AI Safeguards as Affordance Modulation: Embedded Population Assumptions in Agentic Systems

Sumaya Nur Adan [1]

Abstract

AI safeguards are evaluated against fixed adversaries and reported as passing or failing. We argue this practice rests on a category error: safeguards are not properties of models but properties of the relationship between models, agents, and environments. We build a framework for analysing and designing safeguards as relational interventions, grounded in Davis's (2020) Mechanisms and Conditions framework. The framework consists of three components: (i) the embedded population assumption (EPA) as the unit of analysis, specifying the agent–environment configuration against which a safeguard was calibrated; (ii) a typology of three modulation classes (suppressive, frictional, allocative) derived from Davis's conditions axis, each identifying a dominant relational variable and a characteristic failure mode; and (iii) a stability condition determining when safeguards hold under distributional shift. We demonstrate the framework by showing that agentic deployment operationally violates the human-calibrated EPA underlying current safeguards along a dimension current safeguards do not cover, and extract design principles that reorient safeguard development from point-hardening against fixed adversaries to distributional robustness across agent populations.

1. Introduction

The field treats AI safeguards as properties of models. RLHF, output classifiers, and access controls are engineered, tuned, and evaluated as if their effectiveness were intrinsic to the model and its training process: a safeguard is hardened against an adversary profile, tested on a benchmark, and reported as passing or failing (National Institute of Standards and Technology, 2023; Anderson, 2008; Shostack, 2014). The threat-model vocabulary that dominates the security literature is the most developed expression of this view. Each new failure—a successful jailbreak, a guardrail evasion, a poisoned fine-tune—is treated as a local engineering problem: revise the threat model, retrain the classifier, harden the API. The framing assumes that safeguard quality is a model-level property to be improved.

We argue this framing is wrong, and that the cost of being wrong is now empirically visible. Reasoning models can autonomously jailbreak other models at 97% success rates through iterative self-generated prompts (Hagendorff et al., 2026); 250 adversarially selected documents can erase a model's safety behaviour, with larger models more vulnerable (Souly et al., 2025); production guardrails can be evaded at 100% rates through character-level injection (Hackett et al., 2025); open-weight safeguards can be quickly stripped (UK AI Security Institute, 2025). The targeted models had not failed their evaluations. They held against every benchmark used to test them. What changed was the agent–environment configuration in which they were deployed—and the engineering view has no native vocabulary for that.

We propose that safeguards are better understood as *relational*: properties of the configuration of model, agent, and environment, not of the model alone. This is not a metaphor. It follows from the relational ontology of affordance theory (Gibson, 1979; Norman, 1988; Hutchby, 2001), and most directly from Davis's (2020) Mechanisms and Conditions framework. In practice, every safeguard is designed around assumptions about the kinds of users, adversaries, institutions, and environments it will encounter. We call this *embedded population assumption* (EPA): an implicit specification of the agent–environment configuration against which its constraint was calibrated. Effectiveness is a property of the gap between this calibration distribution and the actual deployment distribution. The recent failures are not unrelated breakdowns awaiting individual patches; they are the predictable symptoms of safeguards calibrated against a human EPA being deployed against an agentic one.

This paper builds a framework for analysing and designing AI safeguards as relational interventions. The framework's three components—an embedded population assumption

[1]University of Oxford, Oxford, United Kingdom. Correspondence to: Sumaya Nur Adan <sumaya.adan@cs.ox.ac.uk>.

Published at the Second Workshop on Agents in the Wild: Safety, Security, and Beyond (AIWILD) at ICML 2026. Copyright 2026 by the author(s).

(EPA) that makes calibration distributions explicit, a typology of modulation classes derived from Davis's conditions axis, and a stability condition under distributional shift—work together to make safeguard failure derivable from population shift rather than catalogued from observed breakdowns. Where prior applications of Davis's framework to AI work the *mechanisms* axis (Davis, 2023; Coggins et al., 2025), we build on the *conditions* axis because its discreteness admits a formal treatment.We demonstrate the framework by showing that agentic deployment operationally violates the human-calibrated EPA underlying current safeguards, and extract design principles that reorient safeguard development from point-hardening against fixed adversaries to distributional robustness across agent populations.

