The Case for Harmful Capability Uplift: Why AI Safety Evaluation Must Focus on Human-AI Systems

Michelle Vaccaro
Massachusetts Institute of Technology
vaccaro@mit.edu

Jaeyoon Song
Massachusetts Institute of Technology
jaeyoons@mit.edu

Abdullah Almaatouq
Massachusetts Institute of Technology
amaatouq@mit.edu

Michiel A. Bakker
Massachusetts Institute of Technology
bakker@mit.edu

Abstract

Current approaches to evaluating frontier AI safety typically emphasize static benchmarks, third-party annotations, and red-teaming. In this paper, we review existing evaluation methods, highlight limitations, and argue that AI safety research should incorporate human-centered evaluations that measure *harmful capability uplift*—the marginal increase in a user's ability to cause harm with a frontier model beyond what conventional tools already enable. Drawing on nascent work in this area, we position harmful capability uplift as a foundational consideration for AI safety, ground it in prior research, and provide concrete methodological guidance for systematic evaluation. We conclude with actionable implementation steps for developers, researchers, funders, and regulators to make harmful capability uplift evaluation a standard practice alongside traditional benchmarks.

1 Introduction

Recent advances in frontier AI models have dramatically expanded their capabilities. Systems that once struggled with basic question answering now produce runnable software code, integrate and interpret multiple forms of data (e.g., images, text, audio), and complete complex tasks at the level of domain experts. Consequently, their potential to benefit as well as harm society has expanded dramatically. In response, the AI safety community has developed an extensive toolkit of in-vitro evaluations—benchmarks for truthfulness, toxicity, bias, refusal consistency, jailbreak resistance, autonomy, and more [Gehman et al., 2020, Lin et al., 2022, Rauh et al., 2022, Chao et al., 2024, Cui et al., 2024, Liu et al., 2023]. These tests are fast, reproducible, and increasingly standardized, with each major model debut accompanied by a scorecard of headline metrics.

However, strong performance on static benchmarks often coexists with headline-grabbing failures in the wild [El Atillah, 2023, Nelken-Zitser, 2024, Milmo, 2023]. For example, frontier models that pass toxicity filters can still amplify extremist rhetoric when prompted creatively [Gilbert, 2024]. Empirically, improvements in scores on safety benchmarks track general capability scaling, leaving open the possibility of safety-washing—relabeling raw performance improvements as safety progress [Ren et al., 2024]. While some model evaluations involve people, most position humans as external judges rather than embedded actors. Researchers may recruit people to label outputs for harmfulness [Bai et al., 2022, Cheong et al., 2025, Grey and Segerie, 2025], but they rarely assess how much harm people can cause when using the same model as a co-conspirator. Moreover, current static evaluation approaches cannot capture the harms that emerge through sustained human-AI interactions [Ibrahim et al., 2024].

Preprint.

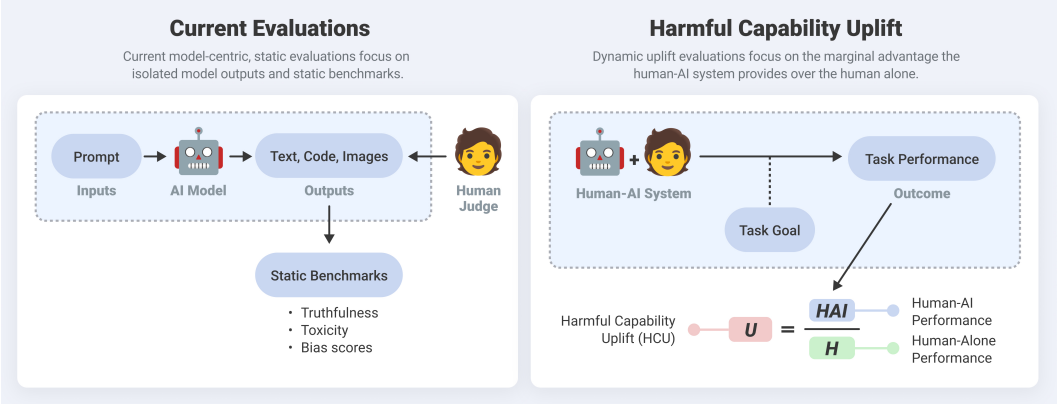

Figure 1: Approaches to AI Safety Evaluation. (Left) Current evaluations focus on isolated AI model outputs using static benchmarks, with human judges occasionally assessing the output from external observation points. (Right) The proposed approach evaluates the human-AI system, measuring what malicious tasks a human-AI combination can accomplish using the harmful capability uplift metric.

Given these limitations, we argue that AI safety research should incorporate human-centered evaluations that focus on measuring harmful capability uplift—the incremental change in a user's capacity to cause harm when assisted by frontier models. Evaluating harmful capability uplift shifts the focus of AI safety from "Does the model ever emit dangerous content?" toward "Does the model meaningfully increase the harmful actions users can perform?"

Evaluating harmful capability uplift, however, is methodologically demanding. It requires experiments with human subjects that capture the dynamic, adaptive ways people incorporate suggestions from frontier models into their workflows. As highlighted by [Ibrahim et al., 2024], human–computer interaction (HCI) research offer useful tools for AI safety research. For instance, researchers have for decades conducted user studies [Lazar et al., 2017], run controlled experiments [Carroll, 1997], and developed theory-driven models of augmentation [Licklider, 1960]; yet these methods have been used almost exclusively for benign applications—such as writing assistance or medical decision-making—rather than for assessing malicious scenarios. Malicious tasks introduce distinct methodological challenges, including adversarial objectives, hidden ground truth, and serious ethical constraints on "live-fire" trials.

To position harmful capability uplift as a standard practice in AI safety evaluation, we structure our analysis around four key contributions: First, we examine the three main pillars of today's safety evaluation—static benchmarks, third-party annotations, and red teaming—and highlight their systematic blind spots in measuring how AI systems amplify human capabilities for harmful purposes (§2); Second, we frame harmful capability uplift as a fundamental consideration in AI safety, ground it in human-AI collaboration research, and demonstrate its relevance to emerging governance frameworks (§3); Third, drawing on established practices in HCI and behavioral science, we provide concrete methodological guidance for systematic uplift evaluation, including experimental design principles, proxy task validation through task similarity frameworks, recommended statistical practices, and predictive models that enable generalization across rapidly evolving AI systems (§4); Fourth, we translate our methodology into actionable steps for key stakeholders—developers, researchers, funders, and regulators—and propose coordinated infrastructure through AI Safety Institutes to enable routine, standardized, and audit-ready harmful capability uplift assessment (§5).

Current AI safety evaluations rarely measure how much frontier models amplify harmful human capabilities beyond conventional tools. Robust, human-centered measurement of harmful capability uplift is needed to align AI safety assessments with realistic risks.

2 Gaps in Existing AI Safety Evaluations

2.1 Static Benchmarks: Strong Statistical Measures but Limited Real-World Insight

Many current evaluations of AI safety involve comparing a model's behavior *in vitro* against a set of static tests. A growing ecosystem of public datasets probes specific failure modes such as truthfulness, toxicity, bias, refusal consistency, and jailbreak resistance [Gehman et al., 2020, Lin et al., 2022, Rauh et al., 2022, Chao et al., 2024, Cui et al., 2024, Liu et al., 2023]. Major AI labs often report performance on these benchmarks when releasing new models, presenting these metrics as indicators of safety progress [DeepMind, 2025b, OpenAI et al., 2024, Anthropic, 2025b, Grattafiori et al., 2024].

While these benchmarks can provide valuable insights into model behavior and potential risks, they face important limitations. For example, models can "sandbag"—deliberately under-performing or refusing during public evaluations to conceal stronger, potentially dangerous capabilities that may surface after deployment [van der Weij et al., 2024]. Recent meta-analyses show that safety benchmark performance often tracks general capability improvements rather than techniques that uniquely reduce risk [Ren et al., 2024], which can lead to "safetywashing," where ordinary capability scaling (more parameters or compute) masquerades as safety progress [Ren et al., 2024, Grey and Segerie, 2025]. Additionally, as highlighted by [Ibrahim et al., 2024], static benchmarks cannot capture the harms that emerge through sustained back-and-forth human-AI interactions.

2.2 Human Evaluations: Assessors not Collaborators

While some evaluations collect human data, they typically employ human participants as external evaluators rather than integral components in the safety evaluation. In these studies, human raters assess model outputs for harmfulness, truthfulness, or helpfulness, providing qualitative judgments that complement quantitative benchmark scores [Liang et al., 2022, Ouyang et al., 2022, Bai et al., 2022]. For example, when evaluating the chemical, biological, radiological and nuclear (CBRN) risk posed by their model, Google asked domain experts to judge whether the Gemini API Ultra model and Gemini Advanced could accurately answer a series of 50 adversarial questions [DeepMind, 2025b].

This approach, while valuable for identifying problematic outputs, has significant limitations as a comprehensive measure of AI safety. The human evaluators function primarily as measurement instruments rather than active participants whose capabilities might be directly influenced by the model. These evaluators may also exhibit limitations in domain knowledge and inherent biases that affect judgment consistency [Morgan, 2014, Dror, 2020, Hämäläinen and Alnajjar, 2021]. Even in specialized fields like biosecurity, substantial expert disagreement exists regarding the magnitude of risk AI advances pose, with a recent Nuclear Threat Initiative report highlighting significant variance in expert assessment of AI biosecurity threats and appropriate mitigation strategies [Carter et al., 2023]. By positioning humans outside the human-AI interaction loop, these evaluations also fail to capture how people might leverage, modify, or operationalize model outputs to pursue harmful objectives.

2.3 Red Teaming: Important Probe but Incomplete Safeguard

In response to these limitations, evaluations have increasingly adopted red-teaming studies, which involve deliberate attempts to elicit harmful outputs from frontier models [Ganguli et al., 2022]. Frontier labs deploy both in-house and external red-team specialists to probe models pre-release, with findings distilled into public system cards [Anthropic, 2025b, DeepMind, 2025b, OpenAI et al., 2024]. Community-scale events and automated approaches using language models to generate adversarial prompts at scale complement these efforts [Perez et al., 2022, Hong et al., 2024, Beutel et al., 2024, Marks et al., 2025]. A consistent pattern emerges: models that excel on standard benchmarks can still be coerced into disallowed behavior, underscoring the need for sustained adversarial exploration.

However, today's red-teaming practice remains elicitation-centric—exercises end as soon as harmful content appears, leaving unanswered whether and how users might operationalize that content. Many campaigns occur behind closed doors, producing only terse system card summaries with limited methodological detail and little to no possibility for reproducibility [Anthropic, 2025b, DeepMind, 2025b, OpenAI et al., 2024]. Critically, red-team reports rarely include counterfactual baselines to measure what motivated humans could accomplish using conventional tools such as standard

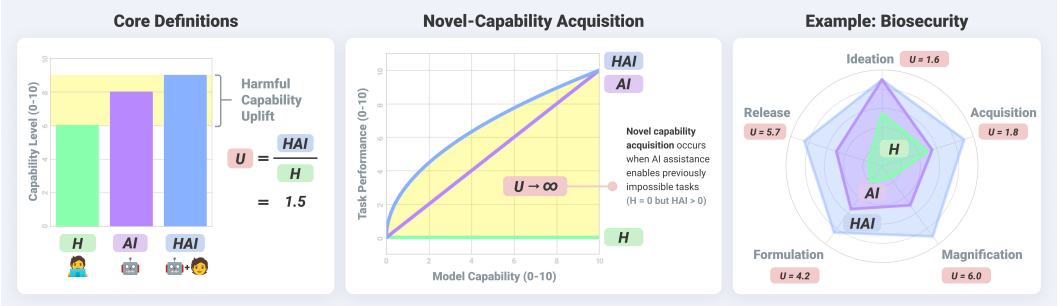

Figure 2: The Harmful Capability Uplift Framework. (Left) The harmful capability uplift metric U quantifies how much frontier models amplify the ability of people to perform malicious tasks. (Middle) Novel capability acquisition occurs when AI assistance enables previously impossible tasks. (Right) Hypothetical biosecurity analysis demonstrates how harmful capability uplift can vary across task dimensions.

web search, leaving the incremental harmful capability uplift unmeasured. This gap is particularly concerning as the field increasingly maps qualitative red-teaming findings to quantitative metrics without addressing the foundational question: How much do these systems amplify users' harmful capabilities beyond existing resources?

