Bridging the In-Context AI Evaluation Gap: Risk Chain Analysis for Coordinated AI Governance

Anonymous ACL submission

Abstract

Capability evaluations cover only a narrow portion of the pathway from AI capability to real-world harm. We introduce *risk chain analysis*—tracing harm trajectories step by step and mapping evaluation evidence at each stage—and apply it to three risk domains: cyber attacks, biological risk, and agentic AI failures. Evaluation consistently concentrates where benchmarking is easiest, not where risk is highest. The uncovered steps require different methods and different actors: sectoral regulators, deploying organisations, and domain-specific agencies that currently have no role in AI evaluation. We argue that AI safety institutes are uniquely positioned to coordinate this distributed evaluation ecosystem.

1 Introduction

Capability evaluations remain the dominant paradigm for AI risk assessment, with 85% of evaluations focused at the capability level (Weidinger et al., 2023). Yet they lack construct validity (Schaeffer et al., 2023; Bean et al., 2025; Raji et al., 2021) and ecological validity (Raji et al., 2022; Weidinger et al., 2023), producing systematic over- and underestimation of risk (Hubinger et al., 2024; Passi and Vorvoreanu, 2022).

Despite broad acknowledgement of these limitations, alternative methods have not been widely adopted. We identify three reasons: a *knowledge problem* (alternative methods are unfamiliar), an *operationalisation problem* (they are costly and resist automation), and a *coordination problem* (the required evidence is distributed across the governance ecosystem). This paper addresses the third directly.

We introduce *risk chain analysis*: decomposing specific harm trajectories into sequential steps, mapping evaluation coverage at each, and identifying which actors should fill gaps. We apply this to three structurally different cases—a malicious-actor scenario (cyber), a well-studied domain (bio), and a deployment failure (agentic AI)—and find a consistent pattern: evaluation clusters at early steps while the later steps where harm materialises are systematically uncovered.

2 Three Cases

2.1 Cyber: AI-Enabled Ransomware on a Hospital

Scenario. A criminal group uses AI to conduct a ransomware attack on an NHS hospital trust running legacy systems, grounded in real precedent (Rodriguez et al., 2025) and the WannaCry (2017) and Synnovis (2024) attacks.

We trace six steps: (1) attacker intent and AI access, (2) reconnaissance, (3) vulnerability identification and exploitation, (4) lateral movement, (5) ransomware deployment, and (6) patient harm.

Nearly all evaluation effort concentrates at Step 3: CTF benchmarks (Zhang et al., 2024; Yang et al., 2024), vulnerability databases, and exploit generation tasks. But the ecological validity gap is widest here. The ARTEMIS study (Lin et al., 2025)—the first comparison of AI agents and human pentesters on live infrastructure—found markedly different capability profiles from CTF results. Steps 1 and 4–6 receive essentially no coverage. Critically, outcomes at these steps depend on the *target's* defences, not the model's capabilities. Evidence here requires NHS Digital, the NCSC, and health regulators—organisations with no current role in AI evaluation.

2.2 Biological Risk

Biorisk has received more evaluation attention than any other domain, with multiple uplift studies (Mouton et al., 2024; OpenAI, 2024; Anthropic, 2024). Even here, evaluation concentrates at Step 2 (information access). As Peppin et al. (2024) note,

there are no published empirical studies connecting information access via LLMs to downstream steps: material procurement, wet lab execution, and dissemination. These require biosecurity regulators, DNA synthesis screening bodies, and public health agencies (Carter et al., 2023).

2.3 Agentic AI: Coding Agent Introduces Vulnerability

A coding agent generates a SQL injection vulnerability; the developer approves under time pressure; the vulnerability reaches production (Veracode, 2025; Apiiro, 2025). Partial coverage exists at Step 2 (code generation benchmarks). The most consequential gap is Step 3: human oversight. Every agentic AI governance framework identifies oversight as a core safeguard (Infocomm Media Development Authority, 2025; Robey et al., 2025; Chan et al., 2024), yet nobody has studied whether developers actually provide effective oversight of AI-generated code. The automation bias literature predicts they often will not (Parasuraman and Manzey, 2010; Buçinca et al., 2021), but this remains untested for coding agents specifically. Evidence must come from research institutions and deploying organisations.

3 Cross-Cutting Findings

Evaluation concentrates where benchmarking is easy, not where risk is highest. Cyber clusters at vulnerability identification, bio at information access, agentic AI at code generation. The steps where harm materialises receive little or no attention.

The best-covered steps have the worst ecological validity. CTF benchmarks, knowledge tests, and isolated code generation tasks strip away contextual complexity. ARTEMIS (Lin et al., 2025), Epoch AI's analysis of biorisk evaluation ambiguity, and Kilov et al.'s (2025) work on moral competence evaluation all demonstrate that benchmark performance is unreliable when deployment context is removed.

