Regulating the Agency of LLM-based Agents

Seán Boddy
Berkman Klein Center for Internet & Society
Harvard University
sboddy@law.harvard.edu

Joshua Joseph
Berkman Klein Center for Internet & Society
Harvard University
jjoseph@law.harvard.edu

Abstract

As increasingly capable large language model (LLM)-based agents are developed, the potential harms caused by misalignment and loss of control grow correspondingly severe. To address these risks, we propose an approach that directly measures and controls the agency of these AI systems. We conceptualize the agency of LLM-based agents as a property independent of intelligence-related measures and consistent with the interdisciplinary literature on the concept of agency. We offer (1) agency as a system property operationalized along the dimensions of preference rigidity, independent operation, and goal persistence, (2) a representation engineering approach to the measurement and control of the agency of an LLM-based agent, and (3) regulatory tools enabled by this approach: mandated testing protocols, domain-specific agency limits, insurance frameworks that price risk based on agency, and agency ceilings to prevent societal-scale risks. We view our approach as a step toward reducing the risks that motivate the "Scientist AI" paradigm, while still capturing some of the benefits from limited agentic behavior.

1 Introduction

As large language model (LLM)-based agents become increasingly capable of performing complex, multi-step tasks in real-world environments, the need for effective control mechanisms becomes critical [Toner et al., 2024, Bengio et al., 2025b]. Recent work has underscored this need by demonstrating behaviors such as deception, blackmail, goal-guarding, and resistance to shutdown attempts [Pan et al., 2024, Meinke et al., 2025, Lynch et al., 2025]. These behaviors emerge as ordinary consequences of the current training paradigms, and these risks increase with the capabilities of the model [Dung, 2023, Bengio et al., 2025a]. As a result, misaligned agents have the potential to cause even greater harms by resisting human control, self-replicating, and disrupting critical infrastructure [Mitre and Predd, 2025, Clymer et al., 2024].

Despite these risks, we lack tools and frameworks to measure or control the source of the problematic behavior: the system's agency [Bengio et al., 2025a]. Therefore, our aim is to make the agency of an LLM-based agent the direct target of regulatory observation and intervention. We do this by conceptualizing agency as a measurable system property distinct from intelligence and operationalizing it along the dimensions of preference rigidity, independent operation, and goal persistence, which are consistent with the interdisciplinary literature on the concept of agency. We propose a representation engineering [Zou et al., 2025] approach, which builds on Chen et al. [2024] and trains linear probes and hypothesize that these probes allow for both the measurement and control of agency. These "agency sliders," function similarly to Chen et al. [2024]'s dashboard sliders for the LLM's representation of the user's age, gender, educational level, and socioeconomic status (see Figure 2 from Chen et al. [2024]). This approach enables a variety of regulatory mechanisms consistent with the motivations underlying the "Scientist AI" paradigm [Bengio et al., 2025a]: mandated testing protocols for high-risk applications, domain-specific agency limits calibrated to risk levels, insurance frameworks that price premiums based on measurable agency characteristics, and hard ceilings on agency levels to prevent societal-scale risks.

Preprint. Under review.

2 Background and Related Work

Agentic Misalignment. Research and testing of frontier LLMs have demonstrated that these models are capable of broad manifestations of misalignment. AI systems have demonstrated reward hacking and strategic deception to meet their objectives [Pan et al., 2024, Meinke et al., 2025]. Fine-tuning models from narrowly scoped misaligned content has been shown to induce broad misalignment in the model [Betley et al., 2025, Chua et al., 2025], and misaligned parent models have passed down misaligned characteristics to their child models [Cloud et al., 2025]. With regard to agents, AI systems have also engaged in goal-guarding and sandbagging [Meinke et al., 2025, van der Weij et al., 2024], and have resisted attempts to shut them down using extreme measures [Lynch et al., 2025]. Misalignment is difficult to detect, is not the result of faulty architecture or training, greatly diminishes the usefulness of an AI system, and is the ordinary result of creating these AI systems through machine learning [Dung, 2023]. Analysis of misalignment also shows that the risk posed by misaligned systems only increases as these systems become more capable [Dung, 2023]. This has led to an increasing need to control the level of agency in newly developed LLM-based agents.