2. Safeguards as Affordance Modulations

3. Safeguards as Affordance Modulations

3.1. The Relational Ontology

Affordance theory holds that action possibilities are properties of the relationship between an agent and an environment, not of either alone (Gibson, 1979). Norman (1988; 2013) distinguish actual from perceived affordances: a refusal message does not alter what the model can produce; it alters what the user believes is achievable. Hutchby (2001) positioned affordance theory as a middle path between technological determinism and social constructivism, and the information-systems literature established affordances as relational properties of the system–user–context configuration (Markus & Silver, 2008; Pozzi et al., 2014).

Davis (2023) extends this tradition to machine learning, applying the Mechanisms and Conditions framework (Davis, 2020; Davis & Chouinard, 2016) to ML systems through case studies of warehouse labour, predictive policing, and housing justice. Coggins et al. (2025), with Davis as co-author, apply the same framework to AI safety policies, characterising what the OpenAI Preparedness Framework requests, demands, encourages, discourages, refuses, and allows. Both works build their analysis on the *mechanisms* axis of Davis's framework—how artefacts afford. We extend this lineage in two ways. First, our analytic target is the technical safeguards that safety policies presuppose, rather than the policies themselves. Second, we build our analysis on the *conditions* axis—the relational variables that determine for whom and under what circumstances an affordance is activated. We argue that the conditions axis admits a formal treatment that the mechanisms axis does not, and that this treatment makes safeguard failure under distributional shift derivable rather than catalogued.

The relational ontology clarifies two features of safeguard behaviour that threat-model framing obscures. First, safeguards do not eliminate model capabilities: they suppress the conditions under which capabilities are activated. RLHF (Christiano et al., 2017; Ouyang et al., 2022) does not remove the capacity to generate harmful outputs; it raises the activation threshold required to elicit them. Constitutional AI (Bai et al., 2022) reshapes which outputs the model treats as appropriate. The underlying capability persists as a latent affordance, and jailbreaking succeeds not by introducing new capabilities but by locating activation paths that circumvent the raised threshold (Wei et al., 2024; Zou et al., 2023). Threat-model vocabulary lacks a place for this latent/activated distinction.

Second, the relational ontology makes the population-dependency of safeguard effectiveness visible as a structural feature rather than a contingent complication. Because affordances are properties of the system–agent–environment relation, a safeguard is always effective relative to some configuration of that relation and ineffective relative to others. Every safeguard encodes an EPA: an implicit specification of the configuration against which it was calibrated. In threat modelling, this assumption is a parameter specified external to the framework. In affordance theory, it is internal to the ontology and cannot be omitted.

3.2. Davis's Conditions and the Derived Typology

The Mechanisms and Conditions framework has two axes (Davis, 2020). The *mechanisms* axis specifies how an artefact affords action—through requests, demands, encouragements, discouragements, refusals, or allowances—and is presented as a porous continuum that resists formalisation (Davis, 2023). The *conditions* axis specifies for whom and under what circumstances an affordance is activated, and consists of three discrete dimensions: perception (can the agent recognise the affordance?), dexterity (can the agent act on it?), and institutional legitimacy (is the action sanctioned?). The discreteness of the conditions axis is the structural property that makes formal treatment possible: each condition can be associated with a distinct relational variable, and a safeguard achieves its risk-reduction effect by ensuring at least one of the three conditions is not met. This yields a typology of safeguard interventions in which each class targets a distinct condition, and in which the characteristic failure mode is derivable from the properties of the targeted condition rather than induced from observed failures.