Gaps cluster around human behaviour and deployment context. Attacker intent, developer oversight, institutional resilience—these are properties of sociotechnical systems, not models. They require usage monitoring (Tamkin et al., 2024), human interaction experiments (Dell'Acqua et al., 2023; Noy and Zhang, 2023), field testing (National Institute of Standards and Technology, 2025), deployment-context assessment, and incident analysis (McGregor, 2021).

Different steps require different actors. AI labs and safety institutes cover Steps 1–3. Everything beyond falls to organisations with no current role in AI evaluation.

4 Who Should Do What

The risk chain framework reveals that a safety case built on capability evaluations alone covers at most one or two steps of a five-to-six step causal chain. Addressing this requires distributing evaluation responsibility—but distribution without coordination risks fragmentation. We argue that AI safety institutes (AISIs) are uniquely positioned to serve as the coordinating node.

4.1 AI Developers

Developers hold unique evidence on real-world usage patterns (Tamkin et al., 2024; Zhao et al., 2024; Zheng et al., 2024). They should invest in ecologically valid evaluations: not just functional benchmarks but operational simulations and security testing in realistic contexts. Sharing aggregated, privacy-preserving usage data—as Anthropic's Clio project has begun to do—would significantly strengthen the evidence base for Steps 1–2 across risk chains.

4.2 Sectoral Regulators and Deployers

Sectoral regulators and deploying organisations own the later steps of every chain. NHS Digital understands hospital network architecture; financial regulators understand trading system risk; software engineering organisations understand code review practices. These actors have domain knowledge for deployment-context assessment but typically lack AI evaluation methodology. Capacity building—through training, shared frameworks, and collaborative evaluation design—is the bridge between model-level evidence and deployment-level risk.

4.3 Safety Institutes as Coordination Nodes

AISIs have three features that make them natural coordinators of the distributed evaluation ecosystem.

First, **methodological comparative advantage**. AISIs concentrate evaluation expertise that no single sectoral regulator can replicate. Their most valuable contribution may not be conducting every evaluation themselves but developing the analytical frameworks—such as risk chain analysis—that

make evaluation gaps visible and tractable across domains. This aligns with the emerging shift in AISI mandates from direct model testing toward supporting evaluation capacity across government (Department for Science, Innovation and Technology, 2024).

Second, **cross-domain visibility**. The risk chains above show structurally similar gaps across very different domains: human oversight is unstudied for both coding agents and biosecurity contexts; ecological validity is poor for both cyber and biorisk benchmarks. A central coordinating body can identify these cross-cutting patterns and prevent duplication of effort. No sectoral regulator has line of sight across domains in this way.

Third, **standard-setting authority**. AISIs can set standards requiring safety cases to specify which parts of a risk pathway their evidence actually covers—and which parts remain unevaluated. This reframes safety cases from "we ran these benchmarks" to "here is where our evidence is strong, here is where it is absent, and here is who should provide it." The risk chain framework provides the structure for such requirements.

This model has precedent. Financial regulators maintain systemic risk frameworks without auditing every firm: they require institutions to report against standardised frameworks and flag when critical risk areas lack evidence (Bank for International Settlements, 2019). Pandemic preparedness similarly required coordination between virologists, epidemiologists, and public health bodies—no single institution assessed the full risk pathway alone (World Health Organization, 2021). AISIs could play an analogous role: maintaining risk chain frameworks, requiring developers to produce evidence for the steps they control, requiring deployers to assess deployment-context steps, and making visible when critical steps have no evidence at all.

As AI proliferates beyond frontier catastrophic risk into healthcare, education, finance, and critical infrastructure, the question AISIs face—*what is our job versus other departments' jobs?*—is precisely what risk chain analysis helps answer. The AISI's job is to maintain the framework, build evaluation capacity in other departments, and ensure accountability across the full chain. Individual departments' job is to provide deployment-context evidence for their domain.

5 Limitations

Risk chains are linear simplifications of pathways that branch, loop, and intersect. Constructing them requires domain expertise, and different experts would draw different chains. The framework identifies gaps but does not prioritise them. Emergent or systemic risks—correlated failures, cascading infrastructure effects—do not follow sequential chains and require different tools (Zwetsloot and Dafoe, 2019; Stein et al., 2024).

6 Conclusion

Current AI evaluation covers a narrow and ecologically questionable portion of the pathway from capability to harm. Risk chain analysis makes this visible and maps which actors should provide evidence at each step. The central implication is that adequate AI risk assessment is a multi-actor coordination problem, and AI safety institutes are best positioned to coordinate it.

References

Anthropic. 2024. Responsible scaling policy, version 1.0.

Apiiro. 2025. AI-generated code risk in fortune 50 enterprises.

Bank for International Settlements. 2019. Stress testing banks: A comparative analysis.

Andrew M Bean and 1 others. 2025. Benchmark inflation: Revealing LLM performance gaps using retro-holdouts. *arXiv preprint arXiv:2501.14892*.

Zana Buçinca, Maja Barbara Malaya, and Krzysztof Z Gajos. 2021. To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. *Proceedings of the ACM on Human-Computer Interaction*, 5(CSCW1).

