EUGENIA KIM

Brooklyn, NY | eugeniavkim [at] gmail [dot] com | eugenia.kim

SELECTED WORK EXPERIENCE

Microsoft — AI Red Team Oct 2020 – Present

AI Safety & Security Researcher II

- Leads research on psychosocial harms in frontier models on the AI Red Team, designing end-to-end studies from threat modeling through mitigation

Software Engineer II (prior role)

- Core contributor to scalable automated red-teaming infrastructure in PyRIT, an open-source Python toolkit, enabling repeatable experimentation across model versions, harm categories, and attack strategies

SELECTED RESEARCH & PUBLICATIONS

[1] XL-SafetyBench: Country-Grounded Cross-Cultural Benchmark for LLM Safety and Cultural Sensitivity Under review. arXiv
Country-grounded benchmark of 5,500 prompts across 10 country–language pairs, evaluating 10 frontier and 27 local LLMs; introduces NSR and CSR metrics, decoupling cultural awareness from jailbreak robustness. In collaboration with Korea AISI.

[2] DisaBench: A Participatory Evaluation Framework for Disability Harms in Language Models Under review. arXiv
Participatory benchmark co-designed with people with lived disability experience of 525 expert-annotated responses showing that general-purpose safety eval systematically misses subtle disability harms only domain experts recognize.

[3] Seeking Late Night Life Lines: Experiences of Conversational AI Use in Mental Health Crisis Accepted to FAccT '26. arXiv
Mixed-methods study of real-world AI use during mental health crises, surfacing patterns of late-night help-seeking and implications for AI safety design.

[4] Expert Evaluation and the Limits of Human Feedback in Mental Health AI Safety Testing Accepted to FAccT '26. arXiv
Shows that expert psychiatrists systematically disagree when evaluating AI responses to mental health scenarios, challenging consensus-based approaches to learning from human feedback. In collaboration with Stanford Center for AI Safety.

[5] From Risk Avoidance to User Empowerment: Reframing Safety in Generative AI for Mental Health Crises 2026. arXiv
Proposes empowerment-oriented design principles for AI crisis support, arguing that avoidant responses harm users who lack alternatives and reduce motivation to seek further help.

[6] A Representation Engineering Perspective on the Effectiveness of Multi-Turn Jailbreaks ICML DIG-BUGS '25. arXiv
Uses representation engineering to show that multi-turn jailbreaks progressively shift model activations into regions perceived as benign, explaining why single-turn defenses fail.

[7] Lessons from Red Teaming 100 Generative AI Products NeurIPS Workshop on Red Teaming GenAI '24. arXiv
Distills eight operational lessons from red-teaming 100+ generative AI systems, covering threat modeling, automation, psychosocial harms, and the role of human judgment.

[8] Taxonomy of Failure Modes in Agentic AI Systems Microsoft AI Red Team Whitepaper, 2025. PDF
Catalogues novel safety and security failure modes unique to autonomous AI agents, including memory poisoning and intent misalignment, with mitigation strategies.

[9] Social and Emotional Uses of AI ACM CHI Workshops '26. Website PDF
Workshop examining how people use AI for social and emotional support, with implications for safety guardrails and wellbeing-aware design.

[10] MLCommons AILuminate Security Jailbreak v0.7 Contributor: attack development & analysis. MLCommons AIRR, 2026. Link
Introduces a mechanism-first taxonomy for single-turn jailbreak attacks and a standardized pipeline for measuring AI resilience under adversarial conditions.

[11] Age Bias in Emotion Detection First author. AIES '21. DOI
First-authored analysis of facial emotion recognition accuracy across age groups, quantifying performance degradation for older adults and proposing bias mitigation strategies.

[12] News Media Framing of Suicide Circumstances and Gender: Mixed Methods Analysis JMIR Mental Health. DOI
Examines gendered stigmatization patterns in news media coverage of suicide, co-authored with CDC researchers.

SELECTED TALKS & PRESENTATIONS

[A] AI Cyber Defense Contest — Invited speaker, National AI CTF, Seoul, South Korea. Nov 2025. Link

[B] UC Berkeley AI Red Teaming Bootcamp — Invited lecturer on automated red teaming for generative models. Aug 2025. Link

[C] Stanford CS521 AI Safety Seminar — Invited speaker: Inside the AI Red Team. Apr 2025. Link

[D] Microsoft Security Communities — Delivered 20+ internal technical talks on red teaming, psychosocial harms, and mitigation.