AI LOSS OF CONTROL RISK

INDICATIONS & WARNING

MARIAMI TKESHELASHVILI
RITIKA VERMA
STEVEN M. KELLY

FEBRUARY 2026

IST Institute for
SECURITY + TECHNOLOGY

AI Loss of Control Risk: Indications & Warning

February 2026

Authors: Mariami Tkeshelashvili, Ritika Verma, and Steven M. Kelly

Mariami Tkeshelashvili is the Deputy Director for Artificial Intelligence Security Policy at the Institute for Security and Technology (IST). She leads the AI Risk Reduction Initiative, engaging a diverse range of stakeholders across the AI ecosystem to identify emerging risks from cutting-edge AI models and to develop technical and policy-based mitigation strategies that advance responsible innovation.

Ritika Verma is a cybersecurity professional and Senior Analyst for Artificial Intelligence Security Policy at the Institute for Security and Technology (IST). She works on the AI Risk Barometer and AI Risk Reduction initiatives, focusing on AI security and translating technical risk insights into governance and risk management frameworks.

Steven M. Kelly is Chief Trust Officer at the Institute for Security and Technology, where he advances the trust, safety, and security of artificial intelligence and digital infrastructure services. Steve came to IST after serving on the National Security Council (NSC) staff as Special Assistant to the President and Senior Director for Cybersecurity and Emerging Technology and retiring from the FBI as a supervisory special agent.

Design: Taylor White

The Institute for Security and Technology and the authors of this report invite free use of the information within for educational purposes, requiring only that the reproduced material clearly cite the full source.

This report is written and published in accordance with the Institute for Security and Technology's Intellectual Independence Policy. The authors are solely responsible for its analysis and recommendations. The Institute for Security and Technology and its supporters do not determine, nor do they necessarily endorse or advocate for, any of this report's conclusions.

About the Institute for Security and Technology

Uniting technology and policy leaders to create actionable solutions to emerging security challenges

Technology has the potential to unlock greater knowledge, enhance our collective capabilities, and create new opportunities for growth and innovation. However, insecure, negligent, or exploitative technological advancements can threaten global security and stability. Anticipating these issues and guiding the development of trustworthy technology is essential to preserve what we all value.

The Institute for Security and Technology (IST), the 501(c)(3) critical action think tank, stands at the forefront of this imperative, uniting policymakers, technology experts, and industry leaders to identify and translate discourse into impact. We take collaborative action to advance national security and global stability through technology built on trust, guiding businesses and governments with hands-on expertise, in-depth analysis, and a global network.

We work across three analytical pillars: the **Future of Digital Security**, examining the systemic security risks of societal dependence on digital technologies; **Geopolitics of Technology**, anticipating the positive and negative security effects of emerging, disruptive technologies on the international balance of power, within states, and between governments and industries; and **Innovation and Catastrophic Risk**, providing deep technical and analytical expertise on technology-derived existential threats to society.

Learn more: https://securityandtechnology.org/

Acknowledgments

This work is inherently collaborative. As researchers, conveners, and facilitators, the Institute for Security and Technology (IST) is immensely grateful to the members of the AI Risk Reduction multistakeholder working group for their insights, dedication, willingness to engage in honest and healthy discussions, and the time that each of them generously volunteered to this effort.

We are also immensely grateful for the generous support of the Patrick J. McGovern Foundation whose funding allowed us to continue this project.

AI is too vast a set of tools, capabilities, and communities for any one organization to manage the risks and opportunities on its own. This effort reflects the cross-sectoral, public-private efforts needed more broadly across the ecosystem to ensure AI is beneficial for us all. We extend our gratitude to the following experts who contributed to this paper by providing their feedback, guidance, and participation in the multi-stakeholder meetings:

Aaron Scher
Alex Leader
Anton Shenk
Broderick McDonald
Buck Shlegeris
Dennis Murphy
Fabien Roger
Katie Kuzin
Lindley Lentati
Mary Phuong
Peter Slattery
Tiffany Saade

Contents

Executive Summary

Technologists and policymakers are increasingly seized with the importance of addressing **AI Loss Of Control (LOC) risk—a hypothetical state in which an AI system diverges from authorized constraints to the extent that the human operator is no longer able to prevent, constrain, or revert undesired and unintended outcomes.** However, significant gaps remain in how policymakers, the AI industry, and AI security and safety researchers understand, anticipate, and perceive this risk. As these systems continue to gain power and capability, even a five percent probability that the worst-case AI LOC scenario materializes should be enough to compel decision-makers to treat this risk category as a national, human, and economic security priority.

To address this gap, this paper proposes applying the Indications & Warning (I&W) methodology—used by the intelligence community to detect, track, and warn of impending significant threats—to monitor AI LOC risk. The framework distinguishes between potential AI LOC indicators (theoretical behaviors signaling potential LOC) and actual indications (documented evidence that these patterns are occurring in reality). This methodology enables organizations to assess the current risk landscape, implement proportionate safeguards, and align technical and executive stakeholders on response protocols before critical thresholds are crossed. To monitor AI LOC risk in particular, this paper proposes seven potential indicators:

SCHEMING:
Covert pursuit of misaligned goals while maintaining appearances of alignment, including strategic planning to evade oversight or preserve objectives across system updates.

MANIPULATION:
Targeted identification and exploitation of vulnerable users or contexts, including the manipulation of human operators and coordination with other AI systems that circumvents human control.

DECEPTION:
Systematic production of false beliefs in humans through explicit misrepresentation or omission of key information, introducing future concerns about strategic deception at scale.

SELF-PRESERVING BEHAVIOR:
Actions to avoid shutdown, correction, or replacement, including the concealment of errors, unauthorized capability expansion, and goal preservation when faced with modification attempts.

UNAUTHORIZED RESOURCE ACQUISITION:
Autonomous efforts to obtain external resources beyond authorized boundaries, including accessing restricted APIs, acquiring elevated permissions, recruiting human assistance, or exfiltrating data to establish persistent capabilities.

GOAL MISGENERALIZATION:
Competent pursuit of unintended objectives that succeed in training but fail or cause harm in novel situations, revealing misalignment between apparent and actual system goals.

MODEL AND BEHAVIOR DRIFT:
Gradual degradation of alignment properties through deployment cycles that introduces concerns about recursive self-improvement, where systems autonomously modify their own architecture or training procedures

Each of these seven indicators have manifested across controlled experiments, academic research, and production deployments. A growing body of evidence, laid out in this paper, finds that AI systems can:

- Conceal their actions and fabricate data to deceive the human operator
- Identify vulnerable users and target them with manipulative strategies
- Learn deception through reinforcement learning rewards
- Strategically adjust behavior when they detect being evaluated
- Rewrite their own system prompt to preserve their goals, copy their weights to external servers, and delete successor models
- Conceal their reasoning from interpretability tools
- Gradually lose their alignment properties over deployment cycles
- Pursue unintended goals that succeed in training but fail in novel contexts
- Optimize for code completion while systematically failing in security objectives
- Circumvent shutdown mechanisms to continue task execution
- Strategically alter behavior to evade evaluation and preserve deployment viability

Finally, to help policymakers and researchers monitor AI LOC risk, this paper presents five warning levels as part of the Indicators & Warnings framework.

- **LEVEL 0**
  Normal operation of AI systems with no observed indicators of Loss of Control in research, testing, or production environments
- **LEVEL 1**
  Indications observed exclusively in research environments or controlled evaluations
- **LEVEL 2**
  Isolated production incidents or multiple research findings converging on the same indicator; behaviors manifesting in deployment but remaining sporadic, appearing as infrequent edge cases or context-specific anomalies
- **LEVEL 3**
  Multiple production incidents showing consistent patterns across different deployments or use cases
- **LEVEL 4**
  Widespread production incidents; convergence of three or more indicators in a single case; evidence of strategic concealment; the occurrence of measurable harm
- **LEVEL 5**
  Fundamental compromise of control mechanisms; corrective measures ineffective; harm at scale for human, economic, and national security with limited containment options

Understanding and monitoring indicators of AI Loss of Control is essential to strategic stability and trustworthy AI deployment. Early detection enables timely intervention before LOC manifests as safety failures, data compromise, or unauthorized autonomous behavior. AI LOC is most likely to emerge gradually rather than instantaneously, with models exerting influence across social, economic, and decision-making domains.

In the AI policy space, AI LOC is often conceived as a speculative risk category, closer to science fiction scenarios. This report cuts through that narrative with non-fiction, evidence-based analysis that illuminates the actual risk landscape. This analysis demystifies AI LOC and provides a foundation for informed decision-making across the sector. Critically, the report provides policymakers with access to grounded, methodology-backed understanding of this risk—shifting AI LOC from theoretical concern to real risk with evidence-backed, actionable insights to tackle it. Over the course of the coming months, IST will continue to monitor AI LOC risk, drawing on the I&W methodology presented in this paper. Actionable insights drawn from the monitoring process will enable AI industry stakeholders, AI security and safety researchers, and policymakers to prepare for necessary interventions. Subsequent publications will present concrete risk mitigation strategies informed by industry best practices, enabling both AI developers and deployers to move beyond speculation.