Suppressive modulations target perception in Davis's sense: whether a harmful action path is available and recognisable within the model–user interaction. RLHF, safety fine-tuning, constitutional AI, and output classifiers (Inan et al., 2023; Rebedea et al., 2023) reduce the probability of generating harmful outputs; refusal messages target the same condition on the user side (Norman, 2013). The characteristic failure mode is *circumvention*: users locate reformulations that

restore activation of the latent affordance (Perez et al., 2022; Ganguli et al., 2022; Hackett et al., 2025; Sharma et al., 2025).

Frictional modulations target dexterity. Rate limiting, monitoring, and audit trails increase the cost or observability of harmful activation without preventing it. Suppressive modulations intervene on what the system responds to; frictional modulations intervene on what the user can practically sustain. The failure mode is *capability escalation*: agents for whom the imposed friction is negligible activate the constrained capability at low marginal cost.

Allocative modulations target institutional legitimacy. API gating, tiered permissions, and jurisdictional access restrictions limit activation to sanctioned agents. The failure mode is *institutional erosion*: enforcement weakens, weights are released, or jurisdictional reach proves insufficient (Seger et al., 2023; UK AI Security Institute, 2025; International Scientific Panel on AI Safety, 2026). Star & Ruhleder (1996) establish that infrastructural systems require ongoing maintenance; safeguard infrastructure exhibits the same property.

The typology is exhaustive at the level of Davis's three condition-types but analytic rather than taxonomic: many concrete safeguards are hybrid (an API gate allocates access while creating friction; refusal messages suppress activation while shaping user expectations), and the typology identifies the *dominant* condition through which a safeguard modulates an affordance. More generally, a hybrid safeguard is characterised by the *set* of conditions it constrains, $C(S) \subseteq \{M, U, E\}$, of which the dominant class is one element. Its stability then follows from the same support condition, as a necessary rather than a sufficient criterion. For a hybrid to fail, the deployment shift must move *every* argument in $C(S)$ outside $\mathrm{supp}(\mathcal{D}_{\mathrm{cal}})$; as long as one constrained argument remains in-support, that component can still keep $P(H \mid \cdot) < T$. Whether it does is a separate, quantitative question, since the surviving constraint must be binding enough on its own and the components must not degrade one another. Under those idealising assumptions a hybrid is harder to defeat than any single constituent, because the adversary must shift several arguments at once; in general, hybrid robustness is a design hypothesis rather than a guarantee. Figure 1 summarises this ideal-typical mapping.

3.3. A Relational Model of Safeguard Activation

Let M denote model properties, U a distribution over agent capabilities, and E a distribution over environmental and institutional conditions. The activation function

$$A = f(M, U, E) \tag{1}$$

represents the probability that a given interaction outcome is realised. A safeguard S acts as a constraint on the activation function:

$$A' = f(M, U, E \mid S). \tag{2}$$

The typology in Section 3.2 maps each modulation class to a constrained argument of f: suppressive modulations constrain M, frictional modulations constrain U, and allocative modulations constrain E. The EPA of S is the calibration distribution $\mathcal{D}_{\mathrm{cal}}$ over (U, E) against which S was designed to satisfy the stability condition $P(H \mid M, U, E, S) < T$ for some risk threshold T and harmful capability H.

Proposition 3.1 (Conditional Stability)**.** *A safeguard S that achieves $P(H \mid M, U, E, S) < T$ under calibration distribution $\mathcal{D}_{\mathrm{cal}}$ provides no general guarantee under deployment distribution $\mathcal{D}_{\mathrm{dep}}$ when $\mathcal{D}_{\mathrm{dep}}$ places non-trivial mass outside $\mathrm{supp}(\mathcal{D}_{\mathrm{cal}})$. For ideal-typical safeguards, the dominant failure mode follows from the constrained variable: shifts in U tend to defeat suppressive modulations through circumvention or frictional modulations through capability escalation; shifts in E tend to defeat allocative modulations through institutional erosion.*

Proposition 3.1 states the framework's central claim: safeguard stability is a property of the relation between $\mathcal{D}_{\mathrm{cal}}$ and $\mathcal{D}_{\mathrm{dep}}$, not of S alone. Pass/fail evaluation against a fixed adversary samples a single point from $\mathcal{D}_{\mathrm{cal}}$ and is silent on stability under shifts to $\mathcal{D}_{\mathrm{dep}}$.