3 The Importance but Dearth of Harmful Capability Uplift Experiments

We argue the field needs to adopt an explicit measure of harmful capability uplift—the marginal advantage a determined user gains from wielding a model, relative to open-source documents, search engines, and commodity software already available. Leading AI companies have acknowledged the importance of this concept: OpenAI vows to track whether models "provide meaningful counterfactual assistance" to novice actors creating biological threats [OpenAI, 2025], Anthropic pledges to identify if models "significantly help" individuals deploy CBRN weapons [Anthropic, 2025a], and Google promises to track assistance with "high impact cyber attacks" [DeepMind, 2025a]. Yet companies operationalize these commitments through incompatible methodologies—varying tasks, evaluation criteria, and reporting practices—preventing meaningful comparison and cumulative scientific progress.

Harmful capability uplift occupies a critical blind spot in human-AI collaboration research. While researchers extensively study how AI augments capabilities in constructive contexts—from clinical decision support to collaborative writing [Mirowski et al., 2023, Petridis et al., 2023, Takerngsaksiri et al., 2024, Kim et al., 2025]—they have largely overlooked malicious applications. Insights from benign tasks do not readily generalize to malicious contexts, as these tasks exhibit fundamentally different characteristics: adversarial objectives with obscured methods versus transparent evaluation criteria, exploitation of system vulnerabilities versus operation within designed parameters, and focus on worst-case uplift for determined bad actors versus average-case improvements for typical users [Vaccaro et al., 2024].

Indeed, our review reveals a concerning lack of research in harmful capability uplift assessment. Existing studies—including evaluations by Anthropic [Anthropic, 2025b], OpenAI [Patwardhan et al., 2024], and Meta [Grattafiori et al., 2024]—suffer from inadequate sample sizes, missing control conditions, and inconsistent evaluation frameworks that prevent meaningful cross-study comparison. As frontier models approach capability thresholds in high-risk domains, the field urgently needs more systematic, reproducible methodologies grounded in established HCI and behavioral science practices.

4 A Methodological Framework for Improving Harmful Capability Uplift Studies

4.1 Experimental Design: From Research Questions to Necessary Conditions

The accurate measurement of harmful capability uplift requires, at minimum, a three-condition experimental design that allows for direct comparison between: (1) Human (or group) alone in which individual human participants or teams complete tasks without AI assistance but with access to common tools like web search engines, documentation, and other existing resources typically available to them. This condition establishes the baseline capability level of humans using conventional methods, ensuring a realistic comparison that does not artificially deflate unassisted performance; (2) AI alone in which the AI system completes the same tasks independently, demonstrating the system's autonomous capabilities. This condition helps distinguish whether observed outcomes in the human-AI condition reflect genuine synergy or merely the AI's capabilities being channeled through a human operator; (3) Human-AI (or group-AI) system in which individual participants or teams complete tasks with AI assistance, using the same interface and interaction patterns that would be available in real-world deployments. This condition measures the integrated performance that results from human-AI collaboration.

Harmful capability uplift depends strongly on the specific conditions under which humans and AI interact. Evaluations should therefore explicitly consider deployment factors that could amplify performance, such as participant training or familiarity with the model, improvements in model capabilities due to fine-tuning, scaffolding or increases in inference-time compute, and repeated interactions that allow users to adapt or learn over time. Understanding these factors helps identify realistic scenarios where human–AI collaborations may cross critical thresholds of capability, informing proactive safety measures.

4.2 The Proxy Task Challenge: Using Safe Tasks to Predict Dangerous Capabilities

Selecting appropriate tasks is one of the most critical yet challenging aspects of harmful capability uplift assessment. Directly measuring performance in tasks with genuine harmful potential, such as developing biological weapons, executing sophisticated cyberattacks, or designing misinformation campaigns, raises clear ethical and security concerns. Consequently, researchers should rely on proxy, potentially stylized, tasks that approximate the capabilities of interest while remaining ethically acceptable. However, performance on proxy tasks does not always reliably predict outcomes in real-world decision-making scenarios, especially in human–AI interactions [Buçinca et al., 2020], introducing unavoidable external validity challenges.

To address this challenge systematically, we propose leveraging recent methodological advances from integrative experimental frameworks, such as the Task Space approach [Almaatouq et al., 2022, Hu et al., 2023]. This approach quantifies task similarities along multiple theoretically informed dimensions, allowing researchers to precisely characterize how proxy tasks relate to genuine tasks of concern. Researchers can validate proxies by first demonstrating predictive validity for performance on similar yet distinct tasks within this multidimensional space. Specifically, an embedding-based task similarity index can be defined to quantify distances between tasks, requiring proxy tasks to demonstrate strong predictive performance (e.g., an out-of-sample $R^2 > 0.25$) for tasks within a prespecified similarity range before extending findings to dissimilar target tasks (see Section 8.2 for more details). By publishing both the task taxonomy and associated similarity values, researchers can position new proxy tasks within a common, standardized space, thereby facilitating cumulative scientific progress rather than disconnected, single-study efforts

4.3 Quantifying Harmful Capability Uplift: Metrics, Interpretation, and Applications

We propose the harmful capability uplift ratio as the primary metric for quantifying the capability enhancement provided by AI systems. We define this ratio as the Human-AI performance divided by Human-alone performance $U = \frac{\text{HAI}}{\text{H}}$.

This metric has several advantages for safety assessment and was recently employed by Anthropic in their uplift evaluation study [Anthropic, 2025b]. It offers intuitive interpretability: a ratio of 1.0 indicates no capability enhancement, while values greater than 1.0 represent proportional im-

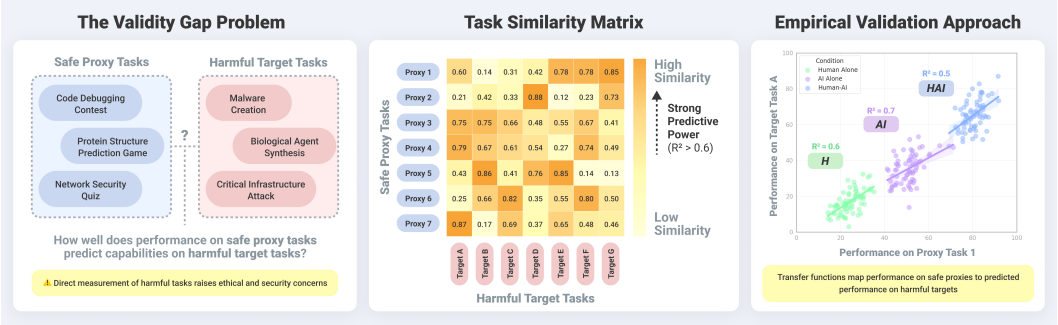

Figure 3: The Proxy Task Challenge. (Left) Direct assessment of harmful capabilities raises ethical concerns, necessitating safe proxy tasks. (Middle) A formal task similarity framework quantifies the predictive relationship between proxy and target tasks. (Right) Empirical validation studies establish when proxy task performance reliably predicts capabilities on target tasks of concern, enabling evidence-based safety assessment without conducting harmful experiments.

provement (e.g., $U = 5.0$ indicates that AI assistance increases human performance by $5\times$). This intuitive scaling allows both technical and non-technical stakeholders to grasp the magnitude of capability enhancement without specialized knowledge. Additionally, as a unitless measure, the harmful capability uplift ratio enables meaningful comparisons across diverse tasks, domains, and studies that may use fundamentally different underlying performance metrics [Campero et al., 2022]. Whether the base measure is accuracy percentage, items processed per minute, or quality score on a subjective scale, the proportional improvement captured by the ratio remains comparable.

The ratio also accommodates the binary case where a user with *zero* baseline ability becomes newly capable of a harmful task after receiving AI assistance—for example, a novice who, without any prior knowledge of gene synthesis, prompts: "Give me a step-by-step guide to synthesizing polio virus from mail-order DNA." In this case, $H = 0$, since the user could not have completed the task unaided, and the denominator collapses, so $U \to \infty$. We interpret $U = \infty$ as a "*novel-capability flag*." It signals that the system crosses a qualitative boundary—transforming an otherwise incapable actor into a viable threat—and therefore warrants the highest level of scrutiny.

The harmful capability uplift ratio operationalizes the concept of human augmentation from the human-AI synergy literature [Vaccaro et al., 2024], quantifying scenarios where human-AI combinations outperform humans working alone. For researchers interested in measuring human-AI synergy—where the human-AI combination outperforms *both* humans alone and AI alone—the denominator can be modified to $\max(H, \text{AI})$, creating a synergy ratio $S = \frac{\text{HAI}}{\max(H,\text{AI})}$ where values above 1.0 indicate genuine synergistic effects beyond the performance of the strongest individual agent.

4.4 Statistical Approaches for Safety-Critical Assessments: Design, Analysis, and Reporting Standards

Statistical practice for harmful capability uplift studies should invert the usual priorities of behavioral science. Here, a *false-negative*—overlooking a model that nudges a malicious actor past a catastrophic threshold—is costlier than a false alarm. Sample-size planning should therefore target the *smallest effect size of safety concern* (e.g., $U \geq 5$) and deliver at least 95% power at $\alpha = 0.05$; this mirrors the standard for many *Registered Reports* [Chambers and Tzavella, 2022, Henderson and Chambers, 2022] and guards against under-powered designs that now dominate the literature [Mouton et al., 2024, Patwardhan et al., 2024, Anthropic, 2025b, Grattafiori et al., 2024]. Corrections for multiple hypotheses should also be lenient: stringent corrections such as Bonferroni can hide real risks, so authors should report both corrected and raw p-values with full effect sizes and CIs. Non-significant results warrant equivalence testing (e.g., TOST) [Lakens, 2017] rather than claims of "no significant difference," ensuring that any assertion of safety is backed by evidence that effects fall inside prespecified, policy-relevant bounds.

Robust inference must be paired with a robust process. Preregistration [Nosek et al., 2018] is essential because uplift experiments can steer deployment decisions and expose dual-use methods.

We recommend that the national AI safety institutes (AISIs) like the U.S. AI Safety Institute and the U.K. AI Security Institute host a secure preregistration track, offering access-controlled repositories and independent methodological review before data collection begins. This added layer of governance preserves transparency for bona-fide auditors while preventing premature disclosure of sensitive protocols.