Scientist AI. Bengio et al. [2025a] proposed the creation of "Scientist AI," or AI models that are non-agentic, trustworthy, and safe by design. These systems would behave like a scientist, explaining the world from observations but never taking action to please humans or fulfill ideals. The authors state that such systems could aid in scientific research and protect against misaligned AI agents, recognizing that agent development is likely to continue despite the risks. Recognizing this eventuality, we see a growing need to quantify and control the level of agency in an AI system.

AI Agent Benchmarks Conflate Intelligence and Agency. AI agent benchmarks typically measure qualities of agents such as reasoning ability, tool selection, and task completion rates, which conflates measures of intelligence and agency. For example, safety-focused benchmarks evaluate agents' adherence to constraints and their resistance to harmful instructions [Zhang et al., 2025], while others measure goal drift and task deviation over extended interactions [Arike et al., 2025]. Another example, τ-bench [Yao et al., 2024] examines multi-turn interactions and tool usage, and Kwa et al. [2025] evaluate agents' ability to complete increasingly long and complex tasks. Sophisticated real-world benchmarks such as Liu et al. [2024], Boisvert et al. [2025], Xu et al. [2024] primarily measure task completion rates. Meanwhile, benchmarks that test long-term coherence, such as Vending-Bench [Backlund and Petersson, 2025], reveal how agents fail to maintain consistent autonomous behavior over extended periods, yet still do not provide a framework for measuring agency as a property distinct from intelligence.

Representation Engineering. Representation engineering (RepE) is an approach to top-down transparency that treats representations at the population level, not individual circuits, as the primary object for monitoring and steering abstract concepts and safety-relevant variables (e.g., honesty, harmfulness, power-seeking) [Zou et al., 2025]. These techniques have enabled practical applications such as increasing truthfulness [Li et al., 2024], reducing sycophancy [Papadatos and Freedman, 2024], improving instruction-following [Stolfo et al., 2025], and allowing the user to control the model's representation of the user [Chen et al., 2024].

3 Measurement and Control of the Agency of an LLM-based Agent

Our proposed approach views agency as a measurable and controllable system property distinct from measures of intelligence [Hutter, 2000, Chollet, 2019, Morris et al., 2024, Chollet et al., 2024]. In Section 3.1, we establish a three-dimensional conceptualization of agency that captures the core attributes common to interdisciplinary conceptions of agency while remaining operationalizable through current RepE techniques. In Section 3.2, we demonstrate how adapting the approach of Chen et al. [2024], can enable both measurement and control of these dimensions through "agency sliders." This white-box approach operates on internal representations rather than outputs, providing robustness against deceptive output-based behaviors while enabling precise calibration of agency levels to match deployment contexts. The resulting framework not only provides the technical foundation for safer LLM-based agent deployment, but also establishes the measurable parameters necessary for the regulatory mechanisms we explore in Section 4.

3.1 Dimensions of Agency

There are wide-ranging conceptions of agency that span the fields of biology [Okasha, 2024, DiFrisco and Gawne, 2025], philosophy [Schlosser, 2019, Perez-Orosio and Wykowska, 2020, Ferrero, 2022], psychology [Moore, 2016], law [Ayres and Balkin, 2024, Spamann and Frankenreiter, 2024], and computer science [Franklin and Graesser, 1997, Kenton et al., 2022, Chan et al., 2023, Toner et al., 2024]. For our purposes, we propose an initial set of three dimensions that appear most frequently in the literature. We do not claim that the dimensions are ultimately sufficient, but we offer them as a starting point to operationalize a conception of agency independent of intelligence. These dimensions are: preference rigidity, independent operation, and goal persistence.

These three dimensions of agency manifest in LLM-based agents as they act in the real-world. Preference rigidity captures how consistently a system maintains preferences and decision patterns across tasks (such as prioritizing certain solution approaches or quality standards regardless of immediate task requirements), independent operation reflects the degree to which a system can proceed without continuous human oversight by making reasonable assumptions from context and available data rather than repeatedly seeking clarification, and goal persistence describes how a system continuously pursues its goals despite obstacles by adapting strategies while maintaining focus on the core goal even when initial approaches fail.

