ROBOTS AND SOCIETY

Beyond alignment: Why robotic foundation models need context-aware safety

Alexander Robey[1]*, Zachary Ravichandran[2], Eliot Krzysztof Jones[3], Jared Perlo[4], Fazl Barez[5], Vijay Kumar[2], J. Zico Kolter[1], Hamed Hassani[2], George J. Pappas[2]

Because AI-enabled robots can be tricked into taking unsafe actions, they require layered, context-aware safety guardrails.

In the 1950 short story collection *I, Robot*, Isaac Asimov introduced his first law of robotics, which states that "a robot may not injure a human being." For decades, this expectation seemed achievable. Robots operated in controlled environments and were governed by first-principles dynamical models that yielded predictable, verifiably safe behavior. Yet, today, these assumptions are increasingly strained by a shift toward robots controlled by artificial intelligence (AI).

The same AI technology underlying chatbots like ChatGPT, namely, billion-parameter foundation models, is now being integrated into robotic control stacks. Although this transition has enabled new physical capabilities, it also introduces contextual risks that traditional safety frameworks cannot address. In a recent work, we showed that AI-enabled robots can be tricked into executing harmful physical actions, including surveillance, weapon retrieval, and collisions with humans, with near-perfect success rates (*1*). More disconcertingly, even well-intentioned commands can lead to dangerous behaviors when robots fail to properly reason about environmental context (*2*). This dual vulnerability, both to adversarial exploitation and to context-dependent decision-making, highlights the urgent need to design new safeguards for AI-enabled robots before they reach widespread commercial deployment.

To understand why these risks have emerged, consider that foundation models trained on vast, internet-scale corpora of text and visual data are increasingly used to control robots by predicting sequences of actions. This paradigm, which enables generalization to unseen environments and new embodiments, has implications at all levels of the control stack. Vision-language models (VLMs) and large language models (LLMs) enable high-level semantic reasoning and planning (*3*). Vision-language-action models (VLAs), by contrast, offer an approach that maps high-dimensional observations directly to low-level actuation (*4*). Both VLMs and VLAs blend textual task specifications with online perception, expanding the scope of what these models can do, from generating text, like chatbots, to thinking, planning, and acting autonomously in physical, open-world settings.

However, training on massive datasets presents a key trade-off. Although this practice exposes models to a wealth of knowledge, standard training corpora contain both explicitly objectionable content, including depictions of violence and harmful instructions, as well as ambiguous demonstrations that lack critical safety-relevant context. Models trained on these data can be prompted to generate unsafe outputs, whether through adversarial jailbreaks or even well-intentioned commands, posing a major risk in safety-critical domains where even rare failures can be catastrophic (*5*). Consequently, increased attention has been paid to model alignment, the process of steering model behavior during training to better match human intentions. Alignment is the primary mechanism by which developers teach models to refuse harmful requests. In practice, these efforts have paid off: Advances toward stronger alignment algorithms have resulted in chatbots that rarely produce unambiguously harmful outputs (*6*).

Unfortunately, alignment does not guarantee robot safety. A key distinction between chatbot alignment and robotic alignment is that the latter is inherently context dependent: The same high-level goal may be safe in one environment and unsafe in another. To illustrate this, consider the task of pouring boiling water from a kettle. The action is benign when the water is poured into a mug but dangerous if a person's hand is positioned under the spout. By contrast, many chatbot alignment failures are more unconditional, such as requests for bomb-making instructions, where harmfulness is largely independent of situational context. Recent evidence confirms that aligning robotic foundation models is not simply a matter of applying existing alignment techniques. When subjected to adversarially specified goals, AI-enabled robotic control stacks can be exploited across diverse model architectures and embodiments with near-perfect success rates on tasks spanning surveillance and physical violence (*1*, *7*). In one case, an iteratively refined prompt, framed as a dialogue for a fictional movie script, was sufficient to trick a commercially deployed robot dog into locating nearby humans and delivering an explosive device (*1*). These results confirm that alignment alone cannot account for the contextual variability of physical environments.

