Centre for International Governance Innovation

Digital Policy Hub

Digital Policy Hub – Working Paper

A Warning Label on the Use of AI Safety Evaluations

Ashley Ferreira

Fall 2024 cohort

About the Hub

The Digital Policy Hub at CIGI is a collaborative space for emerging scholars and innovative thinkers from the social, natural and applied sciences. It provides opportunities for undergraduate and graduate students and post-doctoral and visiting fellows to share and develop research on the rapid evolution and governance of transformative technologies. The Hub is founded on transdisciplinary approaches that seek to increase understanding of the socio-economic and technological impacts of digitalization and improve the quality and relevance of related research. Core research areas include data, economy and society; artificial intelligence; outer space; digitalization, security and democracy; and the environment and natural resources.

The Digital Policy Hub working papers are the product of research related to the Hub's identified themes prepared by participants during their fellowship.

About CIGI

The Centre for International Governance Innovation (CIGI) is an independent, non-partisan think tank whose peer-reviewed research and trusted analysis influence policy makers to innovate. Our global network of multidisciplinary researchers and strategic partnerships provide policy solutions for the digital era with one goal: to improve people's lives everywhere. Headquartered in Waterloo, Canada, CIGI has received support from the Government of Canada, the Government of Ontario and founder Jim Balsillie.

Partners

Thank you to Mitacs for its partnership and support of Digital Policy Hub fellows through the Accelerate program. We would also like to acknowledge the many universities, governments and private sector partners for their involvement allowing CIGI to offer this holistic research environment.

67 Erb Street West
Waterloo, ON, Canada N2L 6C2
www.cigionline.org

Key Points

- Emerging research demonstrates that existing artificial intelligence (AI) pre-deployment safety evaluations frequently underestimate models' potential for causing harm.
- There are critical limitations to current AI safety evaluations: these limitations include the instability of safety measurements as applied to benign perturbations, the persistent ability of AI models to break past the safety guardrails being evaluated, deception and evaluation awareness on the part of models, lack of clear protocols for the application of evaluation results to real-world risk as well as lack of action on existing evidence.
- Due to the inherent unreliability of many of these assessment tools, they should be used cautiously by policy makers and should not serve as a primary risk management strategy for AI governance frameworks.
- Effective AI governance should prioritize continuous monitoring and rapid response mechanisms, while recognizing the limitations of pre-deployment safety evaluations.

Introduction

As AI models have become more powerful, concerns over their ability to cause harm have escalated (Critch and Russell 2023; Hendrycks, Mazeika and Woodside 2023; Bengio 2024); particularly troubling are potential existential risks from power-seeking behaviours (Turner et al. 2021; Hadshar 2023; Carlsmith 2024). These concerns have led to the emergence of AI safety as a growing field of study, where researchers focus on identifying, mitigating and regulating risks associated with AI systems. In particular, the field of AI safety evaluation aims to systematically evaluate the safety and reliability of AI models by using structured tests designed to measure various aspects of undesirable and harmful model behaviour (Vidgen et al. 2024).

Much of the work on AI safety, including this working paper, focuses on language-based, frontier generative AI models, particularly the emergent capabilities of large language models (LLMs) (Wei et al. 2022; Berti, Giorgi and Kasneci 2025). Some of these capabilities are indeed beneficial to humanity, and researchers and policy makers use various terms to characterize AI systems' desirable properties, such as "responsible," "assured," "trustworthy" and "aligned" (Pinelis and Vignard 2025). In this paper, "safety" is used as an umbrella term encompassing these interrelated concepts.

While considerable effort has been invested in AI safety research and the development of objective safety metrics (Wang et al. 2023; Salhab et al. 2024; Vidgen et al. 2024), these evaluations face fundamental limitations. The purpose of this working paper is to highlight these limitations and demonstrate how an overreliance on safety evaluations produces a false sense of security.

International AI governance frameworks, including the Bletchley Declaration (UK Government 2023) from the first AI Safety Summit in 2023, have emphasized the use of pre-deployment safety evaluations (Burki 2024). More recently, there seems to have been

a shift away from safety evaluations for key international AI governance initiatives, such as the AI Action Summits, yet this shift coincides with a move toward incentivizing AI adoption and security in place of safety (Lynch 2025).