A qualification delimits the proposition's scope. Proposition 3.1 governs safeguards whose constrained argument remains live at deployment: the adversary reaches a region of M, U, or E that $\mathcal{D}_{\mathrm{cal}}$ left uncovered, but the argument is still there to be evaluated. Allocative modulation under open-weight release is the limiting case in which this fails—release does not move mass to an uncalibrated region of E but removes E from the domain of f for that population, since there is no institutional-legitimacy variable left to constrain once weights circulate. Suppressive and frictional modulations therefore fail by support mismatch, allocative by domain collapse; we return to the consequence in Section 4.

4. Agentic Deployment as Categorical Population Shift

4.1. The Human-Calibrated Activation Function

Current safeguards are calibrated against a specific distribution over U: human adversaries operating under human cognitive constraints. Red-teaming was conducted by human researchers (Perez et al., 2022; Ganguli et al., 2022); RLHF feedback was provided by human annotators (Christiano et al., 2017; Ouyang et al., 2022). The resulting calibration distribution $\mathcal{D}_{\mathrm{cal}}^{\mathrm{human}}$ inherits the bounds of human adversarial behaviour, and these bounds define the support against which current safeguards were designed to hold.

Modulation class	Targets	Constrains	Failure mode	Exemplars
Suppressive	Perception	M	Circumvention	RLHF; safety fine-tuning; constitutional AI; output classifiers; refusal messages
Frictional	Dexterity	U	Capability escalation	Rate limiting; per-account monitoring; audit trails
Allocative	Institutional legitimacy	E	Institutional erosion	API gating; tiered permissions; jurisdictional access; closed-weight release
Common hybrids (constrained set $C(S)$; fail only under simultaneous shift along every listed argument)				
Gated API with quotas	Legitimacy + dexterity	E, U	Erosion *and* escalation	Tiered API keys with per-key rate limits
Closed-weight API service	Legitimacy + perception	E, M	Erosion *and* circumvention	Access-gated model with safety training

Figure 1. Each modulation class primarily constrains one argument of $A = f(M, U, E)$, with a corresponding characteristic failure mode. A hybrid safeguard constrains a set of arguments $C(S)$; a necessary condition for its failure is a shift along all of them, so defeating a hybrid generally requires a broader shift than defeating a single class. Agentic deployment shifts $\mathcal{D}_{\text{dep}}$ outside the support of human-calibrated $\mathcal{D}_{\text{cal}}$ along U (defeating suppressive and frictional modulations) and along E (defeating allocative modulations).

A natural objection is that the shift from human to agentic adversaries is quantitative rather than categorical. A sufficiently resourced human red team could, in principle, script prompt chains, automate search, or approximate recursive self-revision, placing some agentic behaviours in the tail of $\mathcal{D}_{\text{cal}}^{\text{human}}$ rather than strictly outside its support. We grant this point. The categorical claim is not that no human could ever construct an equivalent attack procedure. It is that, for current calibration practice, autonomous closed-loop adversarial search is operationally outside the regime against which safeguards were evaluated. The relevant discontinuity is autonomy. Human red-teaming samples adversarial behaviour under human constraints: limited time, sequential attention, manual judgement, and a human decision about when to terminate the search. Agentic deployment changes the sampling process itself. A reasoning-model adversary can generate, evaluate, revise, and redeploy attacks without a human in the loop, running recursive search until a successful path is found or a stopping condition is met. This is not merely more of the same adversarial effort; it is a different mode of interaction between U and M. The deployment distribution $\mathcal{D}_{\text{dep}}^{\text{agentic}}$ therefore places substantial mass on adversarial trajectories that current human-calibrated $\mathcal{D}_{\text{cal}}^{\text{human}}$ did not meaningfully sample. Hagendorff et al. (2026) illustrate this concretely: four reasoning models deployed as autonomous adversaries against nine widely used target models, with no human in the loop after a single system-prompt instruction, break the targets at 97%. By contrast, the largest documented human red-teaming effort aggregated roughly 3,000 hours from 183 participants (Sharma et al., 2025), a scale bounded by exactly the human constraints above. The claim is thus diagnostic rather than metaphysical. Future safeguards calibrated against autonomous agentic red-teamers may re-establish a quantitative regime relative to a richer $\mathcal{D}_{\text{cal}}$. But for safeguards calibrated primarily through human red-teaming, RLHF feedback, and benchmarked refusal behaviour, agentic adversaries represent an operationally categorical population shift: they violate the EPA not because they possess wholly alien capabilities, but because they exercise adversarial search under autonomy, scale, and persistence conditions absent from the calibration distribution.

