Statement of Interest:
Scaling Human-Centered AI Red Teaming

Zifei (FeiFei) Han
Department of Computer Science
National University of Singapore
Singapore, Singapore
hanzifeifei@u.nus.edu

Jane L. E
Department of Computer Science
National University of Singapore
Singapore, Singapore
ejane@nus.edu.sg

As AI systems are increasingly deployed in high-stakes domains, organizations are under pressure to scale AI red teaming to meet growing evaluation demands. In practice, red-teaming efforts rely on heterogeneous human contributors (professionals, contractors, and crowd workers) whose technical literacy and domain expertise vary widely, directly shaping evaluation quality and validity[1] . While automation is often proposed as the primary path to scale, recent work shows that red teaming cannot remain effective without sustained investment in human expertise and infrastructure [2] . This motivates my interest in *reframing scalable AI red teaming as a problem of skill formation, judgment calibration, and human support, rather than efficiency alone*.

My interest in this workshop is shaped by **experience across learning science, HCI research, large-scale industry AI deployment, and product design**. I studied with Prof. Ken Koedinger at Carnegie Mellon University on how AI helps novice tutors develop teaching expertise through learning by doing and situated practice. I also worked as an AI consultant at Deloitte Canada supporting organizations through AI adoption and organizational restructuring, where I observed many failures in AI evaluation stemmed not from missing tools, but from insufficient opportunities for stakeholders to learn how systems behave in practice. I have also built revenue-generating systems in learning and HR focused on evaluating, training and supporting human capability at scale. These experiences directly shaped my perspective that human infrastructure can be intentionally designed and scaled, rather than treated as a fixed bottleneck. I am currently a PhD student working with Prof Jane L. E in Human-Centered AI at the National University of Singapore, where I aim to conduct research with real-world safety and implications.

Gaps observed from practice and research. Current approaches to scaling red teaming often emphasize automation and throughput, which can lead to shallow evaluation and overconfidence in system safety. In practice, human judgment remains both the bottleneck and the primary safeguard in red-teaming efforts, making the quality, preparedness, and well-being of evaluators central to effective AI safety outcomes [1] .

In addition, I have observed persistent gaps between what organizations and users believe AI systems can do and how those systems actually behave in real-world deployment. Red-teaming participants often have uneven AI literacy and domain understanding, which directly shapes evaluation validity and trustworthiness. Moreover, current workflows provide limited support for learning, recovery, and judgment calibration over time, despite evidence that AI systems can significantly influence how human skills develop[3] .

These gaps motivate a reframing of red teaming as a learning environment grounded in learning by doing and situated learning principles. Rather than treat red teaming tasks as isolated tests, I argue that participation itself can be reframed as an opportunity for evaluators to develop AI literacy, refine mental models of system behaviour, and build sustained judgement under uncertainty. From this perspective, responsible red teaming is not only about discovering failures, but about training the people who identify, interpret, and act on those failures. This makes the approach both feasible and scalable: *learning science offers well-established methods for developing expertise at scale, which can be adapted to red-teaming workflows through intentional design for feedback, reflection, and progression.*

What I hope to contribute and gain. I hope to contribute to the workshop by offering this learning-oriented, human-centered framing of scalable AI red teaming. I also hope to contribute by helping articulate how educational principles such as learning by doing and situated learning can directly inform red-teaming design. I am eager to engage with practitioners and researchers to better understand how evaluator expertise is currently trained and supported, where key gaps remain, and how collaborative efforts might shape more sustainable and trustworthy red-teaming practices. I see this workshop as an important opportunity to learn from the community and begin forming collaborations as I shape my PhD research agenda in human-centered AI safety and evaluation.

REFERENCES

[1] Alice Qian Zhang, 2024. The Human Factor in AI Red Teaming: Perspectives from Social and Collaborative Computing. *CSCW Companion* (*Nov 2024*). DOI: https://doi.org/10.48550/arXiv.2407.07786

[2] Alice Qian Zhang. 2025. Effective Automation to Support the Human Infrastructure in AI Red Teaming. *ACM Interactions Publication Tech Labor Forum (Aug '25)*. DOI:https://doi.org/10.48550/arXiv.2503.22116

[3] Judy Hanwen Shen and Alex Tamkin. 2026. How AI Impacts Skill Formation *Anthropic*. DOI:https://doi.org/10.48550/arXiv.2601.20245