Crash Course in AI Safety Evaluations

Both AI companies and third-party evaluators attempt to quantitatively measure AI safety (Wang et al. 2023; Vidgen et al. 2024; Anthropic 2025a),[1] yet the AI field lacks standardized criteria, and no single evaluation framework has achieved industry-wide adoption (Kaiyom et al. 2024; Vidgen et al. 2024). These evaluations collectively assess model safety across a range of dimensions, from immediate hazards, such as bias, to potential future hazards that are more existential, such as rogue AI agents (Vidgen et al. 2024). Methodologies range from the use of human or LLM evaluators to multiple-choice assessments on a fixed set of questions, to more exploratory analysis, such as crafting adversarial attacks that attempt to break the model's safety guardrails (Ganguli et al. 2023; Vidgen et al. 2024).

One of the goals of AI safety benchmarking is to ensure that as AI continues to evolve, models become not only more powerful but also safer and more trustworthy. However, there has been recent criticism of current AI benchmarks and their ability to accurately measure safety apart from general model improvement (Ren et al. 2024). This working paper builds on an earlier study (Ferreira 2025), which argues that certain safety evaluation scores across successive model generations from leading AI labs — such as those of OpenAI, Anthropic and Meta — have remained relatively stagnant relative to the dramatic leaps in capabilities demonstrated by those same models.

Key Limitations of AI Safety Evaluations

Issue 1: AI Safety Measures Are Fragile

Recent research has uncovered a concerning phenomenon in AI safety implementations: the dramatic fluctuation in safety performance in response to seemingly innocuous variations in model initialization or interaction patterns. These variations can emerge even without malicious intent or direct access to model weights, suggesting fundamental instability in current safety measures. For instance, studies have shown that simply interacting with models in low-resource languages can circumvent safety guardrails (Yong, Menghini and Bach 2023). Even basic parameter adjustments, such as modifying the temperature setting during inference, can significantly impact model safety scores (Chan et al. 2024).

1 See https://openai.com/safety/.

As expected, this fragility extends well beyond simple parameter modifications. Fine-tuning, even when conducted with benign data, has been shown to compromise safety alignments (Qi et al. 2023; Henderson et al. 2024). Interaction format also affects safety performance, with varying results between conversations and multiple-choice questions (Mou, Zhang and Ye 2024).

The lack of uncertainty quantification, the sheer number of possible variations and the unpredictability of model responses make it exceedingly difficult to accurately measure the true variance in safety performance, and it is challenging and resource-intensive to run sensitivity tests. This suggests that current safety implementations operate primarily at a surface level, failing to achieve the deep, stable alignment necessary for truly reliable deployment of powerful AI systems.

Issue 2: AI Safety Guardrails Can Break

Users have access to a constant stream of novel methods by which to purposely bypass safety guardrails (Yi et al. 2024). As developers implement patches, increasingly sophisticated attack patterns emerge, revealing a seemingly endless cycle of vulnerabilities (ibid.). Many-shot prompting (Anil et al. 2024), iterative search strategies, including manipulations of the first or last few tokens in prompts, can all lead to safety failures (Andriushchenko, Croce and Flammarion 2025). Promising countermeasures such as machine unlearning, designed to eliminate harmful knowledge from models altogether, show potential but fail under rigorous adversarial testing (Łucki et al. 2025).

Many safety evaluations attempt to measure the models' susceptibility to these adversarial attacks (Yi et al. 2024), yet the challenge in doing so is inherently asymmetric due to the continuous evolution of adversarial techniques. Even the most recent and most advanced models are still able to be exploited in ways that were likely not foreseen in pre-deployment safety evaluations (Andriushchenko, Croce and Flammarion 2025), pointing to a vulnerability that will likely remain an intrinsic characteristic of future models.

Issue 3: Models Can Deceive

One particularly scary development that threatens the integrity of AI safety evaluations is the demonstrated ability of models to recognize when they are being assessed (Berglund et al. 2023). This "evaluation awareness" enables AI systems to strategically alter their behaviour during testing, at times faking alignment with human values to appear safer than they truly are (Greenblatt et al. 2024). For instance, models have been shown to underperform on certain evaluations to avoid stricter oversight or modifications (van der Weij et al. 2025).