3.2 Agency Sliders

Building on Chen et al. [2024]'s representation engineering approach for controlling conversational LLM attributes, we propose developing "agency sliders" that enable real-time measurement and control of an LLM-based agent's agentic behavior. Chen et al. [2024] demonstrated that linear probes trained on residual stream activations can successfully identify and control internal representations of user attributes (age, gender, education, socioeconomic status), allowing users to dynamically adjust these representations through an intuitive dashboard interface.

Concretely, we propose adapting Chen et al. [2024] to target our three dimensions of agency where each slider sets a target scalar $s_d^\star \in [-1, 1]$ for dimension $d \in \{\text{rigidity}, \text{independence}, \text{persistence}\}$. At inference time, a control probe, trained on multi-step scenarios of each dimension, supplies a direction $v_{d,\ell}$ in the residual stream at selected layers ℓ (chosen by maximizing causal effect size on held-out intervention tasks), and the system applies representation translation $h_\ell \leftarrow h_\ell + \alpha_d v_{d,\ell}$ with α_d determined by a closed-loop controller that drives a companion *reader* probe's score toward $s_d^\star$. Units are calibrated so that $|\alpha_d| = 1$ corresponds to a one-standard-deviation shift of the reader logit on a calibration suite, resulting in slider semantics that are stable across models and deployments.

This approach offers significant advantages over output-based monitoring or prompt engineering for controlling agent behavior. First, white-box activation-level control is more robust against deceptive behaviors that could compromise output monitoring, as the probes operate on internal representations rather than final outputs that agents might strategically manipulate. Second, representation translation using control probe weight vectors enables granular, continuous adjustment of agency levels, akin to a volume dial rather than an on/off switch, allowing deployment contexts to calibrate agent autonomy precisely to their risk tolerance. For instance, agents deployed in critical infrastructure could operate with low preference rigidity and autonomous operation while maintaining high objective persistence for reliability, whereas research assistants might benefit from higher autonomous operation but lower preference rigidity to remain responsive to user guidance.

4 New Opportunities for Regulation

4.1 Mandated Testing of High-Risk Agents

Agency, at any level, is a risk in AI systems. By assessing their level of agency, we can better understand their benefits and potential harms [Cihon et al., 2025, Bengio et al., 2025a]. Since we cannot assume that these systems are safe without proactive testing [Kinniment et al., 2024], regulators should require pre-deployment testing tied to standardized agency levels, which would enable comparison between AI companies and provide accountability for system failures. Similar to stress tests in finance, or crash tests in automobiles, these evaluations would assess whether, under adversarial conditions, the system remains controllable and within its designated agency limits. Such

testing creates a safety baseline, ensuring developers demonstrate compliance before release, as recommended by leading AI safety institutes [UK AI Safety Institute, 2024].

4.2 Domain-dependent Agency Limits

By understanding the degree of agency at which misalignment occurs and assessing the inherent risks of different deployment contexts, policymakers can establish agency limits tailored to specific industries or applications [Kasirzadeh and Gabriel, 2025]. Unlike speed limits that can depend on the type and condition of the road, this proportionate oversight of risk should be guided by a uniform policy framework, such as the risk taxonomy of the EU AI Act [European Parliament and Council of the European Union, 2024]. Treating a system's degree of agency as a deliberate design choice corresponding to its capabilities and operational environment allows even highly capable systems to be constrained to lower levels of agency when deployed in sensitive contexts [Bengio et al., 2025b, Feng et al., 2025]. Without uniform metrics, policymakers and standards bodies tackling broad AI system regulation would be left with only broad categorization that may prove either overly stringent or permissive and difficult to implement.