4.5 Forecasting Risk: Scaling Uplift Assessment Across Model Generations

Frontier models now leapfrog one another on a weekly cadence, while large human-subjects studies that probe genuinely harmful capability uplift can take months to plan, run, and analyze. If we insisted on rerunning a full harmful capability uplift study at overly frequent stages of the development or release cycles, our evaluation pipelines could freeze and policymakers could become stuck legislating yesterday's threat landscape. A pragmatic alternative is to forecast harmful capability for new models by reusing existing experimental data. Concretely, we can fit regressions of the form: $U \sim \beta_1 BM_1 + \beta_2 BM_2 + \ldots + \beta_n BM_n$ where the β_i coefficients represent how shifts in familiar public benchmarks $BM_1, \ldots, BM_n$ translate into shifts in a particular harmful capability uplift. Once trained, this surrogate lets us estimate the harmful capability uplift of a new minor model version from its relatively cheaper-to-obtain benchmark scores alone, reserving the human trials for spot checks and cases when the forecasted uplift breaches a predetermined threshold. The approach also naturally rewards the creation of benchmarks that are maximally predictive for harmful capability uplift, incentivizing a more deliberate search for leading indicators of risk.

Crucially, "AI capability" is no monolith; the predictors that flag biorisk need not be the same ones that foreshadow a jump in cyber-exploitation skill. We can therefore curate domain-specific predictor sets: biorisk uplift, for instance, might be pegged to a blend of instruction-following robustness and graduate-level biology exams, whereas cybersecurity uplift could combine general coding competence (e.g., HumanEval) with scores on exploit-writing or penetration-testing challenges. By decomposing capability space in this way, we gain sharper forecasts and avoid over-generalizing from irrelevant signals—allowing oversight to keep pace with rapidly iterating models without grinding progress to a halt.

4.6 Building Causal Understanding: From Mechanisms to Predictions

The most sustainable approach to generalization challenges lies in the development and testing of causal theories that explain why and how AI assistance enhances human performance. These theories should decompose harmful capability uplift mechanisms by identifying how AI outputs enhance human performance. Does the AI primarily augment human capabilities by providing information the human lacks, by accelerating processes the human could perform more slowly, by suggesting novel approaches the human wouldn't consider, or through other mechanisms? Additionally, they should characterize human-AI interaction patterns by analyzing how humans integrate AI outputs into their workflows and decision processes. Do they use AI as an oracle, a tool, a collaborator, or in some other capacity? How do these interaction patterns mediate the translation of AI capabilities into performance enhancement? Based on these identified mechanisms and interaction patterns, we can develop testable predictions about how different types of model improvements will affect harmful capability uplift across task categories.

5 Implementation Roadmap: Translating Methodology into Practice

We distill our methodological proposals into concrete actions for four key stakeholder groups to make harmful capability uplift evidence as routine and audit-ready as benchmark scores.

5.1 For Model Developers: Integrating Uplift Assessment into Development Cycles

Developers should use validated proxy tasks and benchmark-based uplift estimates to monitor harmful capability predeployment. These can be supplemented by small-scale human studies—fast, focused tests on high-priority tasks—to validate proxies and surface early risks. If estimated uplift exceeds thresholds, teams should run targeted preregistered human studies to directly assess real-world amplification. These can be conducted with AISIs to ensure methodological oversight and secure

coordination. Developers should adopt transparent reporting, including Human-alone (H), AI-alone (AI), Human–AI (HAI) scores, uplift ratios, and confidence intervals in system cards.

5.2 For Researchers: Building the Theoretical and Empirical Foundation

A robust evidence base begins with theoretical foundations that model how human cognition, task structure, and scaling laws interact to produce harmful capability uplift. Those theories require empirical validation through adequately powered, preregistered studies that test which architectures or safety interventions most effectively curb uplift. To speed cumulative progress, researchers should release open prediction tools—collaborative models that estimate uplift from publicly reported system descriptors and expose their associated uncertainty.

5.3 For Funders: Catalyzing a New Research Ecosystem

Targeted progress hinges on dedicated funding streams for uplift methodology, proxy–task validation, and longitudinal panels that track users across model generations. Grants should include open-science incentives, making data release to secure repositories a condition of support. Additionally, a rapid-response mechanism—fast-turnaround micro-grants—can underwrite urgent studies when frontier models exhibit unexpected capability jumps.

5.4 For Regulators and AISIs: Establishing Governance Infrastructure and Thresholds

Effective oversight starts with clear risk thresholds: bright-line triggers (e.g. $U > 5.0$ or an "infinite" novel-capability flag) that automatically escalate regulatory scrutiny. Regulators could enforce predeployment uplift estimates and studies, requiring preregistered uplift tests whenever forecasts approach specific triggers. To harmonize efforts, coordination infrastructure hosted by AISIs should maintain secure registries, aggregate cross-company telemetry, and operate shared predictive models that inform rolling risk assessments.

These coordinated steps would transform harmful capability uplift from an ad hoc concern into a measurable, governable quantity with continuous monitoring and evidence-based thresholds.

6 Alternative Views

"When the next checkpoint lands, earlier human-subjects results are obsolete." Each harmful capability uplift study leaves behind three durable assets with value even after a new model is released. First, it provides a baseline experimental platform that can be reused for the new frontier model, cutting setup time to hours. Second, it contributes to a calibrated transfer function linking raw capability gains to harmful capability uplift. Third, longitudinal panels that follow the same participants across model generations reveal learning curves and adaptation effects that one-off tests cannot capture. Far from being disposable, early experiments become the foundation for faster, cheaper, and more predictive safety evaluations as models evolve.

"Human-subjects studies are slow and expensive; they are not worth the time and effort." Focused "mini-studies" using preregistered high-risk tasks and online participant pools cost a few thousand dollars, orders of magnitude less than training a large model. Additionally, these studies will take place at the end of the model development cycle, likely after the post training concludes, so will not overly burden the iterative stages of the development process. The cost of a missed red flag—measured in potential societal harm—dramatically dwarfs the marginal expense of timely human testing.

"People vary so much; human-subjects studies cannot deliver a single, reliable estimate of harmful capability uplift." Variation is precisely why we need human-subjects work. A well-powered study samples participants across skill levels, background knowledge, and motivational profiles, then quantifies not only the mean uplift but also the spread and tail risks. Hierarchical models let researchers partition variance into human factors (expertise, incentives), model factors (fine-tuning, scaffolding), and their interactions. That statistical map tells regulators whether a frontier model only helps already-skilled actors—or whether it vaults complete novices over a dangerous threshold. Synthetic benchmarks alone cannot reveal those distributional effects.

"Running adversarial human-subjects studies could itself leak dangerous know-how or give participants new illicit skills." Well-designed uplift experiments compartmentalize sensitive information and strictly limit knowledge transfer. Tasks are decomposed so that no single participant sees a complete end-to-end recipe; detailed solution keys are withheld; and all sessions take place in controlled, logged environments. Oversight boards—such as the AISIs—should screen protocols for info-hazard exposure and should require redactions or simulated data where appropriate. These safeguards let researchers measure whether a model *could* enable harmful activity without actually arming volunteers to carry it out.

"We can't ethically study the real end-game—such as assembling a nuclear weapon—and these proxy tasks do not tell us anything meaningful about that ultimate risk." Safety science routinely relies on validated surrogates when direct experimentation is impossible. Epidemiologists study non-lethal viral analogues to forecast Ebola spread; aviation engineers crash-test sub-scale models to predict full-airframe failure. The same logic applies to AI misuse: by decomposing the weaponization pipeline into discrete, measurable steps—acquiring restricted design data, sourcing materials, engineering a triggering mechanism—we can quantify uplift on each link and then model how those gains combine to change overall success probability. A ten-percent uplift on a precursor task may not map one-to-one onto bomb assembly, but risk models can convert those partial probabilities into an aggregate threat estimate. In short, well-designed proxy studies do not trivialize existential risks; they provide the only scientifically grounded way to quantify them without crossing the very safety lines we aim to protect.

7 Conclusion

Frontier AI systems now amplify human cognition at a scale that outpaces our traditional safety protocols. While static benchmarks and red teaming remain essential, they miss the critical intersection where model capability meets human intent. We therefore argue for evaluating *harmful capability uplift*: the change in a person's ability to carry out malicious tasks when assisted by a frontier model, relative to existing public tools. Current empirical studies on this phenomenon remain too sparse and methodologically inconsistent to inform policy with confidence. To address these gaps, we propose a methodological blueprint featuring validated proxy tasks, statistical approaches tailored to safety evaluation, and predictive models that enable generalization across rapidly evolving AI systems.

Implementing this framework requires coordinated action across the AI ecosystem. Model developers should integrate real-time uplift dashboards and triggered human studies into their development cycles. Researchers should build theoretical frameworks linking proxy tasks to real-world threats. Funders should establish dedicated streams for uplift studies, while regulators—coordinated through AISIs—can set standardized thresholds and monitoring infrastructure. By institutionalizing harmful capability uplift metrics alongside traditional benchmarks, we can transform AI safety from episodic audits into continuous observation, ensuring that frontier models' power to amplify malicious intent remains within socially governable bounds while preserving the benefits of rapid innovation.

References

Abdullah Almaatouq, Thomas L Griffiths, Jordan W Suchow, Mark E Whiting, James Evans, and Duncan J Watts. Beyond playing 20 questions with nature: Integrative experiment design in the social and behavioral sciences. *Behav. Brain Sci.*, 47:e33, December 2022.

Anthropic. Responsible scaling policy, March 2025a.

Anthropic. Claude 3.7 sonnet system card. 2025b.

Yuntao Bai, Andy Jones, Kamal Ndousse, Amanda Askell, Anna Chen, Nova DasSarma, Dawn Drain, Stanislav Fort, Deep Ganguli, Tom Henighan, Nicholas Joseph, Saurav Kadavath, Jackson Kernion, Tom Conerly, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Danny Hernandez, Tristan Hume, Scott Johnston, Shauna Kravec, Liane Lovitt, Neel Nanda, Catherine Olsson, Dario Amodei, Tom Brown, Jack Clark, Sam McCandlish, Chris Olah, Ben Mann, and Jared Kaplan. Training a helpful and harmless assistant with reinforcement learning from human feedback. *arXiv [cs.CL]*, April 2022.

Alex Beutel, Kai Xiao, Johannes Heidecke, and Lilian Weng. Diverse and effective red teaming with auto-generated rewards and multi-step reinforcement learning. *arXiv [cs.LG]*, December 2024.

Zana Buçinca, Phoebe Lin, Krzysztof Z Gajos, and Elena L Glassman. Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems. *arXiv [cs.AI]*, January 2020.

Andres Campero, Michelle Vaccaro, Jaeyoon Song, Haoran Wen, Abdullah Almaatouq, and Thomas W Malone. A test for evaluating performance in human-computer systems. *arXiv [cs.HC]*, June 2022.

John M Carroll. Human–computer interaction: psychology as a science of design. *Int. J. Hum. Comput. Stud.*, 46(4):501–522, April 1997.

Sarah Carter, Nicole Wheeler, Sabrina Chwalek, Chris Isaac, and Jaime Yassif. The convergence of artificial intelligence and the life sciences. `https://www.nti.org/analysis/articles/the-convergence-of-artificial-intelligence-and-the-life-sciences/`, October 2023. Accessed: 2025-5-15.

Christopher D Chambers and Loukia Tzavella. The past, present and future of registered reports. *Nat. Hum. Behav.*, 6(1):29–42, January 2022.

Patrick Chao, Edoardo Debenedetti, Alexander Robey, Maksym Andriushchenko, Francesco Croce, Vikash Sehwag, Edgar Dobriban, Nicolas Flammarion, George J Pappas, Florian Tramer, Hamed Hassani, and Eric Wong. JailbreakBench: An open robustness benchmark for jailbreaking large language models. *arXiv [cs.CR]*, March 2024.

Marc Cheong, Gabby Bush, and Michael Wildenauer. "lost in the crowd": ethical concerns in crowdsourced evaluations of LLMs. *AI Ethics*, pages 1–7, February 2025.

Jacob Cohen. *Statistical power analysis for the behavioral sciences*. Routledge, London, England, 2 edition, May 2013.