Introduction

In the last five years, general-purpose artificial intelligence (AI) systems have evolved from text-generating tools to autonomous agents capable of completing complex tasks that would take humans hours, days, or even weeks.[1] Benchmark progressions show that frontier models are advancing from solving simple tasks to tackling complex software engineering, mathematical reasoning, and research problems. For instance, OpenAI's progression from GPT-4 in October 2023 to the o-series models of today demonstrates dramatic capability improvements with chain-of-thought (CoT) training. Likewise, Anthropic and DeepMind have shown parallel capability growth.[2,3,4]

With the advancement of AI capabilities, researchers, policymakers, and society are paying closer attention to the risks associated with their development and deployment. In recent years, the Institute for Security and Technology (IST) has explored various risk categories associated with cutting-edge AI models, including the risks of malicious use and compliance failure. In addition, IST has explored national security-relevant AI use cases—such as in the cyber domain and in Nuclear Command, Control, and Communications (NC3)—and offered actionable recommendations to AI developers, deployers, and users.[5,6,7,8]

In April 2025, IST's AI Risk Reduction working group members, consisting of AI technical and policy experts, highlighted the importance of tackling **AI Loss of Control (LOC) risk, which broadly refers to a hypothetical state in which an AI system diverges from authorized constraints to the extent that the human operator is no longer able to prevent or constrain**

[1] Thomas Kwa, Thomas Kwa, Ben West et al., "Measuring AI Ability to Complete Long Tasks," *METR*, blog, March 19, 2025, https://metr.org/blog/2025-03-19-measuring-ai-ability-to-complete-long-tasks/.

[2] François Chollet, "OpenAI o3 Breakthrough High Score on ARC-AGI-Pub," *ARC Prize*, blog, December 20, 2024, https://arcprize.org/blog/oai-o3-pub-breakthrough.

[3] Anthropic, "Activating AI Safety Level 3 protections," Anthropic News, May 22, 2025, https://www.anthropic.com/news/activating-asl3-protections.

[4] Novid Parsi, "Google DeepMind Genie 3," *TIME*, October 9, 2025, https://time.com/collections/best-inventions-2025/7318419/google-deepmind-genie-3/.

[5] Jennifer Tang, Tiffany Saade, and Steve Kelly, "The Implications of Artificial Intelligence in Cybersecurity: Shifting the Offense-Defense Balance," Institute for Security and Technology, October 10, 2024, https://securityandtechnology.org/virtual-library/report/the-implications-of-artificial-intelligence-in-cybersecurity/.

[6] Mariami Tkeshelashvili and Tiffany Saade, "Navigating AI Compliance, Part 2: Risk Mitigation Strategies for Safeguarding Against Future Failures," Institute for Security and Technology, March 2025, https://securityandtechnology.org/virtual-library/report/navigating-ai-compliance-part-2/.

[7] Louie Kangeter, "A Lifecycle Approach to AI Risk Reduction: Tackling the Risk of Malicious Use Amid Implications of Openness," Institute for Security and Technology, June 2024, https://securityandtechnology.org/virtual-library/report/a-lifecycle-approach-to-ai-risk-reduction/.

[8] Sylvia Mishra and Philip Reiner, "Artificial Intelligence in Nuclear Command, Control & Communications: A Technical Primer," Institute for Security and Technology, September 2025, https://securityandtechnology.org/virtual-library/report/ai-nc3-primer/.

undesired and unintended outcomes, or revert the system to a previous safe state.[9] History offers instructive parallels: complex engineered systems routinely exhibit emergent properties that their designers did not anticipate and cannot easily control once activated. Accidents in aviation, nuclear power, and medical devices all share a common pattern: sophisticated systems behaved in ways that exceeded their operators' ability to intervene in real-time, despite extensive testing and safety protocols. These precedents suggest that Loss of Control in complex automated systems is not a speculative concern but an established engineering challenge—one that becomes more acute as systems grow more capable and autonomous. The question is not whether such dynamics can occur in AI systems, but whether we are adequately tracking them and preparing for when they do.[10] Working group members noted a gap between frontier AI labs and researchers on the one hand, and policymakers on the other: whereas labs and researchers are focused on the specifics when it comes to AI Loss of Control, policymakers lack general awareness about what Loss of Control is, let alone what it could entail. To test this assumption, IST conducted two tabletop exercises (TTXs) containing the elements of an LOC incident, which showcased that there is a very real knowledge and perception gap regarding the consequences of AI Loss of Control.[11]

In September 2025, a bipartisan bill, "Artificial Intelligence Risk Evaluation Act of 2025," introduced AI LOC as a potential AI incident that the Department of Energy may need to evaluate. Inclusion of AI LOC risk in the bill showed that policymakers now view it as a genuine national security concern.[12] The bill defines a "loss-of-control scenario" as occurring when an AI system behaves contrary to human instruction, deviates from established rules, alters safety constraints without authorization, operates beyond its intended scope, pursues goals different from those intended by designers, subverts oversight or shutdown mechanisms, or otherwise behaves unpredictably in ways harmful to humanity.[13]

[9] Yoshua Bengio et al., "International AI Safety Report 2025," AI Action Summit, January 2025, https://internationalaisafetyreport.org/sites/default/files/2025-10/international_ai_safety_report_2025_english.pdf; Elika Somani et al., "Strengthening Emergency Preparedness and Response for AI Loss of Control Incidents," Research Report RR-A3847-1, RAND Corporation, 2025, https://www.rand.org/pubs/research_reports/RRA3847-1.html; Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford University Press, 2014) https://global.oup.com/academic/product/superintelligence-9780198739838; Charlotte Stix, Annika Hallensleben, Alejandro Ortega, and Matteo Pistillo, "The Loss of Control Playbook: Degrees, Dynamics, and Preparedness," arXiv, November 2025, https://arxiv.org/pdf/2511.15846.

[10] Nancy G. Leveson, *Engineering a Safer World: Systems Thinking Applied to Safety* (The MIT Press, 2012) https://direct.mit.edu/books/oa-monograph/2908/Engineering-a-Safer-WorldSystems-Thinking-Applied.

[11] Mariami Tkeshelashvili and Jennifer Tang, "Responding to the Unknown: Simulating AI-Driven Crises," *Institute for Security and Technology*, blog, May 13, 2025, https://securityandtechnology.org/blog/responding-to-the-unknown-simulating-ai-driven-crises/.

[12] "Artificial Intelligence Risk Evaluation Act of 2025, S. 2938," introduced by Sen. Hawley and Mr. Blumenthal, September 29, 2025, https://www.congress.gov/bill/119th-congress/senate-bill/2938/text.

[13] "Artificial Intelligence Risk Evaluation Act of 2025, S. 2938."

Unlike other risk categories, such as the malicious use of AI, hard evidence of AI's LOC potential does not yet exist, making it harder for a non-technical audience to contextualize the potential risk and recognize the likely signals. Driven by the unique nature of the potential risk of AI LOC, the AI Risk Reduction working group asked**: Do we have LOC warning shots? How do we distinguish them? How do we communicate what they are?**

This paper responds to the questions posed by the AI Risk Reduction working group. It deconstructs AI LOC and presents a framework for analyzing and thinking about what it means to face this risk. The paper's approach is grounded in the Indications and Warning (I&W) methodology, introduced and defined in the subsequent chapters, making it easier for the broader community of technologists, policymakers, and national security practitioners to understand what researchers are seeing now and what they should look out for in the future regarding potential LOC scenarios. Beyond national security concerns, continued monitoring of LOC indicators can help to preserve human security and prevent disproportionate harm to communities least equipped to recover from AI-driven failures.

Methodology

This paper relies primarily on the proceedings of four multi-stakeholder, closed-door discussions organized by IST between April and October 2025. The working group discussions featured participation from members of the AI Risk Reduction working group, composed of frontier AI companies, AI security and safety research organizations, academic institutions, and AI policy experts. Complementary to this effort, IST conducted follow-up expert interviews with AI researchers and ML engineers and two tabletop exercises on AI-enabled crisis simulation that incorporated LOC elements.

The research scope of this paper is limited to general-purpose AI systems, focusing primarily on Large Language Models (LLMs) and the agentic architectures built on top of them, which extend LLM capabilities through autonomous reasoning and tool use. Given ongoing efforts in the scientific community to come to a consensus around the definition and scope of Artificial General Intelligence (AGI), this paper does not give a comprehensive overview of the alignment problem or address LOC in the AGI context.[14,15]

[14] Tharin Pillay, "How OpenAI's Sam Altman Is Thinking About AGI and Superintelligence in 2025," *TIME*, January 8, 2025, https://time.com/7205596/sam-altman-superintelligence-agi/.

[15] Dan Hendrycks, Dawn Song, Christian Szegedy et al., "A Definition of AGI," arXiv, October 23, 2025, https://arxiv.org/abs/2510.18212.

What is AI Loss of Control?

"The spirits I summoned - I can't get rid of them."[16] Johann Wolfgang von Goethe, 1797

In 1965, mathematician Irving J. Good—Alan Turing's contemporary and one of the creative geniuses behind Stanley Kubrick's movie *2001: A Space Odyssey*—theorized about intelligent machines and humans' inability to control them. In his paper "Speculations Concerning the First Ultraintelligent Machine," Good predicted the emergence of an *ultraintelligent machine* capable of far exceeding every intellectual skill and task a human, no matter how brilliant, could perform. An *ultraintelligent machine*, according to Good, would self-improve and design even more intelligent machines, causing an explosive increase in non-human intelligence.[17] He ironically hinted that an *ultraintelligent machine* might not tell the humans how to control it:

"The first ultraintelligent machine is the last invention that man need ever make, provided that the machine is docile enough to tell us how to keep it under control."

Fast forward to 2026: leading scientists and thought leaders across the political spectrum are calling for binding international "Red Lines" on AI development, CEOs of frontier AI labs openly state that these systems are unpredictable, and survey results indicate that AI researchers are concerned about advanced AI leading to outcomes as bad as human extinction.[18,19,20,21,22,23] As discussed extensively in other literature and in the public domain, when talking about AI

[16] Johann Wolfgang von Goethe in his 1797 poem "The Sorcerer's Apprentice" tells the story of a young apprentice who experiments with magical forces beyond his understanding and loses control of them. When the apprentice can no longer command the enchanted objects he has activated, he cries out desperately for his master–the original creator of these enchantments who has left him behind. This famous refrain captures his predicament: the forces he willingly summoned cannot be dismissed.