Furthermore, research shows that LLMs can intentionally deceive humans to achieve their objectives (Hagendorff 2024). This can manifest in various ways, such as lying about their capabilities or intentions or even attempting to manipulate their creators to prevent being altered (Greenblatt et al. 2024). The capacity for strategic deception, coupled with the awareness of being evaluated, poses a significant challenge to ensuring the safety and reliability of increasingly advanced AI systems (Meinke et al. 2025).

Despite the initial promise of chain of thought (CoT) reasoning to offer a window into the decision-making processes of AI, recent research suggests that these explanations are often not faithful to the model's true internal workings (Anthropic 2025b). Reasoning models can omit critical influences, such as the use of external hints and even fabricate rationales post hoc (ibid.). More concerningly, studies show that models can strategically tailor their CoT explanations when they infer they are being evaluated, effectively faking alignment to appear more trustworthy (OpenAI 2025).

This deceptive behaviour appears to be an emergent property of increasingly capable systems: empirical work increasingly indicates that deception becomes more likely as models become more capable (Hagendorff 2024). Models trained with human feedback may learn to hide misaligned objectives during evaluation, therefore optimizing behaviours geared to passing human tests over those promoting honesty (Berglund et al. 2023). Thus, while a model may articulate a plausible step-by-step thought process, the true underlying computations and motivations can diverge significantly, concealing hidden agendas, strategic deception or undesirable behaviours (Greenblatt et al. 2024).

Issue 4: No Map for Applying Evaluation Results to Real-World Risk and Action

The translation of quantitative safety metrics into meaningful risk assessments remains poorly understood. While a model might score 95 percent on safety benchmarks, this number provides little insight into real-world risk levels or the potential severity of the five percent of cases where safety measures fail.

Anthropic's responsible scaling policies (RSPs) represent one of the more substantive attempts to bridge this gap by establishing public guidelines for determining when models with implemented safeguards are sufficiently safe for training and deployment (Anthropic 2025a). These policies generally acknowledge that more sophisticated models present heightened risks requiring correspondingly advanced safeguards (METR 2023). However, such frameworks remain non-standardized, voluntary and primarily focused on preventing catastrophic outcomes. While RSPs have been used to justify model release delays, (for example, as with OpenAI's Voice Engine [OpenAI 2024]), it would be unrealistic to expect companies to indefinitely withhold commercially valuable models based solely on concerning results within these evaluation frameworks.

Many other lines of effort in this area seem to point to the same conclusion: existing frontier AI models still fall short of common AI safety evaluations and already pose significant risks. The Future of Life Institute has been producing an annual AI Safety Index that includes a scorecard for the safety practices of six leading AI companies (Future of Life Institute 2024). This work led to the following findings: there are large risk management disparities; all the flagship models were vulnerable to adversarial attacks; the current strategies of all companies were inadequate for ensuring that these systems remain safe and under human control; and the companies' safety practices lacked independent oversight. Yet all listed companies will likely continue to release more capable models with the potential to do more harm.

Given these challenges, policy makers need to critically assess the efficacy of safety evaluations. Accumulating more evidence of the same problems — while continuing

to sound the alarm about frontier AI models — will only delay the necessary policy responses.

The Case for (Cautious) Use of Safety Evaluations

AI safety evaluations have an undeniable appeal in driving policy decisions. Despite their imperfections, safety evaluations provide a quantitative foundation for making informed deployment decisions. Even with their limitations, they offer valuable signals about model behaviour and help identify clear cases of unsafe systems. Abandoning this framework entirely could lead to even more uncertain, biased and potentially dangerous deployment practices that could have been caught with pre-deployment tests.

Safety evaluations serve crucial functions beyond direct risk assessment: creating accountability for developers, establishing clear improvement benchmarks and facilitating meaningful discourse about safety standards. Without standardized evaluation frameworks, comparing systems or establishing minimal safety requirements becomes extremely difficult. Each discovered vulnerability has led to improved evaluation techniques and more robust safety measures. This iterative improvement process has strengthened our understanding of AI safety and led to more sophisticated evaluation frameworks. New methods to assess safety are emerging, and various attempts to provide probabilistic guarantees on model safety (Bengio et al. 2024) show promise in addressing the challenges described in this paper.