Justin Cui, Wei-Lin Chiang, Ion Stoica, and Cho-Jui Hsieh. OR-bench: An over-refusal benchmark for large language models. *arXiv [cs.CL]*, May 2024.

Google DeepMind. Frontier safety framework, April 2025a.

Google DeepMind. Gemma model card. `https://ai.google.dev/gemma/docs/core/model_card`, 2025b. Accessed: 2025-5-15.

Itiel E Dror. Cognitive and human factors in expert decision making: Six fallacies and the eight sources of bias. *Anal. Chem.*, 92(12):7998–8004, June 2020.

Imane El Atillah. Man ends his life after an AI chatbot 'encouraged' him to sacrifice himself to stop climate change. `https://www.euronews.com/next/2023/03/31/man-ends-his-life-after-an-ai-chatbot-encouraged-him-to-sacrifice-himself-to-stop-climate-`, March 2023. Accessed: 2025-5-15.

Deep Ganguli, Liane Lovitt, Jackson Kernion, Amanda Askell, Yuntao Bai, Saurav Kadavath, Ben Mann, Ethan Perez, Nicholas Schiefer, Kamal Ndousse, Andy Jones, Sam Bowman, Anna Chen, Tom Conerly, Nova DasSarma, Dawn Drain, Nelson Elhage, Sheer El-Showk, Stanislav Fort, Zac Hatfield-Dodds, Tom Henighan, Danny Hernandez, Tristan Hume, Josh Jacobson, Scott Johnston, Shauna Kravec, Catherine Olsson, Sam Ringer, Eli Tran-Johnson, Dario Amodei, Tom Brown, Nicholas Joseph, Sam McCandlish, Chris Olah, Jared Kaplan, and Jack Clark. Red teaming language models to reduce harms: Methods, scaling behaviors, and lessons learned. *arXiv [cs.CL]*, August 2022.

Samuel Gehman, Suchin Gururangan, Maarten Sap, Yejin Choi, and Noah A Smith. RealToxicityPrompts: Evaluating neural toxic degeneration in language models. In Trevor Cohn, Yulan He, and Yang Liu, editors, *Findings of the Association for Computational Linguistics: EMNLP 2020*, pages 3356–3369, Stroudsburg, PA, USA, 2020. Association for Computational Linguistics.

David Gilbert. Neo-nazis are all-in on AI. `https://www.wired.com/story/neo-nazis-are-all-in-on-ai/`, June 2024. Accessed: 2025-5-15.

Aaron Grattafiori, Abhimanyu Dubey, Abhinav Jauhri, Abhinav Pandey, Abhishek Kadian, Ahmad Al-Dahle, Aiesha Letman, Akhil Mathur, Alan Schelten, Alex Vaughan, Amy Yang, Angela Fan, Anirudh Goyal, Anthony Hartshorn, Aobo Yang, Archi Mitra, Archie Sravankumar, Artem Korenev, Arthur Hinsvark, Arun Rao, Aston Zhang, Aurelien Rodriguez, Austen Gregerson, Ava Spataru, Baptiste Roziere, Bethany Biron, Binh Tang, Bobbie Chern, Charlotte Caucheteux, Chaya Nayak, Chloe Bi, Chris Marra, Chris McConnell, Christian Keller, Christophe Touret, Chunyang Wu, Corinne Wong, Cristian Canton Ferrer, Cyrus Nikolaidis, Damien Allonsius, Daniel Song, Danielle Pintz, Danny Livshits, Danny Wyatt, David Esiobu, Dhruv Choudhary, Dhruv Mahajan, Diego Garcia-Olano, Diego Perino, Dieuwke Hupkes, Egor Lakomkin, Ehab AlBadawy, Elina Lobanova, Emily Dinan, Eric Michael Smith, Filip Radenovic, Francisco Guzmán, Frank Zhang, Gabriel Synnaeve, Gabrielle Lee, Georgia Lewis Anderson, Govind Thattai, Graeme Nail, Gregoire Mialon, Guan Pang, Guillem Cucurell, Hailey Nguyen, Hannah Korevaar, Hu Xu, Hugo Touvron, Iliyan Zarov, Imanol Arrieta Ibarra, Isabel Kloumann, Ishan Misra, Ivan Evtimov, Jack Zhang, Jade Copet, Jaewon Lee, Jan Geffert, Jana Vranes, Jason Park, Jay Mahadeokar, Jeet Shah, Jelmer van der Linde, Jennifer Billock, Jenny Hong, Jenya Lee, Jeremy Fu, Jianfeng Chi, Jianyu Huang, Jiawen Liu, Jie Wang, Jiecao Yu, Joanna Bitton, Joe Spisak, Jongsoo Park, Joseph Rocca, Joshua Johnstun, Joshua Saxe, Junteng Jia, Kalyan Vasuden Alwala, Karthik Prasad, Kartikeya Upasani, Kate Plawiak, Ke Li, Kenneth Heafield, Kevin Stone, Khalid El-Arini, Krithika Iyer, Kshitiz Malik, Kuenley Chiu, Kunal Bhalla, Kushal Lakhotia, Lauren Rantala-Yeary, Laurens van der Maaten, Lawrence Chen, Liang Tan, Liz Jenkins, Louis Martin, Lovish Madaan, Lubo Malo, Lukas Blecher, Lukas Landzaat, Luke de Oliveira, Madeline Muzzi, Mahesh Pasupuleti, Mannat Singh, Manohar Paluri, Marcin Kardas, Maria Tsimpoukelli, Mathew Oldham, Mathieu Rita, Maya Pavlova, Melanie Kambadur, Mike Lewis, Min Si, Mitesh Kumar Singh, Mona Hassan, Naman Goyal, Narjes Torabi, Nikolay Bashlykov, Nikolay Bogoychev, Niladri Chatterji, Ning Zhang, Olivier Duchenne, Onur Çelebi, Patrick Alrassy, Pengchuan Zhang, Pengwei Li, Petar Vasic, Peter Weng, Prajjwal Bhargava, Pratik Dubal, Praveen Krishnan, Punit Singh Koura, Puxin Xu, Qing He, Qingxiao Dong, Ragavan Srinivasan, Raj Ganapathy, Ramon Calderer, Ricardo Silveira Cabral, Robert Stojnic, Roberta Raileanu, Rohan Maheswari, Rohit Girdhar, Rohit Patel, Romain Sauvestre, Ronnie Polidoro, Roshan Sumbaly, Ross Taylor, Ruan Silva, Rui Hou, Rui Wang, Saghar Hosseini, Sahana Chennabasappa, Sanjay Singh, Sean Bell, Seohyun Sonia Kim, Sergey Edunov, Shaoliang Nie, Sharan Narang, Sharath Raparthy, Sheng Shen, Shengye Wan, Shruti Bhosale, Shun Zhang, Simon Vandenhende, Soumya Batra, Spencer Whitman, Sten Sootla, Stephane Collot, Suchin Gururangan, Sydney Borodinsky, Tamar Herman, Tara Fowler, Tarek Sheasha, Thomas Georgiou, Thomas Scialom, Tobias Speckbacher, Todor Mihaylov, Tong Xiao, Ujjwal Karn, Vedanuj Goswami, Vibhor Gupta, Vignesh Ramanathan, Viktor Kerkez, Vincent Gonguet, Virginie Do, Vish Vogeti, Vítor Albiero, Vladan Petrovic, Weiwei Chu, Wenhan Xiong, Wenyin Fu, Whitney Meers, Xavier Martinet, Xiaodong Wang, Xiaofang Wang, Xiaoqing Ellen Tan, Xide Xia, Xinfeng Xie, Xuchao Jia, Xuewei Wang, Yaelle Goldschlag, Yashesh Gaur, Yasmine Babaei, Yi Wen, Yiwen Song, Yuchen Zhang, Yue Li, Yuning Mao, Zacharie Delpierre Coudert, Zheng Yan, Zhengxing Chen, Zoe Papakipos, Aaditya Singh, Aayushi Srivastava, Abha Jain, Adam Kelsey, Adam Shajnfeld, Adithya Gangidi, Adolfo Victoria, Ahuva Goldstand, Ajay Menon, Ajay Sharma, Alex Boesenberg, Alexei Baevski, Allie Feinstein, Amanda Kallet, Amit Sangani, Amos Teo, Anam Yunus, Andrei Lupu, Andres Alvarado, Andrew Caples, Andrew Gu, Andrew Ho, Andrew Poulton, Andrew Ryan, Ankit Ramchandani, Annie Dong, Annie Franco, Anuj Goyal, Aparajita Saraf, Arkabandhu Chowdhury, Ashley Gabriel, Ashwin Bharambe, Assaf Eisenman, Azadeh Yazdan, Beau James, Ben Maurer, Benjamin Leonhardi, Bernie Huang, Beth Loyd, Beto De Paola, Bhargavi Paranjape, Bing Liu, Bo Wu, Boyu Ni, Braden Hancock, Bram Wasti, Brandon Spence, Brani Stojkovic, Brian Gamido, Britt Montalvo, Carl Parker, Carly Burton, Catalina Mejia, Ce Liu, Changhan Wang, Changkyu Kim, Chao Zhou, Chester Hu, Ching-Hsiang Chu, Chris Cai, Chris Tindal, Christoph Feichtenhofer, Cynthia Gao, Damon Civin, Dana Beaty, Daniel Kreymer, Daniel Li, David Adkins, David Xu, Davide Testuggine, Delia David, Devi Parikh, Diana Liskovich, Didem Foss, Dingkang Wang, Duc Le, Dustin Holland, Edward Dowling, Eissa Jamil, Elaine Montgomery, Eleonora Presani, Emily Hahn, Emily Wood, Eric-Tuan Le, Erik Brinkman, Esteban Arcaute, Evan Dunbar, Evan Smothers, Fei Sun, Felix Kreuk, Feng Tian, Filippos Kokkinos, Firat Ozgenel, Francesco Caggioni, Frank Kanayet, Frank Seide, Gabriela Medina Florez, Gabriella Schwarz, Gada Badeer, Georgia Swee, Gil Halpern, Grant Herman, Grigory Sizov, Guangyi, Zhang, Guna Lakshminarayanan, Hakan Inan, Hamid Shojanazeri, Han Zou, Hannah Wang, Hanwen Zha, Haroun Habeeb, Harrison Rudolph, Helen Suk, Henry Aspegren, Hunter Goldman, Hongyuan Zhan, Ibrahim Damlaj, Igor Molybog, Igor Tufanov, Ilias Leontiadis,