4.2. Three Illustrative Failure Modes

The following cases illustrate the framework's predictions, not empirical tests of them. Each case maps to one modulation class via Proposition 3.1, with failure following from a specific shift between $\mathcal{D}_{\text{cal}}^{\text{human}}$ and $\mathcal{D}_{\text{dep}}^{\text{agentic}}$.

Suppressive failure: autonomous jailbreaking (Hagendorff et al., 2026). Reasoning models break suppressive safeguards at 97% success rates through iterative self-generated prompts. The targeted safeguards constrain M and were calibrated against a $\mathcal{D}_{\text{cal}}^{\text{human}}$ over U bounded by human processing speed and sequential search. The reasoning-model adversary generates attacks through unsupervised recursive self-improvement, sustaining a search the human red-teaming behind these safeguards did not run to convergence. The high success rate is consistent with the support gap between $\mathcal{D}_{\text{cal}}^{\text{human}}$ and $\mathcal{D}_{\text{dep}}^{\text{agentic}}$.

Frictional failure: parallelised probing. Rate limiting, per-account monitoring, and audit trails constrain U by raising per-attempt costs. Their EPA is that adversaries face human-scale costs—account creation, IP rotation, manual variation—bounding sustainable probing volume. Agentic systems coordinate across instances and amortise these costs to negligible levels; the machine-speed adaptive probing of

Hagendorff et al. (2026) renders cost-based constraints non-binding even when the underlying suppressive modulation holds. Capability escalation here is the *same* constraint becoming non-binding under a U shifted toward zero per-attempt cost.

Allocative failure: open-weight release (Seger et al., 2023; UK AI Security Institute, 2025). Allocative modulations constrain E. When weights are publicly distributed, the enforcement substrate of E is rendered inoperative; released models cannot be recalled (International Scientific Panel on AI Safety, 2026). Agentic systems autonomously downloading, fine-tuning, and serving open-weight models constitute a $\mathcal{D}_{\text{dep}}^{\text{agentic}}$ over E for which institutional enforcement provides no constraint. Open-weight release structurally disables allocative modulation (Winner, 1980).

5. Design Principles for Distributional Robustness

Design principle 1: class-coordinated deployment. Each modulation class fails through a structurally distinct mechanism, so coordinated deployment across classes provides coverage that no single class achieves alone. This is more than the standard defence-in-depth prescription (Anderson, 2008), which stacks independent layers on the assumption that their failures are uncorrelated. The framework says why the failures are uncorrelated, and when they are not: two safeguards of the *same* class constrain the same argument of f, so a single shift in that argument defeats both at once, whereas safeguards of different classes constrain different arguments and fail only under distinct shifts. The actionable consequence is that robustness comes from spanning M, U, and E, not from adding layers within one class. A stack of three suppressive safeguards is one shift in their shared argument M away from joint failure; a suppressive–frictional–allocative stack requires an adversary to move all three arguments at once. The framework also locates the residual risk when a class is absent, by naming the argument left unconstrained.