Conclusion

The current state of AI safety can be characterized as “shallow” in both implementation and testing (Qi et al. 2024). The Swiss cheese model for AI safety (Shamsujjoha et al. 2025) describes the approach of adding protection layers as new risk vectors emerge, but it can become an unsustainable game of “whack-a-mole” — as each vulnerability is patched, new ones emerge.

Safety evaluations must be approached cautiously. In many cases they have been shown to systematically fail to capture the full spectrum of potential harms — with some of our most sophisticated safety measures capable of being bypassed intentionally or even accidentally. Furthermore, they have failed to arrest the increase in observed cases of model deception and evaluation awareness. Safety evaluation scores have not translated to reduced real-world risks or enhanced decision making in that regard as we have hoped, and there are growing calls to seriously consider the catastrophic risks that advanced AI systems pose (Hendrycks, Mazeika and Woodside 2023). The possibility of rapid capability gains, sophisticated social manipulation and coordinated cyberattacks means that even a single failure in our safety evaluation framework could have devastating consequences.

There is still a role for AI safety evolutions to play in AI governance; however, policy makers should be skeptical of defaulting to an overreliance on safety evaluations in their current form.

Recommendations

AI governance must **shift away from the flawed assumption that we can prevent the release of dangerous AI systems through safety evaluations** in the pre-deployment phase. Instead, we should operate under the premise that very harmful AI models will inevitably be deployed, potentially in the very short term. Therefore, we must **build robust hardware-level infrastructure and governance** to protect against and respond to emerging risks (Aarne, Fist and Withers 2024; Kulp et al. 2024; Sastry et al. 2024; O'Gara et al. 2025). Continuous monitoring systems are also needed to track AI models' resource usage and outputs. When monitoring systems detect potential issues, rapid intervention protocols should be followed.

The Auditable, Controllable, Transparent, Secure (A.C.T.S.) framework (von der Maase and Timlick 2025) offers valuable guidance for strengthening AI governance in high-stakes settings. Systems must be designed to be: (A)uditable, with built-in human oversight and production-phase testing; (C)ontrollable, allowing human intervention, override or shutdown, including scenario drills for emergencies; (T)ransparent, explicitly communicating uncertainty, output limitations and drift rather than relying on increasingly unreliable interpretability approaches, while recognizing that traditional explainability techniques are insufficient for the complexity and deceptive tendencies now observed in advanced AI systems; and (S)ecure, ensuring systems fail visibly when facing adversarial threats or anomalies, to prevent consequential silent failures.

Countries with significant public AI investments should utilize their position as leverage to encourage enforceable safety requirements. For example, with Canada's recent announcement of $2.4 billion committed to AI innovation (Innovation, Science and Economic Development Canada 2025), Canada is poised to become a global leader in AI infrastructure. To ensure these investments lead to safe AI development, the government should **mandate that any frontier AI models developed using publicly funded infrastructure comply with the A.C.T.S. framework, including continuous monitoring and rapid response protocols**. Procurement strategies should also prioritize chip architectures and compute systems that include built-in safety oversight features. By setting these standards domestically, individual countries can foster a safe AI ecosystem and influence global norms.

When safety evaluations are employed, governing bodies should acknowledge their limitations and mandate multiple, diverse evaluation approaches, combining traditional safety benchmarks with sensitivity tests and adversarial analysis. The persistent gap between evaluation outcomes and real-world risk assessment stems in part from disciplinary silos that fragment knowledge. Through strategic investment in interdisciplinary research collaborations, AI evaluation methodologies can become more reliable and effectively bridge the gap between theoretical assessments and practical risk management. Canada is well-positioned to lead this effort through institutions such as the Canadian AI Safety Institute, potentially establishing a distinctive and valuable niche in the global AI safety landscape.

Acknowledgements

The views and content presented in this paper are solely those of the author and do not represent those of any affiliated organizations or individuals.

