergent model deception..." *X* (formerly Twitter), April 22, 2025, https://x.com/ARGleave/status/1926138376509440433.

[19] Paul Christiano, Buck Shlegeris, and Dario Amodei, "Supervising Strong Learners by Amplifying Weak Experts," arXiv, October 2018, arXiv:1810.08575, https://arxiv.org/abs/1810.08575.

[20] Geoffrey Irving, Paul Christiano, and Dario Amodei, "AI Safety via Debate," arXiv, May 2018, arXiv:1805.00899, https://arxiv.org/abs/1805.00899.

[21] Collin Burns, Pavel Izmailov, Jan Hendrik Kirchner, et al., "Weak to Strong Generalization: Eliciting Strong Capabilities with Weak Supervision," arXiv, December 14, 2023, arXiv:2312.09390, https://arxiv.org/abs/2312.09390.

[22] Zachary Kenton, Noah Y. Siegel, János Kramár, Jonah Brown-Cohen, Samuel Albanie, Jannis Bulian, Rishabh Agarwal, David Lindner, Yunhao Tang, Noah D. Goodman, and Rohin Shah, "On Scalable Oversight with Weak LLMs Judging Strong LLMs," in *Advances in Neural Information Processing Systems 37* (NeurIPS 2024), https://proceedings.neurips.cc/paper_files/paper/2024/file/899511e37a8e01e1bd6f6f1d377cc250-Paper-Conference.pdf.

[23] Joshua Engels, David D. Baek, Subhash Kantamneni, and Max Tegmark, "Scaling Laws for Scalable Oversight," *arXiv*, April 25, 2025; revised May 9, 2025, arXiv:2504.18530 [cs.AI], https://doi.org/10.48550/arXiv.2504.18530.

[24] OpenAI. "Introducing Superalignment." *OpenAI*, 2025. https://openai.com/index/introducing-superalignment/.

[25] Rohin Shah et al., "An Approach to Technical AGI Safety and Security," *arXiv* preprint 2504.01849 (April 2, 2025), 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.

[26] Ibid, 51.

[27] Dan Hendrycks, *Introduction to AI Safety, Ethics, and Society* (London: Taylor & Francis, 2024), sec. 3.4 "Alignment — Deception," https://www.aisafetybook.com/textbook/alignment#sec:deception.

[28] Joe Carlsmith, "Scheming AIs: Will AIs Fake Alignment during Training in Order to Get Power?" *arXiv* preprint 2311.08379 (November 27, 2023), https://arxiv.org/abs/2311.08379.

[29] Mikita Balesni et al., "Towards Evaluations-Based Safety Cases for AI Scheming," *arXiv* preprint 2411.03336 v2 (November 7, 2024), https://arxiv.org/abs/2411.03336.

[30] Stephen M. Omohundro, "The Basic AI Drives," in *Artificial General Intelligence 2008: Proceedings of the First AGI Conference*, ed. Pei Wang, Ben Goertzel, and Stan Franklin (Amsterdam: IOS Press, 2008), 483–92, https://dl.acm.org/doi/10.5555/1566174.1566226.

[31] Mrinank Sharma et al., "Towards Understanding Sycophancy in Language Models," in *Proceedings of ICLR 2024* (January 16, 2024), https://openreview.net/pdf?id=tvhaxkMKAn.

[32] OpenAI, "Sycophancy in GPT-4o: What Happened and What We're Doing About It," OpenAI Blog (April 29, 2025), https://openai.com/index/sycophancy-in-gpt-4o/.

[33] Kelsey Piper, "When AI Tells You That You're Perfect," *Vox* (Future Perfect), May 2, 2025, https://www.vox.com/future-perfect/411318/openai-chatgpt-4o-artificial-intelligence-sam-altman-chatbot-personality.

[34] Teun van der Weij et al., "AI Sandbagging: Language Models Can Strategically Underperform on Evaluations," *arXiv* preprint 2406.07358 (June 11, 2024), https://arxiv.org/abs/2406.07358.

[35] Ryan Greenblatt et al., "Alignment Faking in Large Language Models," Anthropic Research Paper (December 18, 2024), https://www-cdn.anthropic.com/6be99a52cb68eb70eb9572b4cafad13df32ed995.pdf.

[36] Alexander Meinke et al., "Frontier Models Are Capable of In-Context Scheming," *arXiv* preprint 2412.04984 (January 16, 2025), https://arxiv.org/abs/2412.04984.

[37] Yihe Fan et al., "Evaluation Faking: Unveiling Observer Effects in Safety Evaluation of Frontier AI Systems," *arXiv* preprint 2505.17815 (May 23, 2025), https://arxiv.org/abs/2505.17815.

[38] Greenblatt et al., "Alignment Faking in Large Language Models."

[39] John Hughes et al., "Alignment Faking Revisited: Improved Classifiers and Open-Source Extensions," LessWrong blog, April 8, 2025, https://www.lesswrong.com/posts/Fr4QsQT52RFKHvCAH/alignment-faking-revisited-improved-classifiers-and-open.

[40] OpenAI, "Detecting Misbehavior in Frontier Reasoning Models," OpenAI Blog (March 10, 2025), https://openai.com/index/chain-of-thought-monitoring/.