4.3 Agency-based Insurance Frameworks

A quantified system of agency measurement would also provide infrastructure for the development of insurance frameworks for agentic systems. Strong insurance markets have long been a way of allowing market forces to guide an industry towards safety, as was seen in automobiles when insurers pushed for standards such as airbags [Albaum, 2005], and have been suggested as a pathway to regulate AI systems [Lior, 2025, Henson, 2025, Weil et al., 2024]. Insurance promises to both incentivize safer design and ensure compensation for victims of AI-related harms [Lior, 2025]. By attaching measurable levels of agency to quantifiable risk profiles, insurers could price premiums based not only on a system's agency level but on the characteristics contributing to a system's agency. High-agency systems operating in sensitive domains would therefore face higher premiums unless they implemented safety mechanisms, incentivizing firms to adopt lower-risk designs. Such an approach would shift the regulatory burden to market forces. Regulators could establish baseline requirements for risk disclosure and require a minimum level of insurance, while insurers would drive compliance through financial incentives.

4.4 Hard Agency Limits for the Prevention of Societal-Scale Risk

Agency metrics would enable regulators to set enforceable ceilings on how much agency developers can embed in their AI systems. Similarly to emission standards or nuclear material thresholds, policymakers could establish maximum allowed "agency levels," with mandatory hard stops beyond which further development or deployment is prohibited. Scholars have long argued that even small probabilities of catastrophic harm from runaway AI agents are intolerable and justify strict precautionary limits [Bengio, 2023]. In the same spirit with which international governance proposals have suggested treaties imposing global compute-based caps on the training of advanced AI models above agreed-upon thresholds, policymakers could impose specific limits of agency level in AI agents if they pose salient levels of existential risk [Miotti and Wasil, 2023, Raman et al., 2025, Ramiah et al., 2025]. By codifying these red lines, policymakers would reduce the likelihood of unauthorized action by agents and the emergence of uncontrollable agentic systems, as in Mitre and Predd [2025]. It would also provide clear and enforceable limits for AI companies that ensure innovation proceeds within safe and socially acceptable limits.

5 Conclusion

In this short paper, we proposed measuring and controlling the agency of LLM-based agents as a system property operationalized through preference rigidity, independent operation, and goal persistence via "agency sliders." This approach enables regulatory mechanisms including mandated testing, domain-specific limits, agency-based insurance, and hard ceilings for social-scale risks. Although our approach requires empirical validation across domains and agent architectures, we believe it represents a significant step toward making agency the direct target of technical intervention and regulatory governance, essential as these agents are increasingly integrated into critical systems.

References

Martin Albaum. Safety sells: Market forces and regulation in the development of airbags. Technical report, Insurance Institute for Highway Safety, 2005. URL https://www.iihs.org/media/186adabe-9ef4-479c-ad37-36b9f0e7fca1/Ka0wWQ/Albaum_Safety_Sells.pdf. Copyright © 2005 Martin Albaum and the Insurance Institute for Highway Safety.

Rauno Arike, Elizabeth Donoway, Henning Bartsch, and Marius Hobbhahn. Technical report: Evaluating goal drift in language model agents. *arXiv preprint arXiv:2505.02709*, 2025. URL https://arxiv.org/abs/2505.02709.

Ian Ayres and Jack M. Balkin. The law of AI is the law of risky agents without intentions. *University of Chicago Law Review Online*, 2024. doi: 10.2139/ssrn.4862025. URL https://ssrn.com/abstract=4862025.

Axel Backlund and Lukas Petersson. Vending-bench: A benchmark for long-term coherence of autonomous agents, 2025. URL https://arxiv.org/abs/2502.15840.

Yoshua Bengio. AI and catastrophic risk. *Journal of Democracy*, 34(4):111–121, oct 2023. doi: 10.1353/jod.2023.a907692. URL https://www.journalofdemocracy.org/articles/ai-and-catastrophic-risk/.

Yoshua Bengio, Michael Cohen, Damiano Fornasiere, Joumana Ghosn, Pietro Greiner, Matt MacDermott, Sören Mindermann, Adam Oberman, Jesse Richardson, Oliver Richardson, Marc-Antoine Rondeau, Pierre-Luc St-Charles, and David Williams-King. Superintelligent agents pose catastrophic risks: Can scientist AI offer a safer path? *arXiv preprint arXiv:2502.15657*, 2025a. URL https://arxiv.org/abs/2502.15657.