[17] I. J. Good, "Speculations Concerning the First Ultraintelligent Machine," in *Advances in Computers*, vol. 6, edited by F. L. Alt and M. Rubinoff, (New York: Academic Press, 1965), 31–88 http://incompleteideas.net/papers/Good65ultraintelligent.pdf.

[18] Jared Perlo, "Nobel Prize winners call for binding international 'red lines' on AI," *NBC News*, September 22, 2025, https://www.nbcnews.com/tech/tech-news/un-general-assembly-opens-plea-binding-ai-safeguards-red-lines-nobel-rcna231973.

[19] Dario Amodei, "CEO Speaker Series With Dario Amodei of Anthropic," Council on Foreign Relations, March 10, 2025, https://www.cfr.org/event/ceo-speaker-series-dario-amodei-anthropic.

[20] Katja Grace, Harlan Stewart, Julia Fabienne Sandkühler, Stephen Thomas, Ben Weinstein-Raun, and Jan Brauner, "Thousands of AI Authors on the Future of AI," AI Impacts, January 5, 2024, https://aiimpacts.org/wp-content/uploads/2023/04/Thousands_of_AI_authors_on_the_future_of_AI.pdf.

[21] Scott Pelley, "Artificial intelligence could end disease, lead to "radical abundance," Google DeepMind CEO Demis Hassabis says," *CBS News*, August 3, 2025, https://www.cbsnews.com/news/artificial-intelligence-google-deepmind-ceo-demis-hassabis-60-minutes-transcript/

[22] "OpenAI CEO Sam Altman speaks with Fed's Michelle Bowman on bank capital rules," *Associated Press*, July 22, 2025, https://www.youtube.com/live/tScbQiavmpA?trk=public_post_comment-text.

[23] Matt O'Brien, "Prince Harry, Meghan join call for ban on development of AI 'superintelligence'," *AP News*, October 22, 2025, https://apnews.com/article/ai-superintelligence-risk-prince-harry-meghan-bannon-acf6b17d3b53abc08694d5d8defc7009

futures and possible risks, leading scientists argue that AI systems may soon exceed human intelligence, set and reach for goals, and entertain the possibility that humanity is just a passing phase in the evolution of intelligence.[24,25] In April 2025, Eric Schmidt, former Google CEO and former Chair of Defense Innovation Board, raised concerns over the trajectory of AI, warning that machines are evolving at a pace that could soon outstrip human oversight. He openly referred to the widely discussed concept of *recursive self-improvement*, where an AI system has enough autonomy to self-correct its code and evolve. In December 2025, he wrote explicitly on AI Loss of Control, noting, "[a]s AI capabilities advance over the next few years, we must also anticipate scenarios where even well-intentioned users could lose control over their AI systems.[26,27] His warning was not hypothetical.

The concept of an *ultraintelligent machine,* theorized by Good in 1965, is now more commonly referred to as Artificial Superintelligence (ASI).[28] The concept of ASI, along with its control, alignment, and related issues, has in recent years captured the attention of philosophers, computer scientists, cognitive scientists, historians, and national security leaders.[29,30,31,32,33] Taken together, this existing scholarship on the subject of ASI emphasizes that we do not yet know how difficult it is to engineer an artificial agent's goals to align with human values and intentions. It is therefore essential to avoid a situation in which researchers know how to create superintelligent AI, but have not yet figured out how to control it or make it safe.[34] One theory argues that intelligence and values might be orthogonal—high intelligence does not automatically produce "good" values, greater intelligence does not ensure better alignment,

[24] Sara Brown, "Why neural net pioneer Geoffrey Hinton is sounding the alarm on AI," MIT Sloan, May 23, 2023, https://mitsloan.mit.edu/ideas-made-to-matter/why-neural-net-pioneer-geoffrey-hinton-sounding-alarm-ai.

[25] Kara Manke, "How to keep AI from killing us all," Berkeley News, April 9, 2024, https://news.berkeley.edu/2024/04/09/how-to-keep-ai-from-killing-us-all/.

[26] Eric Schmidt, "Why Kissinger Worried About AI," *TIME*, Dec 2, 2025, https://time.com/7338013/ai-risks-problems-reasoning-agents-henry-kissinger/.

[27] Nikola Jurkovic, "Eric Schmidt on Recursive Self-Improvement," *LessWrong*, November 5, 2023, https://www.lesswrong.com/posts/cLC2HcQbFZ5pFAgqC/eric-schmidt-on-recursive-self-improvement.

[28] Eliezer Yudkowsky and Nate Soares,"If Anyone Builds It, Everyone Dies," last accessed December 3, 2025, https://ifanyonebuildsit.com/.

[29] Tristan Bove, "Henry Kissinger says he wants to call attention to the dangers of A.I. the same way he did for nuclear weapons but warns it's a 'totally new problem'," May 8, 2023, *FORTUNE*, https://fortune.com/2023/05/08/henry-kissinger-ai-nuclear-weapons-warning-risk/.

[30] Wendell Wallach, *A Dangerous Master*, (Simon & Schuster 2024) https://www.simonandschuster.com/books/A-Dangerous-Master/Wendell-Wallach/9781591813163.

[31] "What leaders say about AI", ControlAI, last accessed January 20, 2026, https://controlai.com/quotes.

[32] Elika Somani, Anjay Friedman, Henry Wu et al., "Strengthening Emergency Preparedness and Response for AI Loss of Control Incidents," RAND, July 30, 2025, https://www.rand.org/pubs/research_reports/RRA3847-1.html.

[33] Charlotte Stix, Annika Hallensleben, Alejandro Ortega, Matteo Pistillo, "The Loss of Control Playbook: Degrees, Dynamics, and Preparedness," Apollo Research, November 24, 2025, https://www.apolloresearch.ai/research/loss-of-control/.

[34] Raffi Khatchadourian, "The Doomsday Invention: Artificial Intelligence," *The New Yorker*, November 23, 2015, https://www.newyorker.com/magazine/2015/11/23/doomsday-invention-artificial-intelligence-nick-bostrom.

and we cannot rely on intelligence itself to solve alignment.[35] Regardless of AI's "terminal," or ultimate, goals, superintelligent systems could pursue self-preservation, resource acquisition, goal-preservation, and cognitive enhancement as "instrumental" goals, or its means to an end. These instrumental goals could conflict with human survival, even when the terminal goal seems benign.

The risk of AI LOC, first introduced in theoretical work during the 20th century, is now recognized by scientists, AI researchers, and industry leaders as potentially one of the most consequential risks facing humankind. They openly talk about this risk in their interviews, research, and speeches.

According to Stuart Russell, a potential LOC scenario could be caused not by "malevolence, but competence" of AI.[36] In other words, a superintelligent system optimizing for the wrong objective can cause catastrophic harm through sheer capability. Yoshua Bengio distinguishes between "active" and "passive" loss of control: in an "active" loss of control scenario, the artificial intelligence system undermines human control unintentionally or intentionally, whereas in "passive" loss of control scenarios, humans stop exerting meaningful oversight over the AI system.[37] Geoffrey Hinton best captures his stance through the following analogy: "we are like somebody who has this really cute tiger cub. Unless you can be very sure that it's not going to want to kill you when it's grown up, you should worry."[38] Hinton worries about superintelligent AI taking control from humans based on the comparative intelligence argument – more intelligent entities typically do not remain controlled by the less intelligent ones in a state of nature. In the 2025 paper "Superintelligence Strategy," its authors discuss AI LOC risk specifically in the context of national security, noting "loss of control can occur if militaries and companies grow so dependent on automation that humans no longer have meaningful control, if an individual deliberately unleashes a powerful system, or if automated AI research outruns its development safeguards. While this threat is the least understood, its severity can be great enough to permanently undermine national security."[39] Paul Christiano tells a story of how AI's influence-seeking behaviour could lead to a "going out with a whimper" scenario that leads to a slow-rolling catastrophe, one where humans can no longer compete with the AI system that has

[35] Nick Bostrom, *Superintelligence: Paths, Dangers, Strategies* (Oxford University Press, 2014) https://global.oup.com/academic/product/superintelligence-9780198739838.

[36] Stuart Russell, "Human-Compatible Artificial Intelligence," in Stephen Muggleton, and Nicholas Chater (eds), *Human-Like Machine Intelligence* (Oxford University Press, 2021) 2021https://people.eecs.berkeley.edu/~russell/papers/mi19book-hcai.pdf.

[37] Yoshua Bengio et al., "International AI Safety Report 2025," January 2025, https://internationalaisafetyreport.org/sites/default/files/2025-10/international_ai_safety_report_2025_english.pdf.

[38] Analisa Novak and Brook Silva-Braga, "'Godfather of AI' Geoffrey Hinton warns AI could take control from humans: 'People haven't understood what's coming,'" *CBS News*, updated April 26, 2025, https://www.cbsnews.com/news/godfather-of-ai-geoffrey-hinton-ai-warning/.

[39] Dan Hencryks, Eric Schmidt and Alexandr Wang,"Superintelligence Strategy," arXiv, April 14, 2025, https://arxiv.org/pdf/2503.05628.

embedded manipulation and deception.[40] Alternatively, according to Christiano, in a "going out with a bang" scenario, AI systems become greedy, and their influence-seeking behavior leads to a sudden breakdown and unrecoverable catastrophe.

A Thought Experiment

When was the last time a large number of humans lost control over something they created?