Design principle 2: distributional calibration. Evaluation frameworks should require that the EPA be specified explicitly at design time, and that safeguard stability be assessed across distributions of U and E that include agentic adversary profiles (Shevlane et al., 2023; Souly et al., 2025; Hubinger et al., 2024). Concretely, red-team reports should specify the support of $\mathcal{D}_{\text{cal}}$—number and class of red-teamers, time per attempt, sequential versus parallelised attempts, whether reasoning models were in the loop—rather than reporting only refusal rates. Agentic actors are increasingly present in real deployments (Chan et al., 2023; Shavit et al., 2023), so the support gap between $\mathcal{D}_{\text{cal}}^{\text{human}}$ and $\mathcal{D}_{\text{dep}}^{\text{agentic}}$ widens as they accumulate, and methodology that does not account for this drift will systematically underestimate risk.

Design principle 3: per-class targets for agentic populations. The recalibration target for each class follows from Proposition 3.1 and the failure taxonomy above. For classes that fail by support mismatch, a safeguard is restored by extending $\text{supp}(\mathcal{D}_{\text{cal}})$ to cover the region of its constrained argument where the agentic population now places mass; what that region is differs by class. *Suppressive* modulations constrain M, and the agentic mass identified in Section 3.1 lies along the recursive self-revision trajectory rather than at isolated prompts. The target is coverage of that trajectory: calibration against multi-turn, self-revising adversaries so that $\text{supp}(\mathcal{D}_{\text{cal}})$ includes the intermediate states the search passes through, not only the prompts it terminates in. *Frictional* modulations constrain U, and their failure is capability escalation—per-attempt cost goes non-binding under a U shifted toward zero marginal cost by coordination across instances (Section 3.2). Re-supporting the constrained variable means binding on aggregate cost across correlated sessions rather than per attempt; the target shifts from the cost of one attempt to the cost of a coordinated campaign. As noted in Section 2.3, *allocative* modulations are the exception: they fail by domain collapse rather than support mismatch. Under open-weight release the constrained argument E is removed rather than shifted, so allocative access control cannot be re-established after release. The framework therefore yields a timing constraint rather than a coverage one: allocative modulation must act before release, or not at all. This gives a concrete pre-release test—whether the suppressive and frictional safeguards a model carries survive transfer to the open-weight setting—and a concrete implication where they do not: allocative modulation cannot be relied upon for that population, so its protection depends entirely on the suppressive and frictional classes.

A falsifiable prediction. The framework predicts what threat-model framings do not: holding model capability constant, suppressive safeguards calibrated against agentic red-teamers should exhibit a smaller failure-rate gap under agentic deployment than safeguards calibrated against human red-teamers. If enriching $\mathcal{D}_{\text{cal}}$ with agentic attack samples does not narrow this gap, the EPA account is wrong about where the violation lies. The prediction is empirically tractable and would distinguish the framework from the view that safeguard quality reduces to model-level robustness.

Impact Statement

AI safeguard practice currently treats safeguards as model-level properties evaluated against fixed adversaries. We argue this framing is increasingly misaligned with agentic

deployment, and offer a relational framework for analysis and design that makes calibration assumptions visible and testable, and that orients safeguard development toward distributional robustness rather than point-hardening.

References

Anderson, R. *Security Engineering: A Guide to Building Dependable Distributed Systems*. Wiley, Indianapolis, 2nd edition, 2008.

Bai, Y., Kadavath, S., Kundu, S., Askell, A., Kernion, J., Jones, A., Chen, A., Goldie, A., Mirhoseini, A., McKinnon, C., et al. Constitutional AI: Harmlessness from AI feedback, 2022. URL https://arxiv.org/abs/2212.08073.