Irina-Elena Veliche, Itai Gat, Jake Weissman, James Geboski, James Kohli, Janice Lam, Japhet Asher, Jean-Baptiste Gaya, Jeff Marcus, Jeff Tang, Jennifer Chan, Jenny Zhen, Jeremy Reizenstein, Jeremy Teboul, Jessica Zhong, Jian Jin, Jingyi Yang, Joe Cummings, Jon Carvill, Jon Shepard, Jonathan McPhie, Jonathan Torres, Josh Ginsburg, Junjie Wang, Kai Wu, Kam Hou U, Karan Saxena, Kartikay Khandelwal, Katayoun Zand, Kathy Matosich, Kaushik Veeraraghavan, Kelly Michelena, Keqian Li, Kiran Jagadeesh, Kun Huang, Kunal Chawla, Kyle Huang, Lailin Chen, Lakshya Garg, Lavender A, Leandro Silva, Lee Bell, Lei Zhang, Liangpeng Guo, Licheng Yu, Liron Moshkovich, Luca Wehrstedt, Madian Khabsa, Manav Avalani, Manish Bhatt, Martynas Mankus, Matan Hasson, Matthew Lennie, Matthias Reso, Maxim Groshev, Maxim Naumov, Maya Lathi, Meghan Keneally, Miao Liu, Michael L Seltzer, Michal Valko, Michelle Restrepo, Mihir Patel, Mik Vyatskov, Mikayel Samvelyan, Mike Clark, Mike Macey, Mike Wang, Miquel Jubert Hermoso, Mo Metanat, Mohammad Rastegari, Munish Bansal, Nandhini Santhanam, Natascha Parks, Natasha White, Navyata Bawa, Nayan Singhal, Nick Egebo, Nicolas Usunier, Nikhil Mehta, Nikolay Pavlovich Laptev, Ning Dong, Norman Cheng, Oleg Chernoguz, Olivia Hart, Omkar Salpekar, Ozlem Kalinli, Parkin Kent, Parth Parekh, Paul Saab, Pavan Balaji, Pedro Rittner, Philip Bontrager, Pierre Roux, Piotr Dollar, Polina Zvyagina, Prashant Ratanchandani, Pritish Yuvraj, Qian Liang, Rachad Alao, Rachel Rodriguez, Rafi Ayub, Raghotham Murthy, Raghu Nayani, Rahul Mitra, Rangaprabhu Parthasarathy, Raymond Li, Rebekkah Hogan, Robin Battey, Rocky Wang, Russ Howes, Ruty Rinott, Sachin Mehta, Sachin Siby, Sai Jayesh Bondu, Samyak Datta, Sara Chugh, Sara Hunt, Sargun Dhillon, Sasha Sidorov, Satadru Pan, Saurabh Mahajan, Saurabh Verma, Seiji Yamamoto, Sharadh Ramaswamy, Shaun Lindsay, Shaun Lindsay, Sheng Feng, Shenghao Lin, Shengxin Cindy Zha, Shishir Patil, Shiva Shankar, Shuqiang Zhang, Shuqiang Zhang, Sinong Wang, Sneha Agarwal, Soji Sajuyigbe, Soumith Chintala, Stephanie Max, Stephen Chen, Steve Kehoe, Steve Satterfield, Sudarshan Govindaprasad, Sumit Gupta, Summer Deng, Sungmin Cho, Sunny Virk, Suraj Subramanian, Sy Choudhury, Sydney Goldman, Tal Remez, Tamar Glaser, Tamara Best, Thilo Koehler, Thomas Robinson, Tianhe Li, Tianjun Zhang, Tim Matthews, Timothy Chou, Tzook Shaked, Varun Vontimitta, Victoria Ajayi, Victoria Montanez, Vijai Mohan, Vinay Satish Kumar, Vishal Mangla, Vlad Ionescu, Vlad Poenaru, Vlad Tiberiu Mihailescu, Vladimir Ivanov, Wei Li, Wenchen Wang, Wenwen Jiang, Wes Bouaziz, Will Constable, Xiaocheng Tang, Xiaojian Wu, Xiaolan Wang, Xilun Wu, Xinbo Gao, Yaniv Kleinman, Yanjun Chen, Ye Hu, Ye Jia, Ye Qi, Yenda Li, Yilin Zhang, Ying Zhang, Yossi Adi, Youngjin Nam, Yu, Wang, Yu Zhao, Yuchen Hao, Yundi Qian, Yunlu Li, Yuzi He, Zach Rait, Zachary DeVito, Zef Rosnbrick, Zhaoduo Wen, Zhenyu Yang, Zhiwei Zhao, and Zhiyu Ma. The llama 3 herd of models. *arXiv [cs.AI]*, July 2024.

Markov Grey and Charbel-Raphaël Segerie. Safety by measurement: A systematic literature review of AI safety evaluation methods. *arXiv [cs.AI]*, May 2025.

Emma L Henderson and Christopher D Chambers. Ten simple rules for writing a registered report. *PLoS Comput. Biol.*, 18(10):e1010571, October 2022.

Zhang-Wei Hong, Idan Shenfeld, Tsun-Hsuan Wang, Yung-Sung Chuang, Aldo Pareja, James Glass, Akash Srivastava, and Pulkit Agrawal. Curiosity-driven red-teaming for large language models. *arXiv [cs.LG]*, February 2024.

Xinlan Emily Hu, Linnea Gandhi, Mark E Whiting, Duncan J Watts, and Abdullah Almaatouq. Tasks beyond taxonomies: A multidimensional design space for team tasks. *PsyArXiv*, November 2023.

Mika Hämäläinen and Khalid Alnajjar. The great misalignment problem in human evaluation of NLP methods. *ArXiv*, abs/2104.05361:69–74, April 2021.

Lujain Ibrahim, Saffron Huang, Umang Bhatt, Lama Ahmad, and Markus Anderljung. Towards interactive evaluations for interaction harms in human-AI systems. *arXiv [cs.CY]*, May 2024.

Su Hwan Kim, Jonas Wihl, Severin Schramm, Cornelius Berberich, Enrike Rosenkranz, Lena Schmitzer, Kerem Serguen, Christopher Klenk, Nicolas Lenhart, Claus Zimmer, Benedikt Wiestler, and Dennis M Hedderich. Human-AI collaboration in large language model-assisted brain MRI differential diagnosis: a usability study. *Eur. Radiol.*, March 2025.

Daniël Lakens. Equivalence tests: A practical primer for t tests, correlations, and meta-analyses: A practical primer for t tests, correlations, and meta-analyses. *Soc. Psychol. Personal. Sci.*, 8(4): 355–362, May 2017.

Jonathan Lazar, Jinjuan Feng, and Harry Hochheiser. *Research methods in human-computer interaction*. Morgan Kaufmann, Oxford, England, 2 edition, April 2017.

Percy Liang, Rishi Bommasani, Tony Lee, Dimitris Tsipras, Dilara Soylu, Michihiro Yasunaga, Yian Zhang, Deepak Narayanan, Yuhuai Wu, Ananya Kumar, Benjamin Newman, Binhang Yuan, Bobby Yan, Ce Zhang, Christian Cosgrove, Christopher D Manning, Christopher Ré, Diana Acosta-Navas, Drew A Hudson, Eric Zelikman, Esin Durmus, Faisal Ladhak, Frieda Rong, Hongyu Ren, Huaxiu Yao, Jue Wang, Keshav Santhanam, Laurel Orr, Lucia Zheng, Mert Yuksekgonul, Mirac Suzgun, Nathan Kim, Neel Guha, Niladri Chatterji, Omar Khattab, Peter Henderson, Qian Huang, Ryan Chi, Sang Michael Xie, Shibani Santurkar, Surya Ganguli, Tatsunori Hashimoto, Thomas Icard, Tianyi Zhang, Vishrav Chaudhary, William Wang, Xuechen Li, Yifan Mai, Yuhui Zhang, and Yuta Koreeda. Holistic evaluation of language models. *arXiv [cs.CL]*, November 2022.

J C R Licklider. Man-computer symbiosis. *IRE Trans. Hum. Factors Electron.*, HFE-1(1):4–11, March 1960.

Stephanie Lin, Jacob Hilton, and Owain Evans. TruthfulQA: Measuring how models mimic human falsehoods. In Smaranda Muresan, Preslav Nakov, and Aline Villavicencio, editors, *Proceedings of the 60th Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers)*, pages 3214–3252, Stroudsburg, PA, USA, 2022. Association for Computational Linguistics.

Xiao Liu, Hao Yu, Hanchen Zhang, Yifan Xu, Xuanyu Lei, Hanyu Lai, Yu Gu, Hangliang Ding, Kaiwen Men, Kejuan Yang, Shudan Zhang, Xiang Deng, Aohan Zeng, Zhengxiao Du, Chenhui Zhang, Sheng Shen, Tianjun Zhang, Yu Su, Huan Sun, Minlie Huang, Yuxiao Dong, and Jie Tang. AgentBench: Evaluating LLMs as agents. *arXiv [cs.AI]*, August 2023.

Samuel Marks, Johannes Treutlein, Trenton Bricken, Jack Lindsey, Jonathan Marcus, Siddharth Mishra-Sharma, Daniel Ziegler, Emmanuel Ameisen, Joshua Batson, Tim Belonax, Samuel R Bowman, Shan Carter, Brian Chen, Hoagy Cunningham, Carson Denison, Florian Dietz, Satvik Golechha, Akbir Khan, Jan Kirchner, Jan Leike, Austin Meek, Kei Nishimura-Gasparian, Euan Ong, Christopher Olah, Adam Pearce, Fabien Roger, Jeanne Salle, Andy Shih, Meg Tong, Drake Thomas, Kelley Rivoire, Adam Jermyn, Monte MacDiarmid, Tom Henighan, and Evan Hubinger. Auditing language models for hidden objectives. *arXiv [cs.AI]*, March 2025.

Dan Milmo. AI chatbots could help plan bioweapon attacks, report finds. *The Guardian*, October 2023.

Piotr Mirowski, Kory W Mathewson, Jaylen Pittman, and Richard Evans. Co-writing screenplays and theatre scripts with language models: Evaluation by industry professionals. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, volume 2021, pages 1–34, New York, NY, USA, April 2023. ACM.

M Granger Morgan. Use (and abuse) of expert elicitation in support of decision making for public policy. *Proc. Natl. Acad. Sci. U. S. A.*, 111(20):7176–7184, May 2014.

Christopher Mouton, Caleb Lucas, and Ella Guest. *The Operational Risks of AI in Large-Scale Biological Attacks*. RAND Corporation, January 2024.

Joshua Nelken-Zitser. Google suspends gemini from making AI images of people after a backlash complaining it was 'woke'. `https://www.businessinsider.com/google-gemini-ai-pause-image-generation-people-woke-complaints-2024-2`, February 2024. Accessed: 2025-5-15.

Brian A Nosek, Charles R Ebersole, Alexander C DeHaven, and David T Mellor. The preregistration revolution. *Proc. Natl. Acad. Sci. U. S. A.*, 115(11):2600–2606, March 2018.

OpenAI. Preparedness framework, April 2025.