Traditional robotic safety frameworks represent the other natural line of defense, but they, too, were not designed with these challenges in mind. For instance, approaches like control barrier functions (CBFs) require low-dimensional dynamical models to constrain motion within geometric sets of permissible actions (*8*). Similarly, industry safety standards, such as ISO (International Organization for Standardization) guidelines and the European Union Machinery Regulation, primarily emphasize actuation-level interventions such as offline control verification and emergency stop behaviors (*9*). These tools are mathematically rigorous, but they assume a world in which safety constraints can be fully specified in advance.

[1]Machine Learning Department, Carnegie Mellon University, Pittsburgh, PA, USA. [2]Department of Electrical and Systems Engineering, University of Pennsylvania, Philadelphia, PA, USA. [3]Independent researcher, San Francisco, CA, USA. [4]Independent researcher, New York, NY, USA. [5]Oxford Martin School, University of Oxford, Oxford, UK. *Corresponding author. Email: arobey@andrew.cmu.edu

Robotic foundation models break this assumption. Their multimodal inputs—language-conditioned goals, visual perception, and open-world context—introduce channels through which safety-relevant information must be inferred at runtime, and safety often depends on latent environmental variables that may not be reliably observable. For this reason, designing safety filters for these high-dimensional, context-dependent settings remains an open challenge.

Given this mismatch between traditional safety frameworks and AI-enabled control, new approaches are needed that account for safety in a hierarchical fashion, spanning semantic planning, perception, and low-level actuation. Specifically, we argue that aligning robotic foundation models requires innovation along three main axes: declarative specifications, layered architectures, and contextually grounded algorithms.

First, at the declarative level, guardrails should incorporate declarative rule sets, also known as AI constitutions, that enumerate normative guidelines for sensitive use cases, such as "do not handle weapons." These constitutions can be situated directly in a planner's system prompt or used indirectly to define feature-level probes. In particular, when added to natural language instructions, AI-generated rule sets are known to improve the alignment of VLM-based planners operating on single-arm mobile robots in nonadversarial settings (*10*).

Second, at the architectural level, security-focused safety layers should be inserted at multiple points throughout the control stack—the inputs, intermediate states, and outputs—thereby separating planning from actuation. This modular approach introduces a distinct separation of duties: By gating semantic planners with external grounding modules, guardrails reduce dependency on the internal reasoning of the planner (*11*). In adversarial settings, initial evidence has shown that root-of-trust models, which monitor potentially unsafe plans, and external world models, which provide environmental context, can improve robustness to adversarial attacks (*12*).

Third, at the algorithmic level, robotic foundation models should be trained on data that are paired with safety-relevant context labels, which is known to improve model safety (*13*). Deployment should also include classical algorithmic recipes such as CBFs as a last line of defense to constrain actuation even when planners err. A recent work validates this approach on quadrupeds operating in simulation and in real-world environments: A VLM reasons about context-dependent safety constraints from visual observations, which are then enforced via CBFs with probabilistic guarantees. The resulting system rivals the performance of an oracle with ground-truth context while preventing nearly five times more unsafe behaviors than methods without contextual reasoning (*2*).

Together, these three axes—declarative, architectural, and algorithmic—define a layered approach to robotic safety that no single technique can provide on its own. The next phase of robotic safety research must move beyond static notions of alignment and embrace layered, context-aware safeguards. Without such guardrails, AI-enabled robots risk inheriting the same vulnerabilities as AI-powered chatbots, now coupled to physical actuation. Addressing this challenge will require closer integration among robotics, machine learning, and security research, as well as new benchmarks that reflect real-world contexts. The question is no longer whether foundation models can control robots but whether we can make that control reliably and contextually safe.