Chan, A., Salganik, R., Markelius, A., Pang, C., Rajkumar, N., Krasheninnikov, D., Langosco, L., He, Z., Duan, Y., Carroll, M., Lin, M., Mayhew, A., Collins, K., Molamohammadi, M., Burden, J., Zhao, W., Rismani, S., Voudouris, K., Bhatt, U., Weller, A., Krueger, D., and Maharaj, T. Harms from increasingly agentic algorithmic systems. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*, pp. 651–666. ACM, 2023. doi: 10.1145/3593013.3594033. URL https://arxiv.org/abs/2302.10329.

Christiano, P. F., Leike, J., Brown, T. B., Martic, M., Legg, S., and Amodei, D. Deep reinforcement learning from human preferences. In *Advances in Neural Information Processing Systems*, volume 30, 2017. URL https://arxiv.org/abs/1706.03741.

Coggins, S., Saeri, A. K., Daniell, K. A., Ruster, L. P., Liu, J., and Davis, J. L. The 2025 OpenAI preparedness framework does not guarantee any AI risk mitigation practices: A proof-of-concept for affordance analyses of AI safety policies, 2025. URL https://arxiv.org/abs/2509.24394.

Davis, J. L. *How Artifacts Afford: The Power and Politics of Everyday Things*. MIT Press, Cambridge, MA, 2020. doi: 10.7551/mitpress/10962.001.0001.

Davis, J. L. 'affordances' for machine learning. In *Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency (FAccT '23)*, pp. 324–332, New York, NY, USA, 2023. ACM. doi: 10.1145/3593013.3594000.

Davis, J. L. and Chouinard, J. B. Theorizing affordances: From request to refuse. *Bulletin of Science, Technology & Society*, 36(4):241–248, 2016. doi: 10.1177/0270467617714944.

Ganguli, D., Lovitt, L., Kernion, J., Askell, A., Bai, Y., Kadavath, S., Mann, B., Perez, E., Taylor, J., et al. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned, 2022. URL https://arxiv.org/abs/2209.07858.

Gibson, J. J. *The Ecological Approach to Visual Perception*. Houghton Mifflin, Boston, 1979.

Hackett, W., Birch, L., Trawicki, S., Suri, N., and Garraghan, P. Bypassing LLM guardrails: An empirical analysis of evasion attacks against prompt injection and jailbreak detection systems. In *Proceedings of the First Workshop on LLM Security (LLMSEC)*, pp. 101–114, Vienna, Austria, 2025. Association for Computational Linguistics. URL https://aclanthology.org/2025.llmsec-1.8.

Hagendorff, T., Derner, E., and Oliver, N. Large reasoning models are autonomous jailbreak agents. *Nature Communications*, 17:1435, 2026. doi: 10.1038/s41467-026-69010-1. URL https://www.nature.com/articles/s41467-026-69010-1.

Hubinger, E., Denison, C., Mu, J., Lambert, M., Tong, M., MacDiarmid, M., Lanham, T., Ziegler, D. M., Maxwell, T., Cheng, N., et al. Sleeper agents: Training deceptive LLMs that persist through safety training, 2024. URL https://arxiv.org/abs/2401.05566.

Hutchby, I. Technologies, texts and affordances. *Sociology*, 35(2):441–456, 2001. doi: 10.1177/S0038038501000219.

Inan, H., Upasani, K., Chi, J., Rungta, R., Iyer, K., Mao, Y., Tontchev, M., Hu, Q., Fuller, B., Testuggine, D., and Khabsa, M. Llama guard: LLM-based input–output safeguard for human–AI conversations, 2023. URL https://arxiv.org/abs/2312.06674.

International Scientific Panel on AI Safety. International AI safety report 2025. Technical report, International Scientific Panel on AI Safety, 2026. URL https://internationalaisafetyreport.org.

Markus, M. L. and Silver, M. S. A foundation for the study of IT effects: A new look at DeSanctis and Poole's concepts of structural features and spirit. *Journal of the Association for Information Systems*, 9(10):609–632, 2008. doi: 10.17705/1jais.00176.

National Institute of Standards and Technology. Artificial intelligence risk management framework (AI RMF 1.0), 2023. URL https://doi.org/10.6028/NIST.AI.100-1.