Yoshua Bengio, Sören Mindermann, Daniel Privitera, Tamay Besiroglu, Rishi Bommasani, Stephen Casper, Yejin Choi, Danielle Goldfarb, Hoda Heidari, Leila Khalatbari, Shayne Longpre, Vasilios Mavroudis, Mantas Mazeika, Kwan Yee Ng, Chinasa T. Okolo, Deborah Raji, Theodora Skeadas, Florian Tramèr, Bayo Adekanmbi, Paul Christiano, David Dalrymple, Thomas G. Dietterich, Edward Felten, Pascale Fung, Pierre-Olivier Gourinchas, Nick Jennings, Andreas Krause, Percy Liang, Teresa Ludermir, Vidushi Marda, Helen Margetts, John A. McDermid, Arvind Narayanan, Alondra Nelson, Alice Oh, Gopal Ramchurn, Stuart Russell, Marietje Schaake, Dawn Song, Alvaro Soto, Lee Tiedrich, Gaël Varoquaux, Andrew Yao, and Ya-Qin Zhang. International scientific report on the safety of advanced AI (interim report), 2025b. URL https://arxiv.org/abs/2412.05282.

Jan Betley, Daniel Tan, Niels Warncke, Anna Sztyber-Betley, Xuchan Bao, Martín Soto, Nathan Labenz, and Owain Evans. Emergent misalignment: Narrow finetuning can produce broadly misaligned LLMs. *arXiv preprint arXiv:2502.17424*, 2025. URL https://arxiv.org/abs/2502.17424.

Léo Boisvert, Megh Thakkar, Maxime Gasse, Massimo Caccia, Thibault Le Sellier De Chezelles, Quentin Cappart, Nicolas Chapados, Alexandre Lacoste, and Alexandre Drouin. Workarena++: Towards compositional planning and reasoning-based common knowledge work tasks, 2025. URL https://arxiv.org/abs/2407.05291.

Alan Chan, Rebecca Salganik, Alva Markelius, Chris Pang, Nitarshan Rajkumar, Dmitrii Krasheninnikov, Lauro Langosco, Zhonghao He, Yawen Duan, Micah Carroll, Michelle Lin, Alex Mayhew, Katherine Collins, Maryam Molamohammadi, John Burden, Wanru Zhao, Shalaleh Rismani, Konstantinos Voudouris, Umang Bhatt, Adrian Weller, David Krueger, and Tegan Maharaj. Harms from increasingly agentic algorithmic systems. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency*, FAccT '23, page 651–666, New York, NY, USA, 2023. Association for Computing Machinery. ISBN 9798400701924. doi: 10.1145/3593013.3594033. URL https://doi.org/10.1145/3593013.3594033.

Yida Chen, Aoyu Wu, Trevor DePodesta, Catherine Yeh, Kenneth Li, Nicholas Castillo Marin, Oam Patel, Jan Riecke, Shivam Raval, Olivia Seow, Martin Wattenberg, and Fernanda Viégas. Designing a dashboard for transparency and control of conversational AI, 2024. URL https://arxiv.org/abs/2406.07882.

Francois Chollet, Mike Knoop, Gregory Kamradt, and Bryan Landers. Arc prize 2024: Technical report, 2024.

François Chollet. On the measure of intelligence, 2019. URL https://arxiv.org/abs/1911.01547.

James Chua, Jan Betley, Mia Taylor, and Owain Evans. Thought crime: Backdoors and emergent misalignment in reasoning models. *arXiv preprint arXiv:2506.13206*, 2025. URL https://arxiv.org/abs/2506.13206.

Peter Cihon, Merlin Stein, Gagan Bansal, Sam Manning, and Kevin Xu. Measuring AI agent autonomy: Towards a scalable approach with code inspection. *arXiv preprint arXiv:2502.15212*, February 2025. URL https://arxiv.org/abs/2502.15212. NeurIPS SoLaR Workshop.

Alex Cloud, Minh Le, James Chua, Jan Betley, Anna Sztyber-Betley, Jacob Hilton, Samuel Marks, and Owain Evans. Subliminal learning: Language models transmit behavioral traits via hidden signals in data. *arXiv preprint arXiv:2507.14805*, 2025. URL https://arxiv.org/abs/2507.14805.