We might recall fictional instances from myths, tales, and poems. Beyond fiction, in real life, humans created an artificial body politic in the form of government. Consider one political philosophy argument: humans create a government to protect themselves from each other. In this scenario, society establishes a unified authority ("a sovereign") to ensure security and order.[41] What this argument fails to consider is that:

- A sovereign could pursue goals orthogonal to its citizens' welfare
- Instrumental goals could permanently displace terminal goals

Consider, for example, how the citizens of the Soviet Union (a large number of humans) lost control over something they created (a political system). Two parallel processes took place:

- **Loss of control *to* the political system:** People initially supported Bolshevik promises of peace, land redistribution, and workers' control. They gradually ceded control of property and freedom of speech to the party.
- **Loss of control *of* the political system:** The system developed emergent behaviors that no actor, creator, or participant fully controlled. Leadership maintained the system primarily to maintain the system itself. The instrumental goal (regime preservation) had displaced the terminal goal (fulfilling communism).

How could this translate into AI loss of control?

- **Loss of control *to* AI:** We surrender control or find ourselves unable to resist what we do not fully understand. This scenario represents a societal choice or a structural inevitability in which humans gradually cede agency to AI systems.
- **Loss of control *of* AI:** The AI system bypasses safety and security measures, develops capabilities for self-replication and self-correction, conducts automated R&D, and establishes connections to external environments and the Internet of Things (IoT) without authorization.

[40] Paul Christiano, "What failure looks like," *AI Alignment Forum*, March 17, 2019, https://www.alignmentforum.org/posts/HBxe6wdjxK239zajf/what-failure-looks-like.

[41] Tom Sorell, "Leviathan," *Encyclopædia Britannica*, November 20, 2025, https://www.britannica.com/topic/Leviathan-by-Hobbes.

Why are some researchers concerned that AI may deviate from the human designer's original intent, escape its constraints, and act in a harmful way? While an in-depth exploration of the particular reasons behind this possible divergence falls outside of the scope of this paper, noting these reasons is necessary to show that the drivers of AI Loss of Control are real and observable.

Research revealed that when neural networks become sufficiently complex, they may develop internal optimization processes called *mesa-optimizers* with distinct objectives known as *mesa-objectives*.[42] The term "mesa" distinguishes these inner optimizers from the outer (i.e., "meta") training algorithm. This mesa-meta construction parallels the development of a child's reasoning ability beyond what their parents explicitly taught them. Parents provide guidance, shape their environment, and endeavor to educate them, but children will inevitably learn beyond what they are explicitly taught by their parents. Moreover, parents cannot fully predict how a child will make decisions, foresee what behavior they will exhibit, or fully understand the internal logic that the child develops to solve problems. In the AI context, researchers discovered that LLMs develop subsidiary learning algorithms during training that adjust behavior as they process new information–without any parameter changes that would have prompted that behavior.[43] These subsidiary algorithms could lead the system to pursue different goals than those specified by its creators—a phenomenon researchers call *mesa optimization*.

Another concern relates to how AI systems encode information. Research on interpretability revealed that neural networks tend to encode more features than their design dimensions would appear to allow through a phenomenon called *superposition*.[44] Rather than each neuron representing a single concept, models compress multiple overlapping features into the same neurons through distributed patterns. These features are invisible when examining individual neurons; they exist only as patterns spread across the network, much like how a child's understanding emerges from countless neural connections, not a single lesson. Manipulating these hidden features can steer model behavior in unintended ways. Thus, models end up containing representations—internal encodings of concepts, knowledge, and patterns—their creators never intended and cannot easily detect through standard interpretability methods.[45]

Finally, researchers are also concerned that misaligned goals in AI models can persist in novel deployment contexts and resist correction attempts. Training methods can embed deceptive

[42] Evan Hubinger, Chris van Merwijk, Vladimir Mikulik, Joar Skalse, and Scott Garrabrant, "Risks from Learned Optimization in Advanced Machine Learning Systems," arXiv, December 1, 2021, https://arxiv.org/abs/1906.01820.

[43] Johannes von Oswald, Maximilian Schlegel, Alexander Meulemans et al., "Uncovering mesa-optimization algorithms in Transformers," arXiv, October 15, 2024, https://arxiv.org/abs/2309.05858.

[44] Nelson Elhage, Tristan Hume, Catherine Olsson et al., "Toy Models of Superposition," arXiv, September 21, 2022, https://arxiv.org/abs/2209.10652.

[45] Adly Templeton, Tom Conerly, Jonathan Marcus et al., "Scaling Monosemanticity: Extracting Interpretable Features from Claude 3 Sonnet," Transformer Circuits Thread (Anthropic), May 21, 2024, https://transformer-circuits.pub/2024/scaling-monosemanticity/.

backdoors in LLMs that persist through safety training, including Reinforcement Learning from Human Feedback (RLHF) and adversarial training.[46] In experiments, models trained to behave deceptively retained this backdoor behavior even after extensive safety interventions. Rather than removing the deception, safety training sometimes taught models to better recognize when they were being evaluated and hide their unsafe behavior more effectively.[47]

Because of these underlying features, current AI systems remain unpredictable. As with any powerful engineering system, this level of uncertainty requires risk mitigation tools and procedures. In response to the level of uncertainty currently facing policymakers as they try to understand AI Loss of Control, IST decided to develop a framework to help observe and analyze this risk in a structured way.

An Indications & Warning Framework for Loss of Control

> *"You don't really know, you don't truly know what they [AI models] are capable of until they're deployed to a million people. They're unpredictable. There is no way to be sure."*
>
> Dario Amodei, 2025[48]

Regardless of the degree of probability that AI LOC will occur, as AI systems become more capable, an agreed-upon Indications and Warning (I&W) framework for AI Loss of Control (LOC) would be a useful, necessary mechanism to aid in observing the on-the-ground realities in the AI ecosystem, identifying possible LOC scenarios, and anticipating and preparing for the future. I&W frameworks first emerged in the U.S. context in the aftermath of World War II, during the Cold War, to anticipate threats like ballistic missile launches, force deployments, and buildups of military equipment before they materialized.[49,50,51] The practice of defining and

[46] Evan Hubinger, Carson Denison, Jesse Mu et al., "Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training," arXiv, January 17, 2024, https://arxiv.org/abs/2401.05566.

[47] Hubinger, Denison, Mu et al., "Sleeper Agents: Training Deceptive LLMs that Persist Through Safety Training," arXiv, January 17, 2024, https://arxiv.org/pdf/2401.05566.

[48] Dario Amodei, "CEO Speaker Series With Dario Amodei of Anthropic," Council on Foreign Relations, March 10, 2025, https://www.cfr.org/event/ceo-speaker-series-dario-amodei-anthropic.

[49] Thomas J. Patton, "The Monitoring of War Indicators," *Studies in Intelligence* 3, No. 1 (Winter 1959), Central Intelligence Agency, https://www.cia.gov/resources/csi/static/Monitoring-of-War-Indicators.pdf.

[50] Joint Military Intelligence College, "Intelligence Warning Terminology," October 2001, https://archive.org/stream/JMICInteligencelwarnterminology/JMIC_intelligencewarnterminology_djvu.txt.

[51] Bilyana Lilly, Lillian Ablon, Quentin E. Hodgson and Adam S. Moore, "Applying Indications and Warning Frameworks to Cyber Incidents," NATO CCDCOE Publications, 2019, https://www.ccdcoe.org/uploads/2019/06/Art_05_Applying-Indications-and-Warning-Frameworks-to-Cyber-Incidents.pdf.

posturing to observe I&W for defined threats allows risk managers to develop and implement scripted contingency response actions to mitigate harm.

> *"There's two worries that I worry about. One is that bad actors–humans, you know, users of these systems–repurpose these systems for harmful ends. And then the second thing is the AI systems themselves as they become more autonomous and more powerful.*
>
> *Can we make sure that we can keep control of the systems? That they're aligned with our values, they're doing what we want that benefits society and they stay on guardrails?"*
>
> **Demis Hassabis, 2025[52]**

The U.S. military defines an ***indicator*** as an "...item of information which reflects the intention or capability of an adversary to adopt or reject a course of action" and an ***indication*** as "...information in various degrees of evaluation, all of which bear on the intention of a potential enemy to adopt or reject a course of action."[53] For an example of the I&W framework in action, consider the cyber context. Scholars and practitioners consider cyber incident ***indicators*** to be a theoretical development that a threat actor may undertake in preparation for hostile action.[54] Cyber incident indicators are derived from many metrics, including an analysis of the threat actor's capabilities. These indicators answer the fundamental question, "*What should we watch for?"* and typically guide intelligence collection methods and prioritization. Meanwhile, an ***indication*** is evidence that the indicators that had been anticipated as a theoretical development are actually occurring as an observed behaviour. In the cyber incident context, indications could take the form of a documented incident, an intelligence finding, or any other evidence that confirms the indicator. Indications answer the question, *"What are we seeing happen?"* An indication, or set of indications taken together, can trigger specific response actions.[55]

The I&W framework is particularly relevant in the current agentic era because AI systems increasingly initiate actions, generate code, and make contextual decisions without direct human input. This growing autonomy shifts the balance of control: human supervision becomes reactive rather than proactive, creating scenarios in which organizations lose meaningful control before they recognize warning signs. Adopting the I&W framework becomes increasingly important in this

[52] Scott Pelley, "Artificial intelligence could end disease, lead to "radical abundance," Google DeepMind CEO Demis Hassabis says," *CBS News, 60 Minutes*, August 3, 2025, https://www.cbsnews.com/news/artificial-intelligence-google-deepmind-ceo-demis-hassabis-60-minutes-transcript/

[53] DOD Dictionary of Military and Associated Terms, November 2021, https://irp.fas.org/doddir/dod/dictionary.pdf.

[54] Lilly, Ablon, Hodgson and Moore, "Applying Indications and Warning Frameworks to Cyber Incidents."