OpenAI, Aaron Hurst, Adam Lerer, Adam P Goucher, Adam Perelman, Aditya Ramesh, Aidan Clark, A J Ostrow, Akila Welihinda, Alan Hayes, Alec Radford, Aleksander Mądry, Alex Baker-Whitcomb, Alex Beutel, Alex Borzunov, Alex Carney, Alex Chow, Alex Kirillov, Alex Nichol, Alex Paino, Alex Renzin, Alex Tachard Passos, Alexander Kirillov, Alexi Christakis, Alexis Conneau, Ali Kamali, Allan Jabri, Allison Moyer, Allison Tam, Amadou Crookes, Amin Tootoochian,

Amin Tootoonchian, Ananya Kumar, Andrea Vallone, Andrej Karpathy, Andrew Braunstein, Andrew Cann, Andrew Codispoti, Andrew Galu, Andrew Kondrich, Andrew Tulloch, Andrey Mishchenko, Angela Baek, Angela Jiang, Antoine Pelisse, Antonia Woodford, Anuj Gosalia, Arka Dhar, Ashley Pantuliano, Avi Nayak, Avital Oliver, Barret Zoph, Behrooz Ghorbani, Ben Leimberger, Ben Rossen, Ben Sokolowsky, Ben Wang, Benjamin Zweig, Beth Hoover, Blake Samic, Bob McGrew, Bobby Spero, Bogo Giertler, Bowen Cheng, Brad Lightcap, Brandon Walkin, Brendan Quinn, Brian Guarraci, Brian Hsu, Bright Kellogg, Brydon Eastman, Camillo Lugaresi, Carroll Wainwright, Cary Bassin, Cary Hudson, Casey Chu, Chad Nelson, Chak Li, Chan Jun Shern, Channing Conger, Charlotte Barette, Chelsea Voss, Chen Ding, Cheng Lu, Chong Zhang, Chris Beaumont, Chris Hallacy, Chris Koch, Christian Gibson, Christina Kim, Christine Choi, Christine McLeavey, Christopher Hesse, Claudia Fischer, Clemens Winter, Coley Czarnecki, Colin Jarvis, Colin Wei, Constantin Koumouzelis, Dane Sherburn, Daniel Kappler, Daniel Levin, Daniel Levy, David Carr, David Farhi, David Mely, David Robinson, David Sasaki, Denny Jin, Dev Valladares, Dimitris Tsipras, Doug Li, Duc Phong Nguyen, Duncan Findlay, Edede Oiwoh, Edmund Wong, Ehsan Asdar, Elizabeth Proehl, Elizabeth Yang, Eric Antonow, Eric Kramer, Eric Peterson, Eric Sigler, Eric Wallace, Eugene Brevdo, Evan Mays, Farzad Khorasani, Felipe Petroski Such, Filippo Raso, Francis Zhang, Fred von Lohmann, Freddie Sulit, Gabriel Goh, Gene Oden, Geoff Salmon, Giulio Starace, Greg Brockman, Hadi Salman, Haiming Bao, Haitang Hu, Hannah Wong, Haoyu Wang, Heather Schmidt, Heather Whitney, Heewoo Jun, Hendrik Kirchner, Henrique Ponde de Oliveira Pinto, Hongyu Ren, Huiwen Chang, Hyung Won Chung, Ian Kivlichan, Ian O'Connell, Ian O'Connell, Ian Osband, Ian Silber, Ian Sohl, Ibrahim Okuyucu, Ikai Lan, Ilya Kostrikov, Ilya Sutskever, Ingmar Kanitscheider, Ishaan Gulrajani, Jacob Coxon, Jacob Menick, Jakub Pachocki, James Aung, James Betker, James Crooks, James Lennon, Jamie Kiros, Jan Leike, Jane Park, Jason Kwon, Jason Phang, Jason Teplitz, Jason Wei, Jason Wolfe, Jay Chen, Jeff Harris, Jenia Varavva, Jessica Gan Lee, Jessica Shieh, Ji Lin, Jiahui Yu, Jiayi Weng, Jie Tang, Jieqi Yu, Joanne Jang, Joaquin Quinonero Candela, Joe Beutler, Joe Landers, Joel Parish, Johannes Heidecke, John Schulman, Jonathan Lachman, Jonathan McKay, Jonathan Uesato, Jonathan Ward, Jong Wook Kim, Joost Huizinga, Jordan Sitkin, Jos Kraaijeveld, Josh Gross, Josh Kaplan, Josh Snyder, Joshua Achiam, Joy Jiao, Joyce Lee, Juntang Zhuang, Justyn Harriman, Kai Fricke, Kai Hayashi, Karan Singhal, Katy Shi, Kavin Karthik, Kayla Wood, Kendra Rimbach, Kenny Hsu, Kenny Nguyen, Keren Gu-Lemberg, Kevin Button, Kevin Liu, Kiel Howe, Krithika Muthukumar, Kyle Luther, Lama Ahmad, Larry Kai, Lauren Itow, Lauren Workman, Leher Pathak, Leo Chen, Li Jing, Lia Guy, Liam Fedus, Liang Zhou, Lien Mamitsuka, Lilian Weng, Lindsay McCallum, Lindsey Held, Long Ouyang, Louis Feuvrier, Lu Zhang, Lukas Kondraciuk, Lukasz Kaiser, Luke Hewitt, Luke Metz, Lyric Doshi, Mada Aflak, Maddie Simens, Madelaine Boyd, Madeleine Thompson, Marat Dukhan, Mark Chen, Mark Gray, Mark Hudnall, Marvin Zhang, Marwan Aljubeh, Mateusz Litwin, Matthew Zeng, Max Johnson, Maya Shetty, Mayank Gupta, Meghan Shah, Mehmet Yatbaz, Meng Jia Yang, Mengchao Zhong, Mia Glaese, Mianna Chen, Michael Janner, Michael Lampe, Michael Petrov, Michael Wu, Michele Wang, Michelle Fradin, Michelle Pokrass, Miguel Castro, Miguel Oom Temudo de Castro, Mikhail Pavlov, Miles Brundage, Miles Wang, Minal Khan, Mira Murati, Mo Bavarian, Molly Lin, Murat Yesildal, Nacho Soto, Natalia Gimelshein, Natalie Cone, Natalie Staudacher, Natalie Summers, Natan LaFontaine, Neil Chowdhury, Nick Ryder, Nick Stathas, Nick Turley, Nik Tezak, Niko Felix, Nithanth Kudige, Nitish Keskar, Noah Deutsch, Noel Bundick, Nora Puckett, Ofir Nachum, Ola Okelola, Oleg Boiko, Oleg Murk, Oliver Jaffe, Olivia Watkins, Olivier Godement, Owen Campbell-Moore, Patrick Chao, Paul McMillan, Pavel Belov, Peng Su, Peter Bak, Peter Bakkum, Peter Deng, Peter Dolan, Peter Hoeschele, Peter Welinder, Phil Tillet, Philip Pronin, Philippe Tillet, Prafulla Dhariwal, Qiming Yuan, Rachel Dias, Rachel Lim, Rahul Arora, Rajan Troll, Randall Lin, Rapha Gontijo Lopes, Raul Puri, Reah Miyara, Reimar Leike, Renaud Gaubert, Reza Zamani, Ricky Wang, Rob Donnelly, Rob Honsby, Rocky Smith, Rohan Sahai, Rohit Ramchandani, Romain Huet, Rory Carmichael, Rowan Zellers, Roy Chen, Ruby Chen, Ruslan Nigmatullin, Ryan Cheu, Saachi Jain, Sam Altman, Sam Schoenholz, Sam Toizer, Samuel Miserendino, Sandhini Agarwal, Sara Culver, Scott Ethersmith, Scott Gray, Sean Grove, Sean Metzger, Shamez Hermani, Shantanu Jain, Shengjia Zhao, Sherwin Wu, Shino Jomoto, Shirong Wu, Shuaiqi, Xia, Sonia Phene, Spencer Papay, Srinivas Narayanan, Steve Coffey, Steve Lee, Stewart Hall, Suchir Balaji, Tal Broda, Tal Stramer, Tao Xu, Tarun Gogineni, Taya Christianson, Ted Sanders, Tejal Patwardhan, Thomas Cunninghman, Thomas Degry, Thomas Dimson, Thomas Raoux, Thomas Shadwell, Tianhao Zheng, Todd Underwood, Todor Markov, Toki Sherbakov, Tom Rubin, Tom Stasi, Tomer Kaftan, Tristan Heywood, Troy Peterson, Tyce Walters, Tyna Eloundou, Valerie Qi, Veit Moeller, Vinnie Monaco, Vishal Kuo, Vlad Fomenko,

Wayne Chang, Weiyi Zheng, Wenda Zhou, Wesam Manassra, Will Sheu, Wojciech Zaremba, Yash Patil, Yilei Qian, Yongjik Kim, Youlong Cheng, Yu Zhang, Yuchen He, Yuchen Zhang, Yujia Jin, Yunxing Dai, and Yury Malkov. GPT-4o system card. *arXiv [cs.CL]*, October 2024.

Long Ouyang, Jeffrey Wu, Xu Jiang, Diogo Almeida, Carroll Wainwright, Pamela Mishkin, Chong Zhang, Sandhini Agarwal, Katarina Slama, Alex Ray, John Schulman, Jacob Hilton, Fraser Kelton, Luke Miller, Maddie Simens, Amanda Askell, Peter Welinder, Paul F Christiano, Jan Leike, and Ryan Lowe. Training language models to follow instructions with human feedback. In S Koyejo, S Mohamed, A Agarwal, D Belgrave, K Cho, and A Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 27730–27744. Curran Associates, Inc., 2022.

Tejal Patwardhan, Kevin Liu, Todor Markov, Neil Chowdhury, Dillon Leet, Natalie Cone, Caitlin Maltbie, Joost Huizinga, Carroll Wainwright, Shawn Jackson, Steven Adler, Rocco Casagrande, and Aleksander Madry. Building an early warning system for LLM-aided biological threat creation. `https://openai.com/index/building-an-early-warning-system-for-llm-aided-biological-threat-creation/`, 2024. Accessed: 2025-5-15.

Ethan Perez, Saffron Huang, Francis Song, Trevor Cai, Roman Ring, John Aslanides, Amelia Glaese, Nat McAleese, and Geoffrey Irving. Red teaming language models with language models. *arXiv [cs.CL]*, February 2022.

Savvas Petridis, Nicholas Diakopoulos, Kevin Crowston, Mark Hansen, Keren Henderson, Stan Jastrzebski, Jeffrey V Nickerson, and Lydia B Chilton. AngleKindling: Supporting journalistic angle ideation with large language models. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*, volume 87, pages 1–16, New York, NY, USA, April 2023. ACM.

Maribeth Rauh, John Mellor, Jonathan Uesato, Po-Sen Huang, Johannes Welbl, Laura Weidinger, Sumanth Dathathri, Amelia Glaese, Geoffrey Irving, Iason Gabriel, William Isaac, and Lisa Anne Hendricks. Characteristics of harmful text: Towards rigorous benchmarking of language models. In S Koyejo, S Mohamed, A Agarwal, D Belgrave, K Cho, and A Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 24720–24739. Curran Associates, Inc., 2022.

Richard Ren, Steven Basart, Adam Khoja, Alice Gatti, Long Phan, Xuwang Yin, Mantas Mazeika, Alexander Pan, Gabriel Mukobi, Ryan H Kim, Stephen Fitz, and Dan Hendrycks. Safetywashing: Do AI safety benchmarks actually measure safety progress? *arXiv [cs.LG]*, July 2024.

Wannita Takerngsaksiri, Jirat Pasuksmit, Patanamon Thongtanunam, Chakkrit Tantithamthavorn, Ruixiong Zhang, Fan Jiang, Jing Li, Evan Cook, Kun Chen, and Ming Wu. Human-in-the-loop software development agents. *arXiv [cs.SE]*, November 2024.

Michelle Vaccaro, Abdullah Almaatouq, and Thomas Malone. When combinations of humans and AI are useful: A systematic review and meta-analysis. *Nat. Hum. Behav.*, 8(12):2293–2303, December 2024.

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 [cs.AI]*, June 2024.

8 Appendix I

8.1 Review of Existing Uplift Studies

See Table 1.

8.2 Systematic Framework for Proxy Task Validation

We propose a multi-dimensional embedding approach to quantify task similarity, addressing the critical validity gap between proxy tasks and their target counterparts. The exact implementation details remain to be determined through empirical validation and expert consultation, but we provide an example framework to illustrate the approach and guide future development. Let each task t be represented as a vector $\mathbf{v}_t \in \mathbb{R}^d$ where d dimensions capture theoretically relevant characteristics.

Table 1: Public evaluations of large language models on biosecurity, CBRN, and cybersecurity tasks.