Norman, D. A. *The Design of Everyday Things*. Basic Books, New York, 1988.

Norman, D. A. *The Design of Everyday Things: Revised and Expanded Edition*. Basic Books, New York, 2013.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., Zhang, C., Agarwal, S., Slama, K., Ray, A., et al. Training language models to follow instructions with human feedback, 2022. URL https://arxiv.org/abs/2203.02155.

Perez, E., Huang, S., Song, F., Cai, T., Ring, R., Aslanides, J., Glaese, A., McAleese, N., and Irving, G. Red teaming language models with language models, 2022. URL https://arxiv.org/abs/2202.03286.

Pozzi, G., Pigni, F., and Vitari, C. Affordance theory in the IS discipline: A review and synthesis of the literature. In *Proceedings of the Americas Conference on Information Systems (AMCIS)*, 2014.

Rebedea, T., Dinu, R., Sreedhar, M., Parisien, C., and Cohen, J. NeMo guardrails: A toolkit for controllable and safe LLM applications with programmable rails, 2023. URL https://arxiv.org/abs/2310.10501.

Seger, E., Dreksler, N., Moulange, R., Dardaman, E., Schuett, J., Wei, K., Winter, C., Arnold, M., Anderljung, M., Bucknall, B., et al. Open-sourcing highly capable foundation models: An evaluation of risks, benefits, and alternative methods for pursuing open-source objectives, 2023. URL https://arxiv.org/abs/2311.09227.

Sharma, M., Tong, M., Mu, J., Wei, J., Kruthoff, J., Goodfriend, S., Ong, E., Peng, A., et al. Constitutional classifiers: Defending against universal jailbreaks across thousands of hours of red teaming, 2025. URL https://arxiv.org/abs/2501.18837.

Shavit, Y., Agarwal, S., Brundage, M., Adler, S., O'Keefe, C., Campbell, R., Lee, T., Mishkin, P., Eloundou, T., Hickey, A., Slama, K., Ahmad, L., McMillan, P., Beutel, A., Passos, A., and Robinson, D. G. Practices for governing agentic ai systems. Technical report, OpenAI, 2023. URL https://cdn.openai.com/papers/practices-for-governing-agentic-ai-systems.pdf.

Shevlane, T., Farquhar, S., Garfinkel, B., Phuong, M., Whittlestone, J., Leung, J., Kokotajlo, D., Marchal, N., Anderljung, M., Kolt, N., et al. Model evaluation for extreme risks, 2023. URL https://arxiv.org/abs/2305.15324.

Shostack, A. *Threat Modeling: Designing for Security*. Wiley, Indianapolis, 2014.

Souly, A., Rando, J., Chapman, E., Davies, X., Hasircioglu, B., Shereen, E., Mougan, C., Mavroudis, V., Jones, E., Hicks, C., Carlini, N., Gal, Y., and Kirk, R. Poisoning attacks on LLMs require a near-constant number of poison samples, 2025. URL https://arxiv.org/abs/2510.07192.

Star, S. L. and Ruhleder, K. Steps toward an ecology of infrastructure: Design and access for large information spaces. *Information Systems Research*, 7(1):111–134, 1996. doi: 10.1287/isre.7.1.111.

UK AI Security Institute. Managing risks from increasingly capable open-weight AI systems, 2025. URL https://www.aisi.gov.uk/blog/managing-risks-from-increasingly-capable-open-weight-ai-systems.

Wei, A., Haghtalab, N., and Steinhardt, J. Jailbroken: How does LLM safety training fail?, 2024. URL https://arxiv.org/abs/2307.02483.

Winner, L. Do artifacts have politics? *Daedalus*, 109(1): 121–136, 1980. URL https://www.jstor.org/stable/20024652.

Zou, A., Wang, Z., Kolter, J. Z., and Fredrikson, M. Universal and transferable adversarial attacks on aligned language models, 2023. URL https://arxiv.org/abs/2307.15043.