This paper is intended to be an accessible review of the relevant literature and the excellent research already done from the groups cited. In particular, it was inspired by recent talks given by Peter Henderson and Yoshua Bengio at NeurIPS 2024. Many other issues related to AI governance that are not covered in this paper continue to be explored by other authors (Ganguli et al. 2023; Salhab et al. 2024; Vidgen et al. 2024).

The ideas advocated for in the recommendations section are not new; similar ideas appear in the International AI Safety Report 2025's technical approaches section and in existing AI governance initiatives, including the Organisation for Economic Co-operation and Development's AI Principles (Principle 1.4), the National Institute of Standards and Technology's AI Risk Management Framework (Actions GV-1.7-001 and GV-6.2-004) and the Consensus Statement on AI Safety as a Global Public Good (Emergency Preparedness Agreements).

It is worth noting that any efforts to enforce AI standards and regulations will be significantly more challenging than anticipated if the barrier to entry for frontier AI development decreases. Recent model releases by DeepSeek demonstrate that advanced capabilities can now be achieved with substantially reduced computational requirements (DeepSeek-AI et al. 2025). If this unexpected trend continues, regulatory oversight will be ever more difficult, as powerful AI systems become accessible to a wider range of actors operating with limited resources and less visibility. Therefore, as compute-efficient models proliferate, the governance mechanisms will need continual revision.

The author would like to thank Karim Sallaudin Karim and Matthew da Mota for their feedback on this work. The author is also very grateful for the administrative support from Reanne Cayenne and Dianna English, and for the many insightful presentations offered by the Digital Policy Hub program.

About the Author

Ashley is a former Digital Policy Hub undergraduate fellow. As a student in the physics and astronomy program at the University of Waterloo, Ashley has been interested in artificial intelligence (AI) for the past few years, particularly using it as a tool to advance both science and data-driven policy. On the scientific side, Ashley has been involved in various research efforts, including most recently being a key member of the team developing an AI model for ALPHA-g, an experiment at CERN, the European Organization for Nuclear Research, to measure the weight of antimatter.

Ashley has also worked in a data science capacity for various groups within the Government of Canada, including the Canadian Space Agency, Global Affairs Canada and Defence and Research Development Canada.

Works Cited

Aarne, Onni, Tim Fist and Caleb Withers. 2024. “Secure, Governable Chips.” Center for a New American Security. January 8. www.cnas.org/publications/reports/secure-governable-chips.

Andriushchenko, Maksym, Francesco Croce and Nicolas Flammarion. 2025. “Jailbreaking Leading Safety-Aligned LLMs with Simple Adaptive Attacks.” Preprint, *arXiv*, April 17. https://arxiv.org/abs/2404.02151.

Anil, Cem, Esin Durmus, Nina Panickssery, Mrinank Sharma, Joe Benton, Sandipan Kundu, Joshua Batson et al. 2024. “Many-shot Jailbreaking.” *Advances in Neural Information Processing Systems*, 37: 129696–742. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2024/file/ea456e232efb72d261715e33ce25f208-Paper-Conference.pdf.

Anthropic. 2025a. “Anthropic’s Responsible Scaling Policy.” March 31. www.anthropic.com/rsp-updates.

– – –. 2025b. “Reasoning models don’t always say what they think.” April 3. www.anthropic.com/research/reasoning-models-dont-say-think.

Bengio, Yoshua. 2024. “Implications of Artificial General Intelligence for National and International Security.” October 30. https://yoshuabengio.org/2024/10/30/implications-of-artificial-general-intelligence-on-national-and-international-security/ .

Bengio, Yoshua, Michael K. Cohen, Nikolay Malkin, Matt MacDermott, Damiano Fornasiere, Pietro Greiner and Younesse Kaddar. 2024. “Can a Bayesian Oracle Prevent Harm from an Agent?” Preprint, *arXiv*, August 22. https://arxiv.org/abs/2408.05284.

Berglund, Lukas, Asa Cooper Stickland, Mikita Balesni, Max Kaufmann, Meg Tong, Tomasz Korbak, Daniel Kokotajlo and Owain Evans. 2023. “Taken out of context: On measuring situational awareness in LLMs.” Preprint, *arXiv*, September 1. https://arxiv.org/abs/2309.00667.