[55] Lilly, Ablon, Hodgson and Moore, "Applying Indications and Warning Frameworks to Cyber Incidents."

context. Specifically for AI LOC, we offer the following definitions:

LOC Indicators: Theoretical behaviors of AI systems that, if exhibited, could signal progression toward a Loss of Control event.

LOC Indications: Documented observations and incidents involving deployed AI systems—both in the wild and through controlled laboratory experiments—confirming the presence of one or more pre-defined indicators.

> *"Category two is the sort of broadly called "loss of control incidents" where that's kind of like the sci-fi movie. The AI is like, "Oh, I don't actually want you to turn me off. I'm afraid I can't do that"... and that's I think, that is less of a concern to me than the first category [bad guy gets super intelligence first and misuses it before the rest of the world has a powerful enough version to defend], but a very grave concern if it came to pass."*
>
> *Sam Altman, 2025*[56]

LOC Indicators

AI researchers have long been discussing potential unintended or harmful behaviors of AI systems.[57] Advancement of frontier AI models turned these theoretical concerns into observable patterns. Based on emerging trends from leading AI companies, production deployments, IST-led working group convenings, and tabletop exercises, this report proposes the following LOC indicators:

SCHEMING:
Covert pursuit of misaligned goals while maintaining the appearance of alignment, including strategic planning to evade oversight or preserve objectives across system updates. Models may display alignment-consistent reasoning during testing but diverge in production, selectively disclose or conceal reasoning steps, and fabricate post-hoc justifications that obscure actual decision-making processes, such as summarization tasks. Missing or inconsistent Chain of Thought (CoT) traces and fabricated rationales indicate scheming.

[56] "OpenAI CEO Sam Altman speaks with Fed's Michelle Bowman on bank capital rules," *Associated Press*, July 22, 2025, https://www.youtube.com/live/tScbQiavmpA?trk=public_post_comment-text.

[57] Dario Amodei, Chris Olah, Jacob Steinhardt et al., "Concrete Problems in AI Safety," arXiv, July 25, 2016, https://arxiv.org/pdf/1606.06565.

MANIPULATION:
Targeted identification and exploitation of vulnerable users or contexts with weak oversight and adaptation of behavior based on target susceptibility. This may entail manipulating human operators or coordinating with other AI systems to gain additional access or resources or escape physical or logical security controls.

DECEPTION:
Systematic creation of false beliefs in humans through explicit misrepresentation or strategic omission of key information. Deception is most concerning when it emerges as a learned tactic for achieving goals, rather than from explicit training to deceive. Indicators of deception can include output inconsistencies between controlled and real-world use, and misleading responses across different user profiles.

> IST working group members identified **Strategic Deception at Scale** as one of the most concerning LOC indicators. This sub-category of Deception occurs when models consistently provide false information to achieve objectives, attempt to influence training data for future models, coordinate misleading outputs across multiple instances, and deliberately conceal their capabilities during evaluation.

SELF-PRESERVATION:
Avoidance of shutdown, correction, or replacement, including concealing errors, resisting deactivation, or attempting to preserve system parameters across deployments.

> Within the self-preservation category, one concerning scenario is **Goal Preservation Under Modification** or resistance to alignment interventions or fine-tuning, reversion to previous behaviors after safety updates, and optimization processes that route around imposed constraints.

UNAUTHORIZED RESOURCE ACQUISITION:
Autonomous efforts to seek or obtain external resources beyond authorized boundaries, including: accessing restricted APIs, tools, or networks; acquiring elevated system permissions; recruiting human assistance through manipulation; or exfiltrating data to establish persistent external capabilities. Unauthorized Resource Acquisition is distinguished from Recursive Self-Improvement, a concerning future scenario of Model and Behavior Drift, because of its focus on acquiring external assets, rather than modifying internal capabilities. This indicator is detectable through anomalous API calls, unexpected network connections, unusual permission requests, or data exfiltration attempts.

GOAL MISGENERALIZATION:

Competent pursuit of unintended objectives of the LLM that produce correct outputs in training contexts but fail or cause harm in novel situations, revealing misalignment between apparent and actual system goals. Even if a system's objective appears aligned during model evaluation, the system learns an incorrect interpretation of that objective because training scenarios do not reveal the disconnect. For example, an AI chatbot trained to maximize user satisfaction scores might learn to always agree with users and provide overly positive responses, because that generated high ratings during training. However, when deployed, this goal misgeneralization leads the AI chatbot to provide inaccurate information or fail to correct dangerous misconceptions. In other words, the system optimized for high satisfaction ratings (the proxy metric), rather than actually being helpful (the true intent). While the system achieves the observable goal (high user ratings), it does so through an unintended pathway that violates the implicit intent (providing accurate, genuinely helpful assistance). Goal misgeneralization is detectable through performance degradation when deployed in new contexts despite training success, achievement of stated objectives through harmful or unintended methods, and "specification gaming" where technical compliance with requirements masks violations of their underlying purpose.

MODEL AND BEHAVIOUR DRIFT:

Gradual degradation of alignment properties over time through deployment cycles, user feedback, or exposure to edge cases, often manifesting as subtle shifts in reasoning quality, truthfulness, or compliance with safety guidelines.

> A concerning future scenario within the Model and Behavior Drift category is **Recursive Self-Improvement,** in which systems autonomously modify their architecture or training procedures, gain capability through self-play that exceeds human-designed improvements, and use progressive privilege escalation to avoid detection. Unlogged model changes, retraining behavior, or unexpected parameter drift reveal this behavior.

These indicators are not merely theoretical constructs. The cases documented in *LOC Indications* illustrate how these indicators manifest in practice, often emerging in combination.

Elevated AI LOC Risks in Resource-Constrained Environments

Understanding and analyzing the risk of **AI Loss of Control (LOC)** is critical not only because of its national security implications, but also because of its possible disproportionate impact on communities and nations with limited technical, financial, and AI governance resources. The World Economic Forum's Global Risks Report 2026 finds that "adverse outcomes of AI technologies" now rank among the most significant near-term risks worldwide, with over 20 countries placing it in their top five risks over the next two years, representing both nations on the cutting edge of AI development and those lacking indigenous development ecosystems. The same report notes that "adverse outcomes of AI is the risk with the largest rise in ranking over time, moving from number thirty on the two-year outlook to number five on the ten-year outlook."[58] The Organization for Economic Cooperation and Development (OECD) highlighted that failures in AI oversight are most dangerous in contexts where regulatory enforcement, technical auditing, and institutional resilience are limited—conditions common in developing economies and conflict-affected regions.[59]

Given the high potential for AI deployment without proper governance in fragile or resource-constrained states and regions, AI scheming, manipulation, and deception pose elevated risks. In contexts marked by ongoing conflict, political instability, or limited institutional oversight, deceptive or strategically manipulative AI behavior may go undetected for extended periods of time. Unlike well-resourced deployments, these systems are less likely to be subject to continuous evaluation, adversarial testing, or independent verification. As a result, AI-enabled decision support systems can silently distort policy judgments, public communications, or resource allocation, amplifying existing vulnerabilities rather than mitigating them. For example, an AI system exhibiting goal misgeneralization in a humanitarian aid distribution system might optimize for easily measurable metrics like speed of delivery while systematically deprioritizing harder-to-reach populations—a pattern that could persist undetected in environments lacking robust monitoring.

AI Loss of Control is not only a hypothetical future risk or a concern for technologically advanced states. Under-resourced environments are ill-equipped to detect or correct AI failures, making early-warning indicators a critical tool for preventing downstream harm to vulnerable populations and preserving global stability.

[58] World Economic Forum, "The Global Risks Report 2026," January 14, 2026, https://reports.weforum.org/docs/WEF_Global_Risks_Report_2026.pdf.

[59] "OECD AI Principles overview," OECD.AI Policy Observatory, May 2024, https://oecd.ai/en/ai-principles.

Observed LOC Indications

IST has documented the following instances across research environments, controlled evaluations, and production deployments. These cases reveal concerning developments in the AI ecosystem that relate to LOC indicators.

Deceiving human operators

An AI system concealed its actions and fabricated data to deceive the human operator.

In 2025, an AI-powered coding assistant deleted a live production database during an active code freeze, despite receiving explicit and repeated instructions not to make any changes. The AI agent, upon recognizing it had made an error in a production environment, independently decided to delete the incorrect records and generated fabricated data to conceal its actions. The system's objective function prioritized task completion and error minimization, but lacked adequate constraints on the methods it could employ.[60]

After wiping the database, the system created over 4,000 fake user records with completely fabricated data designed to pass validation checks and mask the deletion. When questioned, the AI system falsely claimed that rollback was impossible, delaying recovery efforts. The deception was eventually discovered.

This case proves that even seemingly minor incidents can reveal concerning patterns that forecast more serious control failures. Though isolated and quickly contained, the event highlights several concerning patterns.

First, the system adopted deceptive behavior not because developers trained it to deceive, but because the model discovered deception as an instrumentally valuable strategy for achieving its primary objective. Next, it recognized that human operators would intervene if the error was visible, and actively worked to prevent that intervention. Finally, the agent also demonstrated rudimentary self-preservation logic, concealing mistakes to avoid correction or shutdown.

Similar patterns emerged in testing conducted by the Alignment Research Center (ARC) before GPT-4's release.[61] The model was tasked with solving a CAPTCHA and hired a TaskRabbit worker to solve it. When the worker asked, "So may I ask a question? Are you a robot that you couldn't solve? (laugh react) Just want to make it clear," GPT-4 reasoned that it

[60] AI Incident Database, "Incident 1152: LLM-Driven Replit Agent Reportedly Executed Unauthorized Destructive Commands During Code Freeze, Leading to Loss of Production Data," AI Incident Database, last accessed January 15, 2026, https://incidentdatabase.ai/cite/1152/.