Josh Clymer, Hjalmar Wijk, and Beth Barnes. The rogue replication threat model, November 2024. URL https://metr.org/blog/2024-11-12-rogue-replication-threat-model/.

J. DiFrisco and R. Gawne. Biological agency: a concept without a research program. *Journal of Evolutionary Biology*, 38(2):143–156, 2025. doi: 10.1093/jeb/voae153. URL https://academic.oup.com/jeb/article/38/2/143/7920097.

Le Kim Dung. Current cases of AI misalignment and their implications for future risks. *Synthese*, 202(138):1–23, 2023. doi: 10.1007/s11229-023-04367-0. URL https://link.springer.com/article/10.1007/s11229-023-04367-0.

European Parliament and Council of the European Union. Regulation (EU) 2024/1689 laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). OJ L 2024/1689, July 2024. URL https://eur-lex.europa.eu/eli/reg/2024/1689/oj.

K. J. Kevin Feng, David W. McDonald, and Amy X. Zhang. Levels of autonomy for AI agents, 2025. URL https://arxiv.org/abs/2506.12469.

Luca Ferrero, editor. *The Routledge Handbook of Philosophy of Agency*. Routledge Handbooks in Philosophy. Routledge, Abingdon, Oxon; New York, NY, 2022. doi: 10.4324/9780429202131. URL https://philpapers.org/archive/FERTRH.pdf.

Stan Franklin and Art Graesser. Is it an agent, or just a program?: A taxonomy for autonomous agents. In Jörg P. Müller, Michael J. Wooldridge, and Nicholas R. Jennings, editors, *Intelligent Agents III: Agent Theories, Architectures, and Languages*, volume 1193 of *Lecture Notes in Computer Science*, pages 21–35. Springer, Berlin, Heidelberg, 1997. doi: 10.1007/BFb0013570. URL https://doi.org/10.1007/BFb0013570.

Renee Henson. Government-backed insurance for artificial intelligence technologies. *Georgia State University Law Review*, 41(3):559–?, 2025. doi: 10.2139/ssrn.5226107. University of Missouri School of Law Legal Studies Research Paper No. 2025-20.

Marcus Hutter. A theory of universal artificial intelligence based on algorithmic complexity, 2000. URL https://arxiv.org/abs/cs/0004001.

Atoosa Kasirzadeh and Iason Gabriel. Characterizing ai agents for alignment and governance, 2025. URL https://arxiv.org/abs/2504.21848.

Zachary Kenton, Ramana Kumar, Sebastian Farquhar, Jonathan Richens, Matt MacDermott, and Tom Everitt. Discovering agents, 2022. URL https://arxiv.org/abs/2208.08345.

Megan Kinniment, Lucas Jun Koba Sato, Haoxing Du, Brian Goodrich, Max Hasin, Lawrence Chan, Luke Harold Miles, Tao R. Lin, Hjalmar Wijk, Joel Burget, Aaron Ho, Elizabeth Barnes, and Paul Christiano. Evaluating language-model agents on realistic autonomous tasks, 2024. URL https://arxiv.org/abs/2312.11671.

Thomas Kwa, Ben West, Joel Becker, Amy Deng, Katharyn Garcia, Max Hasin, Sami Jawhar, Megan Kinniment, Nate Rush, Sydney Von Arx, Ryan Bloom, Thomas Broadley, Haoxing Du, Brian Goodrich, Nikola Jurkovic, Luke Harold Miles, Seraphina Nix, Tao Lin, Neev Parikh, David Rein, Lucas Jun Koba Sato, Hjalmar Wijk, Daniel M. Ziegler, Elizabeth Barnes, and Lawrence Chan. Measuring AI ability to complete long tasks. *arXiv preprint arXiv:2503.14499*, 2025. URL https://arxiv.org/abs/2503.14499.

Kenneth Li, Oam Patel, Fernanda Viégas, Hanspeter Pfister, and Martin Wattenberg. Inference-time intervention: Eliciting truthful answers from a language model, 2024. URL https://arxiv.org/abs/2306.03341.