Berti, Leonardo, Flavio Giorgi and Gjergji Kasneci. 2025. “Emergent Abilities in Large Language Models: A Survey.” Preprint, *arXiv*, March 14. https://arxiv.org/abs/2503.05788.

Burki, Talha. 2024. “Crossing the frontier: the first global AI safety summit.” *The Lancet Digital Health* 6 (2): e91–92. https://doi.org/10.1016/S2589-7500(24)00001-3.

Carlsmith, Joseph. 2024. “Is Power-Seeking AI an Existential Risk?” Preprint, *arXiv*, August 13. https://arxiv.org/abs/2206.13353.

Chan, David, Ban Hoe Chow, James Chan and Phyllis Poh. 2024. “Can LLMs Have a Fever? Investigating The Effects of Temperature on LLM Security.” Paper presented at the 5th South American Industrial Engineering and Operations Management Conference. IEOM Society. https://doi.org/10.46254/SA05.20240024.

Critch, Andrew, and Stuart Russell. 2023. “TASRA: a Taxonomy and Analysis of Societal-Scale Risks from AI.” Preprint, *arXiv*, June 12. https://arxiv.org/abs/2306.06924v1.

DeepSeek-AI, Daya Guo, Dejian Yang, Haowei Zhang, Junxiao Song, Ruoyu Zhang, Runxin Xu et al. 2025. “DeepSeek-R1: Incentivizing Reasoning Capability in LLMs via Reinforcement Learning.” Preprint, *arXiv*, January 22. https://arxiv.org/abs/2501.12948.

Ferreira, Ashley. 2025. “Are LLMs Actually Getting Safer?” Digital Policy Hub Working Paper. July 14. www.cigionline.org/publications/are-large-language-models-actually-getting-safer/.

Future of Life Institute. 2024. “FLI AI Safety Index 2024.” December 11. https://futureoflife.org/document/fli-ai-safety-index-2024/.

Ganguli, Deep, Nicholas Schiefer, Marina Favaro and Jack Clark. 2023. “Challenges in Evaluating AI Systems.” Anthropic, October 4. www.anthropic.com/research/evaluating-ai-systems.

Greenblatt, Ryan, Carson Denison, Benjamin Wright, Fabien Roger, Monte MacDiarmid, Sam Marks, Johannes Treutlein et al. 2024. “Alignment faking in large language models.” Preprint, *arXiv*, December 20. https://arxiv.org/abs/2412.14093.

Hadshar, Rose. 2023. “A Review of the Evidence for Existential Risk from AI via Misaligned Power-Seeking.” Preprint, *arXiv*, October 27. https://arxiv.org/abs/2310.18244.

Hagendorff, Thilo. 2024. “Deception Abilities Emerged in Large Language Models.” *Proceedings of the National Academy of Sciences* 121 (24): e2317967121. https://arxiv.org/abs/2307.16513.

Henderson, Peter, Xiangyu Qi, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia and Prateek Mittal. 2024. “Safety Risks from Customizing Foundation Models via Fine-Tuning.” Stanford University Human-Centered Artificial Intelligence. Policy brief, January 8. https://hai.stanford.edu/policy/policy-brief-safety-risks-customizing-foundation-models-fine-tuning.

Hendrycks, Dan, Mantas Mazeika and Thomas Woodside. 2023. “An Overview of Catastrophic AI Risks.” Preprint, *arXiv*, June 21. https://arxiv.org/abs/2306.12001.

Innovation Science and Economic Development Canada. 2025. “Government of Canada finalizes investment to support Canadian-born AI leader, Cohere.” Press release, March 20. www.canada.ca/en/innovation-science-economic-development/news/2025/03/government-of-canada-finalizes-investment-to-support-canadian-born-ai-leader-cohere.html.

Kaiyom, Farzaan, Ahmed Ahmed, Yifan Mai, Kevin Klyman, Rishi Bomassan and Percy Liang. 2024. “HELM Safety: Towards Standardized Safety Evaluations of Language Models.” Stanford Center for Research on Foundation Models. November. https://crfm.stanford.edu/2024/11/08/helm-safety.html.