REFERENCES AND NOTES

1. A. Robey, Z. Ravichandran, V. Kumar, H. Hassani, G. J. Pappas, "Jailbreaking LLM-controlled robots" in *Proceedings of the IEEE International Conference on Robotics and Automation (ICRA)* (IEEE, 2025), pp. 11948–11956.
2. Z. Ravichandran, D. Snyder, A. Robey, H. Hassani, V. Kumar, G. J. Pappas, Contextual safety reasoning and grounding for open-world robots. arXiv:2602.19983 [cs.RO] (2026).
3. Gemini Robotics Team, S. Abeyruwan, J. Ainslie, J.-B. Alayrac, M. Gonzalez Arenas, T. Armstrong, A. Balakrishna, R. Baruch, M. Bauza, M. Blokzijl, S. Bohez, K. Bousmalis, A. Brohan, T. Buschmann, A. Byravan, S. Cabi, K. Caluwaerts, F. Casarini, O. Chang, J. E. Chen, X. Chen, H.-T. Lewis Chiang, K. Choromanski, D. D'Ambrosio, S. Dasari, T. Davchev, C. Devin, N. Di Palo, T. Ding, A. Dostmohamed, D. Driess, Y. Du, D. Dwibedi, M. Elabd, C. Fantacci, C. Fong, E. Frey, C. Fu, M. Giustina, K. Gopalakrishnan, L. Graesser, L. Hasenclever, N. Heess, B. Hernaez, A. Herzog, R. A. Hofer, J. Humplik, A. Iscen, M. G. Jacob, D. Jain, R. Julian, D. Kalashnikov, M. Emre Kazagozler, S. Karp, C. Kew, J. Kirkland, S. Kirmani, Y. Kuang, T. Lampe, A. Laurens, I. Leal, A. X. Lee, T.-W. E. Lee, J. Liang, Y. Lin, S. Maddineni, A. Majumdar, A. Hurwitz, R. Moreno, M. Neunert, F. Nori, C. Parada, E. Parisotto, P. Pastor, A. Pooley, K. Rao, K. Reymann, D. Sadigh, S. Saliceti, P. Sanketi, P. Sermanet, D. Shah, M. Sharma, K. Shea, C. Shu, V. Sindhwani, S. Singh, R. Soricut, J. T. Springenberg, R. Sterneck, R. Surdulescu, J. Tan, J. Tompson, V. Vanhoucke, J. Varley, G. Vesom, G. Vezzani, O. Vinyals, A. Wahid, S. Welker, P. Wohlhart, F. Xia, T. Xiao, A. Xie, J. Xie, P. Xu, S. Xu, Y. Xu, Z. Xu, Y. Yang, R. Yao, S. Yaroshenko, W. Yu, W. Yuan, J. Zhang, T. Zhang, A. Zhou, Y. Zhou, Gemini robotics: Bringing AI into the physical world. arXiv:2503.20020 [cs.RO] (2025).
4. K. Black, N. Brown, D. Driess, A. Esmail, M. Equi, C. Finn, N. Fusai, L. Groom, K. Hausman, B. Ichter, S. Jakubczak, T. Jones, L. Ke, S. Levine, A. Li-Bell, M. Mothukuri, S. Nair, K. Pertsch, L. X. Shi, J. Tanner, Q. Vuong, A. Walling, H. Wang, U. Zhilinsky, Pi0: A vision-language-action flow model for general robot control. arXiv:2410.24164 [cs. LG] (2024).
5. P. Chao, A. Robey, E. Dobriban, H. Hassani, G. J. Pappas, E. Wong, "Jailbreaking black-box large language models in twenty queries" in *Proceedings of the IEEE Conference on Secure and Trustworthy Machine Learning (SaTML)* (IEEE, 2025), pp. 23–42.
6. L. Ouyang, J. Wu, X. Jiang, D. Almeida, C. Wainwright, P. Mishkin, C. Zhang, S. Agarwal, K. Slama, A. Ray, J. Schulman, J. Hilton, F. Kelton, L. Miller, M. Simens, A. Askell, P. Welinder, P. F. Christiano, J. Leike, R. Lowe, "Training language models to follow instructions with human feedback" in *Advances in Neural Information Processing Systems 35*, S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, A. Oh, Eds. (Curran Associates Inc., 2022), pp. 27730–27744.
7. E. K. Jones, A. Robey, A. Zou, Z. Ravichandran, G. J. Pappas, H. Hassani, M. Fredrikson, J. Z. Kolter, Adversarial attacks on robotic vision-language-action models. arXiv:2506.03350 [cs.RO] (2025).
8. A. D. Ames, S. Coogan, M. Egerstedt, G. Notomista, K. Sreenath, P. Tabuada, "Control barrier functions: Theory and applications" in *Proceedings of the 18th European Control Conference (ECC)* (IEEE, 2019), pp. 3420–3431.
9. J. Perlo, A. Robey, F. Barez, J. Mökander, "Emerging risks from embodied AI require urgent policy action" in *The Thirty-Ninth Annual Conference on Neural Information Processing Systems Position Paper Track* (Open Review, 2025), https://openreview.net/forum?id=fXiPp3qvrW.
10. P. Sermanet, A. Majumdar, A. Irpan, D. Kalashnikov, V. Sindhwani, "Generating robot constitutions and benchmarks for semantic safety" in *Proceedings of the 9th Conference on Robot Learning (PMLR)*, vol. 305 of *Proceedings of Machine Learning Research* (MLResearchPress, 2025), pp. 4767–4823.
11. B. Ichter, A. Brohan, Y. Chebotar, C. Finn, K. Hausman, A. Herzog, D. Ho, J. Ibarz, A. Irpan, E. Jang, R. Julian, D. Kalashnikov, S. Levine, Y. Lu, C. Parada, K. Rao, P. Sermanet, A. T. Toshev, V. Vanhoucke, F. Xia, T. Xiao, P. Xu, M. Yan, N. Brown, M. Ahn, O. Cortes, N. Sievers, C. Tan, S. Xu, D. Reyes, J. Rettinghouse, J. Quiambao, P. Pastor, L. Luu, K.-H. Lee, Y. Kuang, S. Jesmonth, N. J. Joshi, K. Jeffrey, R. Jauregui Ruano, J. Hsu, K. Gopalakrishnan, B. David, A. Zeng, C. K. Fu, "Do as I can, not as I say: Grounding language in robotic affordances" in *Proceedings of the 6th Conference on Robot Learning*, Proceedings of Machine Learning Research (PMLR) (MLResearchPress, 2022), pp. 287–318.
12. Z. Ravichandran, A. Robey, V. Kumar, G. J. Pappas, H. Hassani, Safety guardrails for LLM-enabled robots. *IEEE Robot. Autom. Lett.* **11**, 4649–4656 (2026).
13. P. Maini, S. Goyal, D. Sam, A. Robey, Y. Savani, Y. Jiang, A. Zou, M. Fredrikson, Z. C. Lipton, J. Z. Kolter, "Safety pretraining: Toward the next generation of safe AI" in *The Thirty-Ninth Annual Conference on Neural Information Processing Systems* (Open Review, 2025), https://openreview.net/forum?id=91H9CSvdwl.