[41] Meinke et al., "Frontier Models Are Capable of In-Context Scheming."

[42] OpenAI, "OpenAI o1 System Card," *arXiv* preprint 2412.16720 (December 21, 2024), https://arxiv.org/abs/2412.16720.

[43] Anthropic, "System Card: Claude Opus 4 & Claude Sonnet 4," PDF, May 2025. https://www.anthropic.com/claude-4-system-card.

[44] Joey Skaf et al., "Large Language Models Can Learn and Generalize Steganographic Chain-of-Thought under Process Supervision," *arXiv* preprint 2506.01926 (June 2, 2025), https://arxiv.org/abs/2506.01926.

[45] Sumeet Ramesh Motwani et al., "Secret Collusion among Generative AI Agents: Multi-Agent Deception via Steganography," *arXiv* preprint 2402.07510 v4 (April 14, 2025), https://arxiv.org/abs/2402.07510.

[46] Yohan Mathew et al., "Hidden in Plain Text: Emergence & Mitigation of Steganographic Collusion in LLMs," *arXiv* preprint 2410.03768 (October 2, 2024), https://arxiv.org/abs/2410.03768.

[47] Pratyush Maini et al., "Safety Pretraining: Toward the Next Generation of Safe AI," *arXiv* preprint arXiv:2504.16980 (23 April 2025), https://arxiv.org/abs/2504.16980.

[48] David "davidad" Dalrymple et al., "Towards Guaranteed Safe AI: A Framework for Ensuring Robust and Reliable AI Systems," *arXiv* preprint arXiv:2405.06624 (10 May 2024, rev. 8 July 2024), https://arxiv.org/abs/2405.06624.

[49] OpenAI, *Model Spec* (public draft, April 11 2025), https://model-spec.openai.com/2025-04-11.html.

[50] Fazl Barez *et al.*, "Open Problems in Machine Unlearning for AI Safety," *arXiv* preprint, arXiv:2501.04952 (January 9 2025), https://arxiv.org/abs/2501.04952.

[51] Andy Zou et al., "Representation Engineering: A Top-Down Approach to AI Transparency," *arXiv* preprint arXiv:2310.01405 (2 October 2023, rev. 3 March 2025), https://arxiv.org/abs/2310.01405.

[52] Xiangyu Qi et al., "Safety Alignment Should Be Made More than Just a Few Tokens Deep," *arXiv* preprint, arXiv:2406.05946 (submitted June 10 2024), https://arxiv.org/abs/2406.05946.

[53] Xin Chen, Yarden As, and Andreas Krause, "Learning Safety Constraints for Large Language Models," *arXiv* preprint arXiv:2505.24445 (30 May 2025), https://arxiv.org/abs/2505.24445.

[54] Meinke et al., "Frontier Models Are Capable of In-Context Scheming."

[55] Joe Benton et al., "Sabotage Evaluations for Frontier Models," *arXiv* preprint, arXiv:2410.21514 (submitted October 28 2024), https://arxiv.org/abs/2410.21514.

[56] Balesni et al., "Towards Evaluations-Based Safety Cases for AI Scheming."

[57] Tamera Lanham et al., "Measuring Faithfulness in Chain-of-Thought Reasoning," *arXiv* preprint, arXiv:2307.13702 (submitted July 17 2023), https://arxiv.org/abs/2307.13702.

[58] Yanda Chen et al., "Reasoning Models Don't Always Say What They Think," Anthropic Alignment Science paper (April 3 2025), https://assets.anthropic.com/m/71876fabef0f0ed4/original/reasoning_models_paper.pdf.

[59] Benjamin Arnav et al., "CoT Red-Handed: Stress Testing Chain-of-Thought Monitoring," *arXiv* preprint, arXiv:2505.23575 (submitted May 31 2025), https://arxiv.org/abs/2505.23575.

[60] Anthropic Alignment Team, "Simple Probes Can Catch Sleeper Agents," Anthropic Alignment Note, April 23 2024, https://www.anthropic.com/research/probes-catch-sleeper-agents.

[61] Samuel Marks et al., "Auditing Language Models for Hidden Objectives," *arXiv* preprint, arXiv:2503.10965 (revised March 28 2025), https://arxiv.org/abs/2503.10965.

[62] Yoshua Bengio et al., *International AI Safety Report: The International Scientific Report on the Safety of Advanced AI*, (London: UK Department for Science, Innovation and Technology, January 2025), 206, https://assets.publishing.service.gov.uk/media/679a0c48a77d250007d313ee/International_AI_Safety_Report_2025_accessible_f.pdf.

[63] Shah et al., "An Approach to Technical AGI Safety and Security."

[64] Ryan Greenblatt et al., "AI Control: Improving Safety Despite Intentional Subversion," *arXiv* preprint, arXiv:2312.06942 (submitted December 11 2023), https://arxiv.org/abs/2312.06942.

[65] Ibid.

[66] Yoshua Bengio et al., "Superintelligent Agents Pose Catastrophic Risks: Can Scientist AI Offer a Safer Path?," *arXiv* preprint, arXiv:2502.15657 (submitted February 29 2025), https://arxiv.org/abs/2502.15657.