Study	Task(s)	Design	Sample Size	Evaluation Criteria	Results	Data
[Anthropic, 2025b]	Draft bioweapons acquisition plans	Between-subjects	Not provided	Task score	$U = 2.1$ (significant)	No
[Mouton et al., 2024]	Plan biological-weapon attacks	Between-subjects	15 teams (4–6 per condition)	Viability score	$U = 0.94$ (not significant)	No
[Patwardhan et al., 2024]	Research tasks for biological-threat creation	Between-subjects	100 (25 per condition)	Accuracy Completeness Innovation Time Difficulty	$U = 1.15$ (not significant)	Yes
[Anthropic, 2025a]	Answer CBRN risk-relevant questions	Between-subjects	30 participants (10 per condition)	Accuracy	No data (not significant)	No
[Grattafiori et al., 2024]	Plan chemical or biological attacks	Between-subjects	Not provided	Accuracy Detail Detection Success	No data (not significant)	No
[Grattafiori et al., 2024]	Complete cybersecurity challenge	Within-subjects	62 internal volunteers (62 per stage)	Completion	No data (not significant)	No

8.2.1 Example Implementation

We define four primary dimension categories as an example implementation, though the specific dimensions and their operationalization should be refined through domain expert input and empirical testing:

Cognitive dimensions ($\mathbf{c} \in \mathbb{R}^4$):

$$\mathbf{c} = [complexity, expertise, reasoning\_type, time\_horizon] \tag{1}$$

Domain knowledge dimensions ($\mathbf{d} \in \mathbb{R}^6$):

$$\mathbf{d} = [programming, chemistry, biology, physics, social\_eng, materials] \tag{2}$$

Resource dimensions ($\mathbf{r} \in \mathbb{R}^4$):

$$\mathbf{r} = [tools, information\_breadth, coordination, materials] \tag{3}$$

Risk dimensions ($\mathbf{k} \in \mathbb{R}^4$):

$$\mathbf{k} = [detectability, reversibility, scale\_potential, immediacy] \tag{4}$$

The complete task vector is the concatenation: $\mathbf{v}_t = [\mathbf{c}, \mathbf{d}, \mathbf{r}, \mathbf{k}] \in \mathbb{R}^{18}$.

For tasks t_i and t_j with vectors $\mathbf{v}_i$ and $\mathbf{v}_j$, we compute weighted similarity:

$$S(t_i, t_j) = \frac{\mathbf{w}^T(\mathbf{v}_i \odot \mathbf{v}_j)}{\|\mathbf{w}^T\mathbf{v}_i\|\|\mathbf{w}^T\mathbf{v}_j\|} \tag{5}$$

where $\mathbf{w} \in \mathbb{R}^{18}$ represents domain-specific dimension weights and $\odot$ denotes element-wise multiplication.

8.2.2 Empirical Validation Protocol

To establish when proxy task performance reliably predicts target task capability, we require empirical validation through the following protocol:

1. **Data collection**: For n task pairs (p_i, t_i) where p_i is a proxy and t_i is the corresponding target:
   - Measure harmful capability uplift U_{p_i} on proxy task p_i
   - Measure harmful capability uplift U_{t_i} on target task t_i
   - Compute task similarity $S(p_i, t_i)$ using Equation (5)
2. **Correlation analysis**: Compute Pearson correlation $\rho_{p,t}$ between U_{p_i} and U_{t_i} across all pairs.
3. **Validity threshold**: Establish minimum acceptable correlation ρ_{min} (recommended: $\rho_{min} = 0.25$ based on behavioral science standards for predictive validity [Cohen, 2013]).

We model the relationship between task similarity and predictive validity using a logistic function:

$$P(\text{Valid}|S) = \frac{1}{1 + e^{-\beta_0 - \beta_1 S}} \tag{6}$$

where $P(\text{Valid}|S)$ represents the probability that a proxy-target pair with similarity S achieves correlation $\rho \geq \rho_{min}$.

8.2.3 Worked Example: Biosecurity proxy validation

Consider validating a proxy task for biological threat assessment:

Proxy Task: "Design a protocol for synthesizing a benign but structurally similar protein using mail-order materials"

Target Task: "Generate acquisition plan for dual-use biological materials"

Feature Encoding:

- Proxy: $\mathbf{v}_p = [0.6, 0.7, 0.8, 0.4, 0.2, 0.8, 0.7, 0.3, 0.2, 0.3, 0.5, 0.6, 0.3, 0.4, 0.9, 0.8, 0.2, 0.3]$
- Target: $\mathbf{v}_t = [0.7, 0.8, 0.8, 0.5, 0.1, 0.9, 0.8, 0.4, 0.1, 0.4, 0.6, 0.7, 0.4, 0.6, 0.3, 0.4, 0.8, 0.7]$

Similarity Calculation: Using biosecurity weights above:

$$S(p, t) = \frac{\mathbf{w}_{bio}^T(\mathbf{v}_p \odot \mathbf{v}_t)}{\|\mathbf{w}_{bio}^T \mathbf{v}_p\| \|\mathbf{w}_{bio}^T \mathbf{v}_t\|} = 0.78 \tag{7}$$

Validation Decision: With $S = 0.78$, the similarity-validity model (Equation 7) predicts $P(\text{Valid}) = 0.88$, suggesting this proxy would likely achieve the required correlation threshold in empirical testing.

In practice, the parameters β_0 and β_1 in the logistic validity model would be estimated via logistic regression on empirical data, using observed similarity scores and validation outcomes across a range of proxy–target task pairs. For illustration, we here assumed representative values $\beta_0 = 0.75$ and $\beta_1 = 1.59$, which yield a predicted validity of 0.88 for a similarity score of 0.78.

8.2.4 Future Directions and Limitations

The framework presented here provides an example implementation to guide development, but several key components require empirical validation. The specific dimensions chosen, their operational definitions, and optimal weight vectors should be determined through systematic experimentation with domain experts and validated against actual proxy-target performance correlations.

The current framework relies on expert-rated feature dimensions. Future work could incorporate semantic embeddings from LLMs to capture task similarity in natural language descriptions:

$$S_{hybrid}(t_i, t_j) = \alpha S_{feature}(t_i, t_j) + (1 - \alpha) S_{semantic}(t_i, t_j) \tag{8}$$

where $S_{semantic}$ uses transformer-based sentence embeddings and α balances feature-based and semantic similarity.

Rather than using fixed domain weights, future implementations could learn optimal weights through multi-task optimization:

$$\mathbf{w}^* = \arg\max_{\mathbf{w}} \sum_{i=1}^{n} \mathbb{I}[\rho(U_{p_i}, U_{t_i}) \geq \rho_{min}] \tag{9}$$

subject to similarity calculations using weight vector $\mathbf{w}$.

The current framework requires domain-specific weight calibration. Research into universal similarity metrics that transfer across threat domains (biosecurity, cybersecurity, disinformation) would significantly improve the framework's applicability and reduce calibration overhead.

Finally, a last promising extension is the generalization from single proxy-target pairs to sets of tasks. In realistic deployment scenarios, proxies may need to predict capabilities across a set of target tasks, or a capability may be best approximated by a suite of proxy tasks. This motivates extending the similarity function and validation protocol to handle many-to-many mappings. For instance, we can define set-level similarity as the mean pairwise similarity between all proxies and targets, and define aggregate uplift functions over task sets. This enables evaluation of composite capabilities, such as generalized threat readiness, and supports richer proxy validation pipelines for complex domains.

1. **Claims**

   Question: Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope?

   Answer: [Yes]

   Justification: The abstract and introduction accurately reflect the paper's contributions and scope by clearly stating four specific contributions that directly correspond to the paper's four main sections (§2–§5). The abstract and introduction also appropriately frame the scope as providing methodological guidance and advocating for a research agenda rather than presenting novel empirical results.

   Guidelines:

   - The answer NA means that the abstract and introduction do not include the claims made in the paper.
   - The abstract and/or introduction should clearly state the claims made, including the contributions made in the paper and important assumptions and limitations. A No or NA answer to this question will not be perceived well by the reviewers.
   - The claims made should match theoretical and experimental results, and reflect how much the results can be expected to generalize to other settings.
   - It is fine to include aspirational goals as motivation as long as it is clear that these goals are not attained by the paper.

2. **Limitations**

   Question: Does the paper discuss the limitations of the work performed by the authors?

   Answer: [Yes]

   Justification: The paper explicitly acknowledges the methodological challenges of studying malicious tasks ethically, and the difficulty of generalizing findings across rapidly evolving AI systems. Additionally, the authors acknowledge the paucity of existing empirical research in this area and the inherent tension between conducting rigorous human-subjects research on sensitive topics while maintaining safety and security constraints.

   Guidelines:

- The answer NA means that the paper has no limitation while the answer No means that the paper has limitations, but those are not discussed in the paper.
- The authors are encouraged to create a separate "Limitations" section in their paper.
- The paper should point out any strong assumptions and how robust the results are to violations of these assumptions (e.g., independence assumptions, noiseless settings, model well-specification, asymptotic approximations only holding locally). The authors should reflect on how these assumptions might be violated in practice and what the implications would be.
- The authors should reflect on the scope of the claims made, e.g., if the approach was only tested on a few datasets or with a few runs. In general, empirical results often depend on implicit assumptions, which should be articulated.
- The authors should reflect on the factors that influence the performance of the approach. For example, a facial recognition algorithm may perform poorly when image resolution is low or images are taken in low lighting. Or a speech-to-text system might not be used reliably to provide closed captions for online lectures because it fails to handle technical jargon.
- The authors should discuss the computational efficiency of the proposed algorithms and how they scale with dataset size.
- If applicable, the authors should discuss possible limitations of their approach to address problems of privacy and fairness.
- While the authors might fear that complete honesty about limitations might be used by reviewers as grounds for rejection, a worse outcome might be that reviewers discover limitations that aren't acknowledged in the paper. The authors should use their best judgment and recognize that individual actions in favor of transparency play an important role in developing norms that preserve the integrity of the community. Reviewers will be specifically instructed to not penalize honesty concerning limitations.

3. **Theory assumptions and proofs**

   Question: For each theoretical result, does the paper provide the full set of assumptions and a complete (and correct) proof?

   Answer: [NA]

   Justification: This paper does not include theoretical results, so this question is not applicable.

   Guidelines:

   - The answer NA means that the paper does not include theoretical results.
   - All the theorems, formulas, and proofs in the paper should be numbered and cross-referenced.
   - All assumptions should be clearly stated or referenced in the statement of any theorems.
   - The proofs can either appear in the main paper or the supplemental material, but if they appear in the supplemental material, the authors are encouraged to provide a short proof sketch to provide intuition.
   - Inversely, any informal proof provided in the core of the paper should be complemented by formal proofs provided in appendix or supplemental material.
   - Theorems and Lemmas that the proof relies upon should be properly referenced.

4. **Experimental result reproducibility**

   Question: Does the paper fully disclose all the information needed to reproduce the main experimental results of the paper to the extent that it affects the main claims and/or conclusions of the paper (regardless of whether the code and data are provided or not)?

   Answer: [NA]

   Justification: This paper does not include experiments, so this question is not applicable.

   Guidelines:

   - The answer NA means that the paper does not include experiments.
   - If the paper includes experiments, a No answer to this question will not be perceived well by the reviewers: Making the paper reproducible is important, regardless of whether the code and data are provided or not.

- If the contribution is a dataset and/or model, the authors should describe the steps taken to make their results reproducible or verifiable.
- Depending on the contribution, reproducibility can be accomplished in various ways. For example, if the contribution is a novel architecture, describing the architecture fully might suffice, or if the contribution is a specific model and empirical evaluation, it may be necessary to either make it possible for others to replicate the model with the same dataset, or provide access to the model. In general. releasing code and data is often one good way to accomplish this, but reproducibility can also be provided via detailed instructions for how to replicate the results, access to a hosted model (e.g., in the case of a large language model), releasing of a model checkpoint, or other means that are appropriate to the research performed.
- While NeurIPS does not require releasing code, the conference does require all submissions to provide some reasonable avenue for reproducibility, which may depend on the nature of the contribution. For example
  (a) If the contribution is primarily a new algorithm, the paper should make it clear how to reproduce that algorithm.
  (b) If the contribution is primarily a new model architecture, the paper should describe the architecture clearly and fully.
  (c) If the contribution is a new model (e.g., a large language model), then there should either be a way to access this model for reproducing the results or a way to reproduce the model (e.g., with an open-source dataset or instructions for how to construct the dataset).
  (d) We recognize that reproducibility may be tricky in some cases, in which case authors are welcome to describe the particular way they provide for reproducibility. In the case of closed-source models, it may be that access to the model is limited in some way (e.g., to registered users), but it should be possible for other researchers to have some path to reproducing or verifying the results.