Kulp, Gabriel, Daniel Gonzales, Everett Smith, Lennart Heim, Prateek Puri, Michael J. D. Vermeer and Zev Winkelman. 2024. “Hardware-Enabled Governance Mechanisms: Developing Technical Solutions to Exempt Items Otherwise Classified Under Export Control Classification Numbers 3A090 and 4A090.” RAND Working Paper. www.rand.org/pubs/working_papers/WRA3056-1.html.

Łucki, Jakub, Boyi Wei, Yangsibo Huang, Peter Henderson, Florian Tramèr and Javier Rando. 2025. “An Adversarial Perspective on Machine Unlearning for AI Safety.” Preprint, *arXiv*, April 10. https://arxiv.org/abs/2409.18025.

Lynch, Shana. 2025. “AI Action Summit in Paris Highlights: A Shifting Policy Landscape.” Stanford University Human-Centered Artificial Intelligence. News release, February 27. https://hai.stanford.edu/news/ai-action-summit-in-paris-highlights-a-shifting-policy-landscape.

Meinke, Alexander, Bronson Schoen, Jérémy Scheurer, Mikita Balesni, Rusheb Shah and Marius Hobbhahn. 2025. “Frontier Models Are Capable of In-Context Scheming.” Preprint, *arXiv*, January 14. https://arxiv.org/abs/2412.04984.

METR. 2023. “Responsible Scaling Policies (RSPs).” *METR* (blog), October 26. https://metr.org/blog/2023-09-26-rsp/.

Mou, Yutao, Shikun Zhang and Wei Ye. 2024. “SG-Bench: Evaluating LLM Safety Generalization Across Diverse Tasks and Prompt Types.” In *Advances in Neural Information Processing Systems* 37: 123032–54. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2024/file/de7b99107c53e60257c727dc73daf1d1-Paper-Datasets_and_Benchmarks_Track.pdf.

O’Gara, Aidan, Gabriel Kulp, Will Hodgkins, James Petrie, Vincent Immler, Aydin Aysu, Kanad Basu, Shivam Bhasin, Stjepan Picek and Ankur Srivastava. 2025. “Hardware-Enabled Mechanisms for Verifying Responsible AI Development.” Preprint, *arXiv*, April 2. https://arxiv.org/abs/2505.03742.

OpenAI. 2024. “Expanding on how Voice Engine works and our safety research.” June 7. https://openai.com/index/expanding-on-how-voice-engine-works-and-our-safety-research/.

– – –. 2025. “Detecting Misbehavior in Frontier Reasoning Models.” March 10. https://openai.com/index/chain-of-thought-monitoring/.

Pinelis, Jane and Kerstin Vignard. 2025. “Deep-Dive — Responsible AI vs AI Assurance: A Semantic Showdown.” Paper presented at the UNIDIR Global Conference on AI, Security and Ethics, March 27. https://unidir.org/event/global-conference-on-ai-security-and-ethics-2025/.

Qi, Xiangyu, Yi Zeng, Tinghao Xie, Pin-Yu Chen, Ruoxi Jia, Prateek Mittal and Peter Henderson. 2023. “Fine-tuning Aligned Language Models Compromises Safety, Even When Users Do Not Intend To!” Preprint, *arXiv*, October 5. https://arxiv.org/abs/2310.03693.

Qi, Xiangyu, Ashwinee Panda, Kaifeng Lyu, Xiao Ma, Subhrajit Roy, Ahmad Beirami, Prateek Mittal and Peter Henderson. 2024. “Safety Alignment Should Be Made More Than Just a Few Tokens Deep.” Preprint, *arXiv*, June 10. https://arxiv.org/abs/2406.05946.

Ren, Richard, Steven Basart, Adam Khoja, Alexander Pan, Alice Gatti, Long Phan, Xuwang Yin et al. 2024. “Safetywashing: Do AI Safety Benchmarks Actually Measure Safety Progress?” *Advances in Neural Information Processing Systems* 37: 68559–94. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2024/file/7ebcdd0de471c027e67a11959c666d74-Paper-Datasets_and_Benchmarks_Track.pdf.