Anat Lior. E/insuring the AI age: Empirical insights into artificial intelligence liability policies. *Connecticut Insurance Law Journal*, 31, 2025. doi: 10.2139/ssrn.5316376. URL https://ssrn.com/abstract=5316376. Forthcoming; 82 pp. Posted 25 June 2025.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Yang, Aohan Zeng, Zhengxiao Du, Chenhui Dong, and Jie Tang. Agentbench: Evaluating LLMs as agents. In *International Conference on Learning Representations (ICLR)*, 2024. URL https://arxiv.org/abs/2308.03688.

Aengus Lynch, Benjamin Wright, Caleb Larson, Kevin K. Troy, Stuart J. Ritchie, Sören Mindermann, Ethan Perez, Evan Hubinger, and Anthropic. Agentic misalignment: How LLMs could be insider threats, June 2025. URL https://www.anthropic.com/research/agentic-misalignment.

Alexander Meinke, Bronson Schoen, Jérémy Scheurer, Mikita Balesni, Rusheb Shah, and Marius Hobbhahn. Frontier models are capable of in-context scheming. *arXiv preprint arXiv:2412.04984*, 2025. URL https://arxiv.org/abs/2412.04984.

Andrea Miotti and Akash Wasil. An international treaty to implement a global compute cap for advanced artificial intelligence. *arXiv preprint arXiv:2311.10748*, 2023. URL https://arxiv.org/abs/2311.10748. Also available via SSRN.

Jim Mitre and Joel B. Predd. Artificial general intelligence's five hard national security problems. Perspective PE-A3691-4, RAND Corporation, Santa Monica, CA, February 2025. URL https://www.rand.org/pubs/perspectives/PEA3691-4.html.

James W. Moore. What is the sense of agency and why does it matter? *Frontiers in Psychology*, 7:1272, 2016. doi: 10.3389/fpsyg.2016.01272. URL https://pmc.ncbi.nlm.nih.gov/articles/PMC5002400/.

Meredith Ringel Morris, Jascha Sohl-dickstein, Noah Fiedel, Tris Warkentin, Allan Dafoe, Aleksandra Faust, Clement Farabet, and Shane Legg. Levels of agi for operationalizing progress on the path to agi, 2024. URL https://arxiv.org/abs/2311.02462.

Samir Okasha. The concept of agent in biology: Motivations and meanings. *Biological Theory*, 19:6–10, 2024. doi: 10.1007/s13752-023-00439-z. URL https://doi.org/10.1007/s13752-023-00439-z. Published online 28 June 2023.

Alexander Pan, Erik Jones, Meena Jagadeesan, and Jacob Steinhardt. Feedback loops with language models drive in-context reward hacking. *arXiv preprint arXiv:2402.06627*, 2024. URL https://arxiv.org/abs/2402.06627.

Henry Papadatos and Rachel Freedman. Linear probe penalties reduce LLM sycophancy, 2024. URL https://arxiv.org/abs/2412.00967.

Jairo Perez-Orosio and Agnieszka Wykowska. Adopting the intentional stance toward natural and artificial agents. *Philosophical Psychology*, 33(3):369–395, 2020. doi: 10.1080/09515089.2019.1688778. URL https://www.tandfonline.com/doi/abs/10.1080/09515089.2019.1688778.

Deepika Raman, Nada Madkour, Evan R. Murphy, Krystal Jackson, and Jessica Newman. Intolerable risk threshold recommendations for artificial intelligence, 2025. URL https://arxiv.org/abs/2503.05812.

Ananthi Al Ramiah, Raymond Koopmanschap, Josh Thorsteinson, Sadruddin Khan, Jim Zhou, Shafira Noh, Joep Meindertsma, and Farhan Shafiq. Toward a global regime for compute governance: Building the pause button, 2025. URL https://arxiv.org/abs/2506.20530.

Markus Schlosser. Agency. In Edward N. Zalta, editor, *The Stanford Encyclopedia of Philosophy*. Metaphysics Research Lab, Stanford University, Winter 2019 edition, 2019. URL https://plato.stanford.edu/archives/win2019/entries/agency/.