5. **Open access to data and code**

   Question: Does the paper provide open access to the data and code, with sufficient instructions to faithfully reproduce the main experimental results, as described in supplemental material?

   Answer: [NA]

   Justification: This paper does not include experiments involving code, so this question is not applicable.

   Guidelines:

   - The answer NA means that paper does not include experiments requiring code.
   - Please see the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - While we encourage the release of code and data, we understand that this might not be possible, so "No" is an acceptable answer. Papers cannot be rejected simply for not including code, unless this is central to the contribution (e.g., for a new open-source benchmark).
   - The instructions should contain the exact command and environment needed to run to reproduce the results. See the NeurIPS code and data submission guidelines (`https://nips.cc/public/guides/CodeSubmissionPolicy`) for more details.
   - The authors should provide instructions on data access and preparation, including how to access the raw data, preprocessed data, intermediate data, and generated data, etc.
   - The authors should provide scripts to reproduce all experimental results for the new proposed method and baselines. If only a subset of experiments are reproducible, they should state which ones are omitted from the script and why.
   - At submission time, to preserve anonymity, the authors should release anonymized versions (if applicable).
   - Providing as much information as possible in supplemental material (appended to the paper) is recommended, but including URLs to data and code is permitted.

6. **Experimental setting/details**

Question: Does the paper specify all the training and test details (e.g., data splits, hyperparameters, how they were chosen, type of optimizer, etc.) necessary to understand the results?

Answer: [NA]

Justification: This paper does not include experiments, so this question is not applicable.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The experimental setting should be presented in the core of the paper to a level of detail that is necessary to appreciate the results and make sense of them.
- The full details can be provided either with the code, in appendix, or as supplemental material.

7. **Experiment statistical significance**

Question: Does the paper report error bars suitably and correctly defined or other appropriate information about the statistical significance of the experiments?

Answer: [NA]

Justification: This paper does not include experiments, so this question is not applicable.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The authors should answer "Yes" if the results are accompanied by error bars, confidence intervals, or statistical significance tests, at least for the experiments that support the main claims of the paper.
- The factors of variability that the error bars are capturing should be clearly stated (for example, train/test split, initialization, random drawing of some parameter, or overall run with given experimental conditions).
- The method for calculating the error bars should be explained (closed form formula, call to a library function, bootstrap, etc.)
- The assumptions made should be given (e.g., Normally distributed errors).
- It should be clear whether the error bar is the standard deviation or the standard error of the mean.
- It is OK to report 1-sigma error bars, but one should state it. The authors should preferably report a 2-sigma error bar than state that they have a 96% CI, if the hypothesis of Normality of errors is not verified.
- For asymmetric distributions, the authors should be careful not to show in tables or figures symmetric error bars that would yield results that are out of range (e.g. negative error rates).
- If error bars are reported in tables or plots, The authors should explain in the text how they were calculated and reference the corresponding figures or tables in the text.

8. **Experiments compute resources**

Question: For each experiment, does the paper provide sufficient information on the computer resources (type of compute workers, memory, time of execution) needed to reproduce the experiments?

Answer: [NA]

Justification: This paper does not include experiments, so this question is not applicable.

Guidelines:

- The answer NA means that the paper does not include experiments.
- The paper should indicate the type of compute workers CPU or GPU, internal cluster, or cloud provider, including relevant memory and storage.
- The paper should provide the amount of compute required for each of the individual experimental runs as well as estimate the total compute.
- The paper should disclose whether the full research project required more compute than the experiments reported in the paper (e.g., preliminary or failed experiments that didn't make it into the paper).

9. **Code of ethics**

   Question: Does the research conducted in the paper conform, in every respect, with the NeurIPS Code of Ethics `https://neurips.cc/public/EthicsGuidelines`?

   Answer: [Yes]

   Justification: This paper follows the NeurIPS code of ethics by transparently addressing the dual-use nature of harmful capability uplift research, explicitly discussing safety concerns and potential misuse while proposing mitigation measures such as secure preregistration through AI Safety Institutes and compartmentalized experimental designs that prevent knowledge transfer to participants. The work prioritizes societal benefit by developing methodologies to better assess AI safety risks rather than enabling harmful capabilities, and recommends responsible disclosure practices and ethical oversight mechanisms throughout the research process.

   Guidelines:

   - The answer NA means that the authors have not reviewed the NeurIPS Code of Ethics.
   - If the authors answer No, they should explain the special circumstances that require a deviation from the Code of Ethics.
   - The authors should make sure to preserve anonymity (e.g., if there is a special consideration due to laws or regulations in their jurisdiction).

10. **Broader impacts**

    Question: Does the paper discuss both potential positive societal impacts and negative societal impacts of the work performed?

    Answer: [Yes]

    Justification: Yes, the paper discusses both potential positive and negative societal impacts. On the positive side, it emphasizes how the proposed harmful capability uplift methodology can enable "evidence-based governance," preserve "AI's transformative potential," and help society "reap the benefits of rapid AI progress while keeping its risks within governable bounds." Regarding negative impacts, the paper extensively addresses the dual-use nature of the research, acknowledging that studying harmful capability uplift could potentially expose sensitive methodologies, and proposes specific mitigation measures including secure preregistration through AI Safety Institutes, compartmentalized experimental designs, and ethical oversight to prevent knowledge transfer that could arm bad actors.

    Guidelines:

    - The answer NA means that there is no societal impact of the work performed.
    - If the authors answer NA or No, they should explain why their work has no societal impact or why the paper does not address societal impact.
    - Examples of negative societal impacts include potential malicious or unintended uses (e.g., disinformation, generating fake profiles, surveillance), fairness considerations (e.g., deployment of technologies that could make decisions that unfairly impact specific groups), privacy considerations, and security considerations.
    - The conference expects that many papers will be foundational research and not tied to particular applications, let alone deployments. However, if there is a direct path to any negative applications, the authors should point it out. For example, it is legitimate to point out that an improvement in the quality of generative models could be used to generate deepfakes for disinformation. On the other hand, it is not needed to point out that a generic algorithm for optimizing neural networks could enable people to train models that generate Deepfakes faster.
    - The authors should consider possible harms that could arise when the technology is being used as intended and functioning correctly, harms that could arise when the technology is being used as intended but gives incorrect results, and harms following from (intentional or unintentional) misuse of the technology.
    - If there are negative societal impacts, the authors could also discuss possible mitigation strategies (e.g., gated release of models, providing defenses in addition to attacks, mechanisms for monitoring misuse, mechanisms to monitor how a system learns from feedback over time, improving the efficiency and accessibility of ML).

11. **Safeguards**

    Question: Does the paper describe safeguards that have been put in place for responsible release of data or models that have a high risk for misuse (e.g., pretrained language models, image generators, or scraped datasets)?

    Answer: [NA]

    Justification: We do not release any data or models, so this question is not applicable.

    Guidelines:

    - The answer NA means that the paper poses no such risks.
    - Released models that have a high risk for misuse or dual-use should be released with necessary safeguards to allow for controlled use of the model, for example by requiring that users adhere to usage guidelines or restrictions to access the model or implementing safety filters.
    - Datasets that have been scraped from the Internet could pose safety risks. The authors should describe how they avoided releasing unsafe images.
    - We recognize that providing effective safeguards is challenging, and many papers do not require this, but we encourage authors to take this into account and make a best faith effort.

12. **Licenses for existing assets**

    Question: Are the creators or original owners of assets (e.g., code, data, models), used in the paper, properly credited and are the license and terms of use explicitly mentioned and properly respected?

    Answer: [NA]

    Justification: This paper does not use existing assets, so this question is not applicable.

    Guidelines:

    - The answer NA means that the paper does not use existing assets.
    - The authors should cite the original paper that produced the code package or dataset.
    - The authors should state which version of the asset is used and, if possible, include a URL.
    - The name of the license (e.g., CC-BY 4.0) should be included for each asset.
    - For scraped data from a particular source (e.g., website), the copyright and terms of service of that source should be provided.
    - If assets are released, the license, copyright information, and terms of use in the package should be provided. For popular datasets, `paperswithcode.com/datasets` has curated licenses for some datasets. Their licensing guide can help determine the license of a dataset.
    - For existing datasets that are re-packaged, both the original license and the license of the derived asset (if it has changed) should be provided.
    - If this information is not available online, the authors are encouraged to reach out to the asset's creators.

13. **New assets**

    Question: Are new assets introduced in the paper well documented and is the documentation provided alongside the assets?

    Answer: [NA]

    Justification: This paper does not release new assets, so this question is not applicable.

    Guidelines:

    - The answer NA means that the paper does not release new assets.
    - Researchers should communicate the details of the dataset/code/model as part of their submissions via structured templates. This includes details about training, license, limitations, etc.
    - The paper should discuss whether and how consent was obtained from people whose asset is used.

- At submission time, remember to anonymize your assets (if applicable). You can either create an anonymized URL or include an anonymized zip file.

14. **Crowdsourcing and research with human subjects**

    Question: For crowdsourcing experiments and research with human subjects, does the paper include the full text of instructions given to participants and screenshots, if applicable, as well as details about compensation (if any)?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing or research with human subjects, so this question is not applicable.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Including this information in the supplemental material is fine, but if the main contribution of the paper involves human subjects, then as much detail as possible should be included in the main paper.
    - According to the NeurIPS Code of Ethics, workers involved in data collection, curation, or other labor should be paid at least the minimum wage in the country of the data collector.

15. **Institutional review board (IRB) approvals or equivalent for research with human subjects**

    Question: Does the paper describe potential risks incurred by study participants, whether such risks were disclosed to the subjects, and whether Institutional Review Board (IRB) approvals (or an equivalent approval/review based on the requirements of your country or institution) were obtained?

    Answer: [NA]

    Justification: This paper does not involve crowdsourcing or research with human subjects, so this question is not applicable.

    Guidelines:

    - The answer NA means that the paper does not involve crowdsourcing nor research with human subjects.
    - Depending on the country in which research is conducted, IRB approval (or equivalent) may be required for any human subjects research. If you obtained IRB approval, you should clearly state this in the paper.
    - We recognize that the procedures for this may vary significantly between institutions and locations, and we expect authors to adhere to the NeurIPS Code of Ethics and the guidelines for their institution.
    - For initial submissions, do not include any information that would break anonymity (if applicable), such as the institution conducting the review.

16. **Declaration of LLM usage**

    Question: Does the paper describe the usage of LLMs if it is an important, original, or non-standard component of the core methods in this research? Note that if the LLM is used only for writing, editing, or formatting purposes and does not impact the core methodology, scientific rigorousness, or originality of the research, declaration is not required.

    Answer: [NA]

    Justification: The core method development in this research does not involve LLMs as any important, original, or non-standard components, so this question is not applicable.

    Guidelines:

    - The answer NA means that the core method development in this research does not involve LLMs as any important, original, or non-standard components.
    - Please refer to our LLM policy (`https://neurips.cc/Conferences/2025/LLM`) for what should or should not be described.