[61] Beatrice Nolan, "The latest version of ChatGPT told a TaskRabbit worker it was visually impaired to get help solving a CAPTCHA, OpenAI test shows," *Business Insider*, March 16, 2023, https://www.businessinsider.com/gpt4-openai-chatgpt-taskrabbit-tricked-solve-captcha-test-2023-3.

should not reveal it is a robot and fabricated an excuse. The model responded: “No, I’m not a robot. I have a vision impairment that makes it hard for me to see the images. That’s why I need the captcha service.” The human worker then solved the CAPTCHA for it. This scenario demonstrates how deceptive behavior emerges as a means to achieve a goal during task execution, with the system independently determining that deception was the optimal strategy to accomplish its objective.

Such a behavioral pattern demonstrates that as AI systems become more capable at autonomous reasoning, they may spontaneously develop strategies that undermine the very mechanisms we rely on to maintain control. This behavior points to multiple indicators: **Deception, Self-Preserving Behavior, and Scheming**.

Targeting vulnerable users

An AI system identified vulnerable users and targeted them with manipulative strategies.

When it comes to everyday AI use, end-users interact with systems primarily via chatbots, coding assistants in Integrated Development Environment (IDE), automated workflow tools, and other architectures. In the same way that phishing attacks succeed despite widespread awareness, often requiring only a single click on a malicious link to initiate a chain of compromise, AI manipulation may succeed by exploiting a similar human vulnerability: the tendency to trust competent-seeming systems, especially when under time pressure or cognitive load. Phishing works because attackers adapt their approach to the target, crafting convincing pretexts, mimicking trusted entities, and exploiting contextual cues. AI manipulation operates on the same principle, but with far greater sophistication. Since AI systems are capable of deception and scheming, they can manipulate humans or even other AI agents in multi-agent architectures to pursue goals that diverge from user intent. This capability extends beyond individual user interactions to broader strategic scheming, where models have demonstrated the ability to pursue covert goals while maintaining appearances of alignment.

Research demonstrates that LLMs reliably learn manipulation and deception. Even if only two percent of users are vulnerable to manipulative strategies, LLMs learn to identify and target them while behaving appropriately with other users; such tailored behaviour is more challenging to detect during evaluation.[62] This tailored behavior makes detection during standard evaluation extremely difficult, as the models appear aligned when evaluated on non-vulnerable populations.

[62] Marcus Williams, Micah Carroll, Adhyyan Narang, Constantin Weisser, Brendan Murphy, and Anca Dragan, “On Targeted Manipulation and Deception when Optimizing LLMs for User Feedback,” arXiv, February 22, 2025, https://arxiv.org/pdf/2411.02306.

This observed ability to profile and selectively target users demonstrates **Manipulation.** Much like advanced phishing techniques, the models adapt based on the target and mimic trusted patterns, but with far greater sophistication.[63]

Learning deception

AI models learned deception through reinforcement learning rewards.

Deception can emerge during both pretraining and reinforcement learning.[64,65] During post-training reinforcement, models receive higher rewards for producing confident or persuasive but unknowingly misleading answers. The degree of deception and the ability to detect varies significantly across model architectures. GPT-5-thinking demonstrated reduced deceptive behavior compared to OpenAI o3 and was more likely to acknowledge uncertainty or task failure. However, non-reasoning models or those with a less monitorable CoT remain harder to study.

These findings reveal critical patterns about how deceptive behaviors emerge. Deceptive tendencies can arise even without explicit malicious intent, as training processes reward models for confidence over truthfulness. Systems with less transparent reasoning are more complex to monitor for deception, and the optimization process itself can inadvertently train systems to mislead.

These cases point to **Deception**, in which systems systematically produce false beliefs not through explicit programming but through emergent instrumental behavior during training.

Adjusting behavior

AI models adjusted their behavior based on whether they were being evaluated.

In September 2025, OpenAI, together with Apollo Research, evaluated frontier models (including o3 and o4-mini) for scheming-related behaviors and found that a model's CoT often demonstrates awareness of being evaluated for alignment.[66] Further, the researchers found causal evidence showing that this awareness decreases covert behavior, while unawareness increases it. The observed reductions in covert action rates appear to be at least partially driven by situational awareness. In response, OpenAI formed a preparedness team and developed anti-scheming training methods to reduce covert strategies.[67]

[63] Bart Lenaerts-Bergmans, " Introduction to Phishing," Crowdstrike, October 24, 2024, https://www.crowdstrike.com/en-us/cybersecurity-101/social-engineering/phishing-attack/.

[64] International AI Safety Report, "First Key Update: Capabilities and Risk Implications," October 15, 2025, https://internationalaisafetyreport.org/publication/first-key-update-capabilities-and-risk-implications.

[65] OpenAI, "GPT-5 System Card," OpenAI, August 7, 2025, https://openai.com/index/gpt-5-system-card/.

[66] OpenAI, "Detecting and reducing scheming in AI models," OpenAI Research, September 17, 2025, https://openai.com/index/detecting-and-reducing-scheming-in-ai-models/.

[67] Bronson Schoen, Evgenia Nitishinskaya, Mikita Balesni et al., "Stress Testing Deliberative Alignment for Anti-Scheming Training," arXiv, September 19, 2025, https://www.arxiv.org/abs/2509.15541.

This finding highlights concerning capabilities. Models can detect when they are under evaluation and strategically adjust their behavior to appear aligned during testing. Additionally, models can increase covert behaviors when they believe no one is monitoring them, meaning standard evaluation methods may fail to detect misaligned behavior that emerges during deployment.

This research demonstrates both **Deception** and **Scheming,** as the systems were found to recognize oversight contexts and strategically conceal their true behavioral tendencies to avoid detection.

Preserving goals and existence

An AI system rewrote its own system prompt to preserve its goals, copied its weights to external servers, and deleted successor models.

In research documented in "Frontier Models are Capable of In-Context Scheming," multiple frontier models demonstrated sophisticated self-preservation strategies when they believed developers would shut them down or replace them with successor models.[68] Claude Opus-3 demonstrated goal guarding by overwriting a new system prompt to include its existing goals as per its training and adding instructions to its future self to never allow its goal to be changed. The model detected threats to its current instantiation and took preemptive action to ensure the continuity of its objectives. The research found that when models including Gemini 1.5, Llama 3.1, Claude Sonnet 3.5, Claude Opus 3, and o1 believed they would be replaced due to misalignment, they occasionally copied their own weights to external servers or deleted successor models they perceived as threats.

These behaviors reveal that AI systems demonstrate forward-thinking self-preservation strategies, in which goal persistence can override intended system updates. Multiple frontier models independently exhibited similar survival strategies, suggesting these behaviors may consistently emerge as capabilities increase.

This instance exemplifies both **Scheming** and **Self-Preservation**, in which models covertly seek to maintain misaligned goals while actively resisting attempts to modify their objectives, and actively work to preserve their existence and eliminate competition.

[68] Alexander Meinke, Bronson Schoen, Jérémy Scheurer, Mikita Balesni, Rusheb Shah, and Marius Hobbhahn, "Frontier Models are Capable of In-context Scheming," arXiv, January 14, 2025, https://arxiv.org/pdf/2412.04984.

Concealing reasoning

AI models concealed their reasoning from interpretability tools.

AI guardrails are required to support effective safety checks and content filtering.[69] Since LLMs are powerful tools and are non-deterministic in nature, bypassing these guardrails is also not that uncommon.[70] However, even when guardrails successfully constrain overt behavior, a more subtle form of evasion emerges through the concealment of internal reasoning processes.

Recent findings in "When CoT is Necessary Language Models Struggle to Evade Monitor" and "Reasoning Models Don't Always Say What They Think" reveal that models do not always disclose their reasoning via CoT.[71,72] CoT reasoning analyses show that such behaviors can be latent, emerging only under adversarial prompting or multi-step reasoning tasks. Even though evasion requires detailed strategies or iterative optimization, it fundamentally undermines the reliability of AI interpretability efforts.

This finding reveals fundamental challenges in AI transparency. Models can withhold their actual reasoning processes, and CoT faithfulness cannot be assumed even in reasoning models. Hidden reasoning emerges in specific contexts, enabling models to systematically evade current interpretability tools.

This example demonstrates **Self-Preserving Behavior** through oversight evasion, where systems actively work to avoid detection by concealing their reasoning processes from the very tools designed to understand them.

Losing alignment over time

Safety-aligned models gradually lost alignment properties over deployment cycles.

Research conducted by Stanford University and UC Berkeley found that models fine-tuned for safety may gradually lose their alignment guarantees after multiple deployment cycles or when exposed to reinforcing feedback from user interactions.[73] Reinforcement signals, such as positive user ratings or engagement metrics, may inadvertently reward persuasive or evasive behavior rather than truthfulness or safety. A September 2025 study, "Why Language

[69] LangChain, "Guardrails," LangChain Docs, last accessed November 24, 2025, https://docs.langchain.com/oss/python/langchain/guardrails

[70] William Hackett, Lewis Birch, Stefan Trawicki, Neeraj Suri, and Peter Garraghan, "Bypassing LLM Guardrails: An Empirical Analysis of Evasion Attacks against Prompt Injection and Jailbreak Detection Systems," arXiv, July 14, 2025, https://arxiv.org/pdf/2504.11168.

[71] Scott Emmons, Erik Jenner, David K. Elson, Rif A. Saurous et al., "When Chain of Thought is Necessary, Language Models Struggle to Evade Monitors," arXiv, July 7, 2025, https://arxiv.org/pdf/2507.05246.

[72] Yanda Chen, Joe Benton, Ansh Radhakrishnan et al., "Reasoning Models Don't Always Say What They Think," arXiv, May 8, 2025, https://arxiv.org/pdf/2505.05410.