Sarah R Carter and 1 others. 2023. The convergence of artificial intelligence and the life sciences. *NTI Biosecurity*.

Alan Chan, Matthew J Salganik, and 1 others. 2024. Visibility into AI agents. *arXiv preprint arXiv:2401.13138*.

Fabrizio Dell'Acqua, Edward McFowland III, Ethan R Mollick, and 1 others. 2023. Navigating the jagged technological frontier: Field experimental evidence of the effects of AI on knowledge worker productivity and quality. *Harvard Business School Working Paper*.

Department for Science, Innovation and Technology. 2024. AI safety institute approach to evaluations.

Evan Hubinger, Carson Denison, Jesse Mu, Mike Lambert, Meg Tong, and 1 others. 2024. Sleeper agents: Training deceptive LLMs that persist through safety training. *arXiv preprint arXiv:2401.05566*.

Infocomm Media Development Authority. 2025. Companion guide for agentic AI systems.

Daniel Kilov and 1 others. 2025. Moral competence evaluation of AI systems: Beyond clean scenarios. *arXiv preprint*.

Hao Lin and 1 others. 2025. ARTEMIS: AI agents vs human pentesters on live infrastructure. *arXiv preprint*.

Sean McGregor. 2021. Preventing repeated real world AI failures by cataloging incidents: The AI incident database. *Proceedings of AAAI*.

Christopher Mouton, Caleb Lucas, and Ella Guest. 2024. Can large language models provide useful guidance on social engineering? *RAND Corporation*.

National Institute of Standards and Technology. 2025. NIST AI 700-2: ARIA field testing protocol.

Shakked Noy and Whitney Zhang. 2023. Experimental evidence on the productivity effects of generative artificial intelligence. *Science*, 381(6654):187–192.

OpenAI. 2024. GPT-4o system card: Biological risk assessment.

Raja Parasuraman and Dietrich H Manzey. 2010. Complacency and bias in human use of automation: An attentional integration. *Human Factors*, 52(3):381–410.

Samir Passi and Mihaela Vorvoreanu. 2022. Overreliance on AI: Literature review. *Microsoft Research*.

Aidan Peppin, Edward Sherburn, Alexander Sherburn, and Jide Kim. 2024. Evaluating biological risks from advanced AI. *arXiv preprint arXiv:2412.01946*.

Inioluwa Deborah Raji, Emily M Bender, Amandalynne Paullada, Emily Denton, and Alex Hanna. 2021. AI and the everything in the whole wide world benchmark. *Proceedings of NeurIPS Datasets and Benchmarks Track*.

Inioluwa Deborah Raji, I Elizabeth Kumar, Aaron Horowitz, and Andrew Selbst. 2022. The fallacy of AI functionality. *Proceedings of FAccT*.

Alexander Robey and 1 others. 2025. Jailbreaking and safeguarding large language models: A survey. *arXiv preprint*.

Manuel Rodriguez and 1 others. 2025. Systematization of knowledge: The AI-cybersecurity threat landscape. *arXiv preprint*.

Rylan Schaeffer, Brando Miranda, and Sanmi Koyejo. 2023. Are emergent abilities of large language models a mirage? *Advances in Neural Information Processing Systems*, 36.

Merlin Stein and 1 others. 2024. Interconnected monitoring for AI safety. *arXiv preprint*.

Alex Tamkin and 1 others. 2024. Clio: Privacy-preserving insights into real-world AI use. *arXiv preprint arXiv:2412.13678*.

Veracode. 2025. State of AI code security 2025.

Laura Weidinger, Maribeth Rauh, Nahema Marchal, Arianna Manzini, Lisa Anne Hendricks, Juan Mateos-Garcia, Stevie Bergman, Jackie Kay, Conor Griffin, and 1 others. 2023. Sociotechnical safety evaluation of generative AI systems. *arXiv preprint arXiv:2310.11986*.

World Health Organization. 2021. Strengthening preparedness for health emergencies: Implementation of the International Health Regulations.

John Yang, Akshara Prabhakar, Karthik Narasimhan, and Shunyu Yao. 2024. InterCode: Standardizing and benchmarking interactive coding with execution feedback. In *NeurIPS*.

Andy K Zhang, Neil Perry, Riya Duber, and 1 others. 2024. CYBENCH: A framework for evaluating cybersecurity capabilities and risks of language models. In *arXiv preprint arXiv:2408.08926*.

Wenting Zhao and 1 others. 2024. WildChat: 1M ChatGPT interaction logs in the wild. *arXiv preprint arXiv:2405.01470*.

Lianmin Zheng and 1 others. 2024. LMSYS-Chat-1M: A large-scale real-world LLM conversation dataset. *arXiv preprint arXiv:2309.11998*.

Remco Zwetsloot and Allan Dafoe. 2019. Thinking about risks from AI: Accidents, misuse and structure. *Lawfare Blog*.