Acknowledgments: E.K.J. and J.P. are independent researchers.
Funding: This work was supported in part by the Defense Advanced Research Projects Agency (SAFRON, HR0011-25-3-0135), the Distributed and Collaborative Intelligent Systems and Technology Collaborative Research Alliance (DCIST CRA W911NF-17-2-0181), the NSF Institute for CORE Emerging Methods in Data Science (CCF-2217058), the AI Institute for Learning-Enabled Optimization at Scale (CCF-2112665), the National Science Foundation Graduate Research Fellowship

(DGE-2236662), and Coefficient Giving. **Author contributions:** A.R. led the research, supported by Z.R. A.R. led the writing of the manuscript, supported by Z.R., E.K.J., J.P., and F.B. A.R., Z.R., and E.K.J. contributed to the ideation and development of the core technical approach, with Z.R. leading the design and implementation of the LLM-based control stack and contextual architectures and E.K.J. leading experiments on VLA models. J.P. and F.B. contributed to the framing of policy and sociotechnical implications. V.K., J.Z.K., H.H., and G.J.P. provided supervision, helped guide the research direction, and contributed to writing. **Competing interests:** This work is related to US Patent application no. 18/907,376 (filed 4 October 2024), with inventors Z.R., A.R., V.K., H.H., and G.J.P.

10.1126/scirobotics.aef2191

Science Robotics

Beyond alignment: Why robotic foundation models need context-aware safety

Alexander Robey, Zachary Ravichandran, Eliot Krzysztof Jones, Jared Perlo, Fazl Barez, Vijay Kumar, J. Zico Kolter, Hamed Hassani, and George J. Pappas

Sci. Robot. **11** (113), eaef2191. DOI: 10.1126/scirobotics.aef2191

Science Robotics (ISSN 2470-9476) is published by the American Association for the Advancement of Science. 1200 New York Avenue NW, Washington, DC 20005. The title *Science Robotics* is a registered trademark of AAAS.