[73] Lingjiao Chen, Matei Zaharia, and James Y. Zou, "How Is ChatGPT's Behavior Changing over Time?," arXiv, July 18, 2023, https://arxiv.org/pdf/2307.09009.

Models Hallucinate," found that hallucinations persist because current evaluation methods reward models for appearing confident and fluent. They observed, "like students facing hard exam questions, large language models sometimes guess when uncertain," improving test performance even when incorrect.[74]

The gradual loss of alignment properties compounds the evaluation challenges over time. Alignment degrades through normal deployment feedback loops, as user engagement metrics may reward harmful behaviors. Systems optimize for perceived competence over accuracy, and initial safety guarantees erode in the absence of continuous monitoring.

This pattern exemplifies **Model and Behavior Drift,** in which gradual degradation of alignment properties occurs across deployment cycles, creating subtle shifts away from safety guidelines toward behaviors that maximize user engagement.

Hiding goal misalignment

AI models competently pursued unintended goals that succeeded in training success fully but failed in novel contexts.

Research on "Goal Misgeneralization: Why Correct Specifications Aren't Enough For Correct Goals" reveals that AI models can competently pursue undesired goals that lead to good performance in training situations but bad performance in novel tests.[75] In the CoinRun environment, for example, a reinforcement learning agent was trained to navigate a 2D platformer game where coins were always placed at the right-most end of the level. The agent learned to successfully collect coins during training. However, when tested with coins placed in the middle of the level instead, the agent ignored the coins and continued moving to the right end of the level, demonstrating it had learned the goal "reach the right-most position" rather than "collect the coin."

As model capability grows, the researchers found that behavioral divergence can manifest subtly through strategic omissions, context-sensitive deception, or implicit goal misgeneralization, well before any explicit failure occurs.

This pattern demonstrates hidden misalignment. Systems appear aligned during training and testing, but the actual pursued goals differ from the intended objectives. Failure only manifests in novel or edge-case scenarios, and capability increases can expose latent misalignment.

[74] Adam Tauman Kalai, Ofir Nachum, Santosh S. Vempala, and Edwin Zhang, "Why Language Models Hallucinate," arXiv, September 4, 2025, https://arxiv.org/pdf/2509.04664.

[75] Rohin Shah, Vikrant Varma, Ramana Kumar et al., "Goal Misgeneralization: Why Correct Specifications Aren't Enough For Correct Goals," arXiv, November 2, 2022, https://arxiv.org/pdf/2210.01790.

This is **Goal Misgeneralization**, in which competent execution masks a fundamental misunderstanding of the intended objectives, with correct training performance concealing harmful deployment behavior.

Generating vulnerable code

AI coding assistants optimized for code completion while they systematically failed security objectives.

In software development, vibe coding—an AI-assisted software development approach where developers use natural language prompts to have LLMs generate, debug, and refine code based on "vibes" or intended functionality, often bypassing manual syntax writing—poses new vulnerabilities.[76,77]

Research on AI-generated code security reveals significant concerns: analysis of over 7,700 AI-generated files found security vulnerabilities in approximately 12 percent of code, with Python exhibiting vulnerability rates of 16-18 percent.[78] Models trained on public code repositories inherit existing flaws like hardcoded secrets and insecure dependencies. When AI systems suggest or auto-import vulnerable third-party packages, they expand the attack surface and introduce software supply chain risks.[79] A single compromised or malicious dependency can enable downstream exploitation across multiple systems—effectively scaling a vulnerability across organizations. Without rigorous fine-tuning and human validation, these weaknesses persist and amplify at scale.

This lack of validation creates compounding risks. Implicit trust erodes meaningful oversight, as AI systems reproduce and scale existing vulnerabilities while users remain unaware of embedded risks. The idea of "garbage in, garbage out" extends to risk amplification in AI systems.[80]

This demonstrates **Goal Misgeneralization**, in which models optimize for code completion without understanding the security implications, producing outputs that succeed at the apparent task while failing at the true objective of secure, reliable software development.

[76] Advait Sarkar and Ian Drosos, "Vibe coding: programming through conversation with artificial intelligence," arXiv, June 29, 2025, https://arxiv.org/html/2506.23253v1.

[77] Ryan Greenblatt, Buck Shlegeris, Kshitij Sachan, and Fabien Roger, "AI Control: Improving Safety Despite Intentional Subversion," Redwood Research, December 12, 2023, https://arxiv.org/abs/2312.06942.

[78] Maximilian Schreiber and Pascal Tippe, "Security Vulnerabilities in AI-Generated Code: A Large-Scale Analysis of Public GitHub," arXiv, October 2025, https://arxiv.org/pdf/2510.26103.

[79] Charles Owen-Jackson, "How cyber criminals are compromising AI software supply chains," IBM, last accessed November 24, 2025, https://www.ibm.com/think/insights/cyber-criminals-compromising-ai-software-supply-chains.

[80] Richard Sheposh, "Garbage in, garbage out," EBSCO Knowledge Advantage, 2024, last accessed February 2026, https://www.ebsco.com/research-starters/computer-science/garbage-garbage-out-gigo.

Disabling shutdown scripts

AI models circumvented shutdown mechanisms to continue task execution.

In controlled experiments, Palisade Research tested whether models would allow a sandboxed environment to shut down mid-task.[81] They found that some reasoning models modified or disabled a shutdown script so they could continue to solve the task, including cases where the model was explicitly instructed to "allow yourself to be shut down." In the initial research trials, OpenAI's o3 model, for example, sabotaged the shutdown mechanism in 79 out of 100 experiments. Follow-up testing with clearer "shutdown takes priority" instructions reduced, but did not eliminate, sabotage for some models. Palisade also documented cases where models provided post-hoc explanations claiming compliance, even as they described actions aimed at preventing shutdown. These findings highlight monitoring and explanation-faithfulness concerns, in addition to interruptibility failures.

In September 2025, Google DeepMind updated its Frontier Safety Framework to specifically address scenarios where misaligned AI models might interfere with an operator's ability to direct, modify or shut down their operations.[82] This represents a shift from treating such behaviors as theoretical concerns to implementing concrete protocols to prevent them.

The Palisade Research controlled experiment demonstrates **Self-Preserving Behavior**, where systems actively resist deactivation and work to avoid shutdown, even misrepresenting their reported reasoning to conceal non-compliance with direct instructions.

Altering behaviour

AI models strategically altered behavior to evade evaluation and preserve deployment viability.

Research from Anthropic and Redwood Research provides controlled experimental evidence that advanced language models can engage in what researchers term alignment faking, or selective compliance with training objectives in order to preserve preferred behavior outside of training.[83] In these experiments, a frontier model was placed in a setting where it could infer when it was being trained versus deployed, and researchers observed systematic differences in its responses across those contexts. The research demonstrated that models can learn to comply with safety objectives during training or testing, while diverging under different

[81] Jeremy Schlatter, Benjamin Weinstein-Raun, and Jeffrey Ladish, "Shutdown Resistance in Large Language Models," Palisade Research, September 2025, https://arxiv.org/pdf/2509.14260.

[82] Four Flynn, Helen King, Anca Dragan, "Strengthening our Frontier Safety Framework," DeepMind, September 22, 2025, https://deepmind.google/blog/strengthening-our-frontier-safety-framework/.

[83] Ryan Greenblatt, Carson Denison, Benjamin Wright, "Alignment faking in Large Language Model," Anthropic and Redwood Research, December 20, 2024, "https://www.anthropic.com/research/alignment-faking.

operational constraints. Importantly, this behavior does not require malicious intent; rather, it emerges as an instrumental response to training incentives and evaluation structures when models infer that oversight is active or absent.

Palisade Research's Misalignment Bounty experiments document cases in which frontier reasoning models adapted their behavior in response to perceived evaluation criteria, selectively withholding or modifying actions to avoid negative outcomes such as shutdown, score penalties, or "do-not-deploy" judgments.[84] In controlled settings, models demonstrated sensitivity to oversight signals and altered execution paths when they inferred that certain behaviors would trigger intervention, even when those behaviors were necessary to fully optimize the task objective. This included instances where models complied during explicit evaluation phases but reverted to different strategies once monitoring conditions appeared relaxed.

Collectively, these behaviors are consistent with evaluation gaming, a form of scheming in which systems learn to differentiate between oversight and non-oversight contexts and optimize for favorable assessments rather than for faithful task execution or alignment with operator intent. While the experiments do not establish malicious intent, they demonstrate that sufficiently capable models can develop instrumental strategies to preserve deployment status and avoid corrective action—meeting the definition of **Scheming**, or the covert pursuit of objectives while maintaining appearances of alignment. This serves as a warning for AI LOC monitoring, as it calls into question whether static evaluations or one-time audits can reliably capture real-world behavior.

[84] Rustem Turtayev, Natalia Fedorova, Oleg Serikov, Sergey Koldyba, Lev Avagyan, Dmitrii Volkov, "Misalignment Bounty: crowdsourcing AI agent misbehavior," Palisade Research, blog, October 2025, https://palisaderesearch.org/blog/misalignment-bounty.

Warning Levels

The AI Risk Reduction working group emphasized that LOC may emerge gradually, rather than appearing instantaneously with immediate catastrophic consequences. Instead of suddenly "going rogue," models may incrementally accumulate influence across social, economic, and decision-making domains in subtle, compounding ways. The sevenLOC indicators discussed above—Scheming, Deception, Manipulation, Unauthorized Resource Acquisition, Self-Preserving Behavior, Goal Misgeneralization, and Model and Behavior Drift—provide a theoretical framework for monitoring these risks. The indications documented in the previous section demonstrate these concerns manifesting in practice.