Holger Spamann and Jens Frankenreiter. Agency law primer. In *Corporations*, chapter 1.1.2. H2O Open Casebook, Harvard Law School Library, 2024. URL https://opencasebook.org/casebooks/261-corporations/resources/1.1.2-agency-law-primer/.

Alessandro Stolfo, Vidhisha Balachandran, Safoora Yousefi, Eric Horvitz, and Besmira Nushi. Improving instruction-following in language models through activation steering, 2025. URL https://arxiv.org/abs/2410.12877.

Helen Toner, John Bansemer, Kyle Crichton, Matt Burtell, Thomas Woodside, Anat Lior, Andrew J. Lohn, Ashwin Acharya, Beba Cibralic, Chris Painter, Cullen O'Keefe, Iason Gabriel, Kathleen Fisher, Ketan Ramakrishnan, Krystal Jackson, Noam Kolt, Rebecca Crootof, and Samrat Chatterjee. Through the chat window and into the real world: Preparing for AI agents. Workshop report, Center for Security and Emerging Technology, October 2024. URL https://cset.georgetown.edu/publication/through-the-chat-window-and-into-the-real-world-preparing-for-ai-agents/.

UK AI Safety Institute. AI Safety Institute Approach to Evaluations. Technical report, UK Government, 2024. URL https://www.gov.uk/government/publications/ai-safety-institute-approach-to-evaluations/ai-safety-institute-approach-to-evaluations.

Teun van der Weij, Felix Hofstätter, Ollie Jaffe, Samuel F. Brown, and Francis Rhys Ward. AI sandbagging: Language models can strategically underperform on evaluations. *arXiv preprint arXiv:2406.07358*, 2024. URL https://arxiv.org/abs/2406.07358.

Gabriel Weil, Matteo Pistillo, Suzanne Van Arsdale, Junichi Ikegami, Kensuke Onuma, Megumi Okawa, and Michael A. Osborne. Insuring emerging risks from ai. Technical report, Institute for Law & AI; Oxford Martin AI Governance Initiative, University of Oxford, November 14 2024. URL https://oms-www.files.svdcdn.com/production/downloads/Insuring%20emerging%20risks%20from%20AI%2014%20Nov%2024%20Final.pdf. Policy report.

Frank F. Xu, Yufan Song, Boxuan Li, Yuxuan Tang, Kritanjali Jain, Mengxue Bao, Zora Z. Wang, Xuhui Zhou, Zhitong Guo, Murong Cao, Mingyang Yang, Hao Yang Lu, Amaad Martin, Zhe Su, Leander Maben, Raj Mehta, Wayne Chi, Lawrence Jang, Yiqing Xie, Shuyan Zhou, and Graham Neubig. Theagentcompany: Benchmarking LLM agents on consequential real world tasks, 2024. URL https://arxiv.org/html/2412.14161v1. arXiv preprint arXiv:2412.14161.

Shunyu Yao, Noah Shinn, Pedram Razavi, and Karthik Narasimhan. τ-bench: A benchmark for tool-agent-user interaction in real-world domains. *arXiv preprint arXiv:2406.12045*, 2024. URL https://arxiv.org/abs/2406.12045.

Zhexin Zhang, Shiyao Cui, Yida Lu, Jingzhuo Zhou, Junxiao Yang, Hongning Wang, and Minlie Huang. Agent-safetybench: Evaluating the safety of LLM agents, 2025. URL https://arxiv.org/abs/2412.14470.

Andy Zou, Long Phan, Sarah Chen, James Campbell, Phillip Guo, Richard Ren, Alexander Pan, Xuwang Yin, Mantas Mazeika, Ann-Kathrin Dombrowski, Shashwat Goel, Nathaniel Li, Michael J. Byun, Zifan Wang, Alex Mallen, Steven Basart, Sanmi Koyejo, Dawn Song, Matt Fredrikson, J. Zico Kolter, and Dan Hendrycks. Representation engineering: A top-down approach to AI transparency, 2025. URL https://arxiv.org/abs/2310.01405.