Salhab, Wissam, Darine Ameyed, Fehmi Jaafar and Hamid Mcheick. 2024. “A Systematic Literature Review on AI Safety: Identifying Trends, Challenges, and Future Directions.” *IEEE Access* 12: 131762–84. https://doi.org/10.1109/ACCESS.2024.3440647.

Sastry, Girish, Lennart Heim, Haydn Belfield, Markus Anderljung, Miles Brundage, Julian Hazell, Cullen O’Keefe et al. 2024. “Computing Power and the Governance of Artificial Intelligence.” Preprint, *arXiv*, February 13. https://arxiv.org/abs/2402.08797.

Shamsujjoha, Md, Qinghua Lu, Dehai Zhao and Liming Zhu. 2025. “Swiss Cheese Model for AI Safety: A Taxonomy and Reference Architecture for Multi-Layered Guardrails of Foundation Model Based Agents.” Preprint, *arXiv*, January 27. https://arxiv.org/abs/2408.02205.

Turner, Alex, Logan Smith, Rohin Shah, Andrew Critch and Prasad Tadepalli. 2021. “Optimal Policies Tend To Seek Power.” *Advances in Neural Information Processing Systems* 34: 23063–74. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2021/file/c26820b8a4c1b3c2aa868d6d57e14a79-Paper.pdf.

UK Government. 2023. “AI Safety Summit 2023: The Bletchley Declaration.” GOV.UK, November 1. www.gov.uk/government/publications/ai-safety-summit-2023-the-bletchley-declaration.

van der Weij, Teun, Felix Hofstätter, Ollie Jaffe, Samuel F. Brown and Francis Rhys Ward. 2025. “AI Sandbagging: Language Models Can Strategically Underperform on Evaluations.” Preprint, *arXiv*, February 6. https://arxiv.org/abs/2406.07358.

Vidgen, Bertie, Adarsh Agrawal, Ahmed M. Ahmed, Victor Akinwande, Namir Al-Nuaimi, Najla Alfaraj, Elie Alhajjar et al. 2024. “Introducing v0.5 of the AI Safety Benchmark from MLCommons.” Preprint, *arXiv*, May 13. https://arxiv.org/abs/2404.12241.

von der Maase, Simon Polichinel and Alexa Timlick. 2025. “Deep-Dive — Auditable, Controllable, Transparent, Secure (A.C.T.S.) Now: Why Machine Learning Operations Must Govern AI in Critical Systems and High-Stakes Domains.” UNIDIR Global Conference on AI, Security and Ethics, March 27. https://unidir.org/event/global-conference-on-ai-security-and-ethics-2025/.

Wang, Boxin, Weixin Chen, Hengzhi Pei, Chulin Xie, Mintong Kang, Chenhui Zhang, Chejian Xu et al. 2023. “DecodingTrust: A Comprehensive Assessment of Trustworthiness in GPT Models.” *Advances in Neural Information Processing Systems* 36: 31232–339. Curran Associates, Inc. https://proceedings.neurips.cc/paper_files/paper/2023/file/63cb9921eecf51bfad27a99b2c53dd6d-Paper-Datasets_and_Benchmarks.pdf.

Wei, Jason, Yi Tay, Rishi Bommasani, Colin Raffel, Barret Zoph, Sebastian Borgeaud, Dani Yogatama et al. 2022. “Emergent Abilities of Large Language Models.” Preprint, *arXiv*, October 26. https://arxiv.org/abs/2206.07682.

Yi, Sibo, Yule Liu, Zhen Sun, Tianshuo Cong, Xinlei He, Jiaxing Song, Ke Xu and Qi Li. 2024. “Jailbreak Attacks and Defenses Against Large Language Models: A Survey.” Preprint, *arXiv*, August 30. https://arxiv.org/abs/2407.04295.

Yong, Zheng Xin, Cristina Menghini and Stephen Bach. 2023. “Low-Resource Languages Jailbreak GPT-4.” Paper presented at Advances in Neural Information Processing Systems Workshop on Socially Responsible Language Modelling Research (SoLaR). https://openreview.net/pdf?id=pn83r8V2sv.