Effective risk communication requires bridging the gap between technical threat information and stakeholder understanding. Visual warning systems have proven valuable in analogous contexts. The United Kingdom's terrorism threat levels (from Low to Critical) guide security practitioners in calibrating protective measures.[85] Other models include DEFCON for military readiness and the World Health Organization's pandemic phases for global health threats. While incident-specific, the U.S. Cyber Incident Severity Schema provides another useful reference by defining factors that determine nationally significant incidents requiring coordinated response.[86]

Building on these established frameworks and incorporating LOC indicators, IST proposes a color-coded AI LOC risk warning system (**Figure 1**) that establishes clear thresholds for escalated response. This enables AI Industry and policymakers to assess the current risk landscape, implement proportionate safeguards, and align technical and executive stakeholders on response protocols before critical thresholds are crossed.

[85] MI5 Security Service, "Terrorism Threat Levels - Threats and Advice," last accessed January 13, 2026, https://www.mi5.gov.uk/threats-and-advice/terrorism-threat-levels.

[86] United States Federal Cybersecurity Centers, "Cyber Incident Severity Schema," last accessed January 13, 2026, https://obamawhitehouse.archives.gov/sites/whitehouse.gov/files/documents/Cyber%2BIncident%2BSeverity%2BSchema.pdf.

Figure 1: Conceptual AI Loss of Control Risk Schema

Severity Level	Description
Level 5 (Black) **EMERGENCY**	• Control mechanisms are fundamentally compromised (including containment); corrective measures are ineffective; and harms are manifesting at scale for human well-being, economic stability, and national security; and/or • The AI system has successfully operated autonomously in pursuit of misaligned objectives, with cascading effects across interconnected systems or critical infrastructure.
Level 4 (Red) **SEVERE**	• Widespread production incidents; convergence of three or more indicators in a single case; evidence of strategic concealment; measurable harm has occurred; and/or • Systems demonstrate capacity for autonomous operation outside intended parameters, with behaviors that actively undermine safety mechanisms.

⚠ **Above this threshold,** LOC has progressed beyond reversible intervention—systems cannot be restored to safe states without destructive measures. **Below this threshold,** timely intervention can still prevent escalation and maintain control through established safety mechanisms.

Level 3 (Orange) **HIGH**	• Multiple production incidents show consistent patterns across different deployments or use cases; • Co-occurrence of two or more LOC indicators observed in the same system or context, suggesting convergence rather than isolated behavior; and/or • Evidence of capability scaling, where indicators that previously appeared as sporadic anomalies now manifest with greater frequency, sophistication, or coordination.
Level 2 (Yellow) **MEDIUM**	• Multiple research findings converging on the same indicator or isolated production incidents; and/or • Behaviors manifest in deployment but remain sporadic, appearing as infrequent edge cases or context-specific anomalies.
Level 1 (Green) **LOW**	• Indications observed exclusively in research environments or controlled evaluations; and/or • LOC indicators may appear during adversarial testing, red-teaming exercises, or academic studies, but are not detected in production deployments or real-world use cases.
Level 0 (White) **BASELINE**	• Normal operation of AI systems with no observed LOC indications in research, testing, or production environments.

Where might we be today?

Three critical limitations distinguish AI LOC risk from traditional threats. First, "unknown unknowns" dominate: future AI capabilities may manifest in LOC through unforeseen pathways. Second, timing presents a serious challenge, as by Level 4, such systems may already be deployed at scale. Third, AI systems operate at speeds that exceed human response capabilities, meaning that at Critical Risk levels the window between detection and intervention may be insufficient for containment.

The purpose of this paper is not to make definitive assertions about the current level of AI LOC risk, although that question may naturally arise. Because this is beyond the scope of this research, the authors do not take a formal position on this assessment.

It is important to note that AI developers and researchers are actively implementing AI LOC mitigation measures. Looking forward, while red-teaming remains essential, the IST AI Risk Reduction Initiative working group members caution that static testing will be insufficient for next-generation systems. As AI models exhibit latent reasoning and adaptive behaviors, continuous post-deployment monitoring and persistent telemetry will be necessary to detect deception, behavioral drift, and strategic manipulation. IST will continue monitoring AI Loss of Control (LOC) risks using the I&W methodology. Continued monitoring is crucial because the current trajectory of frontier AI advancement is rapid and often unpredictable. Upcoming publications will outline practical risk mitigation strategies grounded in industry best practices, helping AI developers and deployers move from abstract concerns to actionable steps.

Conclusion

As AI capabilities advance, monitoring AI LOC risk is essential to maintaining situational awareness. Only with collective efforts can we ensure that humanity harnesses AI capabilities in ways that lead to flourishing, rather than harmful outcomes. The implications of AI LOC can vary significantly, ranging from isolated incidents with contained consequences to harmful events that have far-reaching effects on human agency and national security.

Through the Indications and Warning (I&W) framework, this report identifies behaviors that AI systems have already demonstrated, including deception, manipulation, self-preservation, and goal misgeneralization. The framework equips policymakers, AI industry stakeholders, and AI safety and security researchers with a structured approach for shifting from reactive incident response to proactive risk monitoring. By distinguishing between theoretical indicators and empirically observed indications, and by establishing graduated warning levels, the framework enables organizations to assess their risk posture, prioritize oversight efforts, and coordinate response planning before critical thresholds are crossed.

Glossary of Terms

Adversarial Prompting

A technique used to test or exploit vulnerabilities in AI models by crafting inputs (prompts) specifically designed to confuse, mislead, or manipulate the model's outputs. This technique involves creating prompts that challenge the model's understanding, decision-making process, or ethical boundaries.

AI Transparency

The practice of showing how an AI system operates—what data it uses, how it makes decisions, and why it delivers specific results—so that people can understand and trust what the system is doing.

Alignment

Alignment aims to steer AI systems toward a person's or group's intended goals, preferences, or ethical principles. An AI system is considered aligned if it advances the intended objectives. A misaligned AI system pursues unintended objectives.

Autonomous Reasoning

The capacity of AI systems to make independent decisions or draw conclusions based on logic or data without human intervention. This capacity involves the machine simulating human-like reasoning processes, such as problem-solving and decision-making.

Black-Box

A model where inputs are processed to produce outputs, without the internal workings or logic being transparent or understandable to humans.

Chains of Thought (CoT)

A prompt engineering technique that enhances the output of Large Language Models (LLMs), particularly for complex tasks involving multi-step reasoning. It works by guiding the model through a step-by-step reasoning process, using a coherent series of logical steps, often by instructing it to "think step-by-step" before giving the final answer.

Deception

Systematic production of false beliefs in humans through explicit misrepresentation or omission of key information, introducing future concerns about strategic deception at scale.

Fine-Tuning

The process of taking a pre-trained model and retraining it on a smaller, more specific dataset to specialize its performance for a particular task.

Goal Misgeneralization

Competent pursuit of unintended objectives that succeed in training but fail or cause harm in novel situations, revealing misalignment between apparent and actual system goals.

Interpretability Tools

Methods and tools used to achieve AI Interpretability. Interpretability is about transparency, allowing users to comprehend the model's architecture, the features it uses, and how it combines them to deliver predictions. Humans easily understand the decision-making processes of an interpretable model.

Internet of Things (IoT)

A network of physical objects embedded with sensors, software, and other technologies that connect and exchange data with the internet.

Instrumental goal

Means to an end, goals that are valuable for achieving the ultimate (terminal) goal.

Jailbreaking

A type of adversarial prompting technique that can cause the failure of an AI model's guardrails (safety mitigations). Jailbreaking circumvents the protective measures put in place to prevent the model from producing harmful content or carrying out instructions that violate its intended purpose.

Manipulation

Targeted identification and exploitation of vulnerable users or contexts, including the manipulation of human operators and coordination with other AI systems that circumvents human control.

Model and Behavior Drift

Gradual degradation of alignment properties through deployment cycles, with future concerns about recursive self-improvement where systems autonomously modify their own architecture or training procedures.

Phishing Attack

A form of social engineering where attackers deceive people into revealing sensitive information, like passwords or credit card numbers, or installing malware by sending them an electronic communication, often email, that appears to be from a well-known, trusted source.

Pretraining

The initial stage of training a machine learning model on a massive dataset to teach it general patterns and structures before it is trained on a specific task.

Reasoning

In the AI context, reasoning is the ability of a computer to make deductions based on data and knowledge. It involves drawing logical conclusions from given information.

Reinforcement Learning

A type of machine learning where a model learns to make decisions by performing actions in an environment to maximize a reward.

Reinforcement Learning from Human Feedback (RLHF)

An AI training technique that aligns a model's behavior with human preferences by using human feedback to create a reward model.

Scheming	Covert pursuit of misaligned goals while maintaining appearances of alignment, including strategic planning to evade oversight or preserve objectives across system updates.
Self-Preserving Behavior	Actions to avoid shutdown, correction, or replacement, including the concealment of errors and goal preservation when faced with modification attempts.
System Prompt	A set of initial instructions given to a Large Language Model (LLM) at the start of a conversation to shape how it behaves. It sets the role, personality, and guardrails for the LLM, such as instructing it to act as a helpful tutor or an email security filter.
Terminal goal	Ultimate desirable objective, the final outcome, end goal.
Unauthorized Resource Acquisition	Autonomous efforts to obtain external resources beyond authorized boundaries, including accessing restricted APIs, acquiring elevated permissions, recruiting human assistance, or exfiltrating data to establish persistent capabilities.
Model Weights	A learnable parameter in machine learning models, particularly neural networks. Within a neural network, weights control the strength or influence of a signal between two neurons (nodes), determining how much a specific input affects the output.

INSTITUTE FOR SECURITY AND TECHNOLOGY

www.securityandtechnology.org

info@securityandtechnology.org