

Gabriel Mukobi

Web: gabrielmukobi.com | Email: gmukobi@cs.berkeley.edu | Mobile: [360.525.7299](tel:360.525.7299) | GitHub: [mukobi](https://github.com/mukobi) | LinkedIn: [gabrielmukobi](https://www.linkedin.com/in/gabrielmukobi)

Summary:

Researcher, engineer, and leader passionate about research, governance, and field-building to reduce risks from advanced AI systems. Experienced in machine learning research, software engineering, and leadership in both small-team and large-company environments.

Experience:

Technical Staff Member, U.S. AI Safety Institute - July 2024–Present - San Francisco, CA - nist.gov/aisi

Advancing the science, standards, and coordination of AI safety.

Technology and Security Policy Fellow, RAND Corporation - Apr 2024–July 2024 - Washington, DC - rand.org

Informing and improving AI governance decision-making and policy through research.

Technical AI Safety Research Fellow, Existential Risk Alliance - Jul 2023–Sep 2023 - Cambridge, UK - erafellowship.org

Benchmarked cooperative AI capabilities with language models in multi-agent environments; also a [Krueger AI Safety Lab](https://www.kruegerai.com/) Intern.

Gameplay Engineering Intern, Respawn Entertainment - Jun 2022–Sep 2022 - Remote - ea.com

Engineered core gameplay and AI features on Respawn's unreleased Star Wars bounty hunter first-person shooter.

Gameplay Engineering Intern, Riot Games - Jun 2021–Sep 2021 - Remote - riotgames.com

Designed and implemented core features as a Software Engineering Intern on the gameplay team of [2XK0](https://www.riotgames.com/en/games/2xk0) (prev. Project L).

Research Programmer Intern and Tools Programmer Intern, Epic Games - Jun 2020–Jan 2021 - Remote - [unrealengine.com](https://www.unrealengine.com)

Created deep reinforcement learning samples, a machine learning plugin, and virtual production tools in Unreal Engine.

Google Engineering Practicum Intern, Google Cloud Platform - Jun 2019–Sep 2019 - Seattle, WA - github.com/knative-portability

Developed full-stack [open-source applications](https://github.com/knative-portability) as proof of portability for [Knative](https://www.knative.dev/), a platform for serverless containerized workloads.

Selected Projects:

Escalation Risks from Language Models in Military and Diplomatic Decision-Making - Oct 2023–Jan 2024 - [Paper](#), [GitHub](#)

Co-first author. Evaluating the risks from autonomous language model decision-makers in escalating international conflicts. Accepted to ACM FAccT 2024, MASEC NeurIPS 2023 workshop (spotlight).

Welfare Diplomacy: Benchmarking Language Model Cooperation - Jun 2023–Sep 2023 - [Paper](#), [GitHub](#)

First author. Multi-agent LLM evaluations in a novel general-sum variant of Diplomacy that better incentivizes and measures cooperation. Accepted to the SoLaR NeurIPS 2023 workshop, in review at ICML 2024.

SuperHF: Supervised Iterative Learning from Human Feedback - Jan 2023–Sep 2023 - [Paper](#), [GitHub](#)

First author. Alternative to RLHF using supervised learning instead of RL. Accepted to the SoLaR NeurIPS 2023 workshop.

Societal Adaptation to Advanced AI - Jan 2024–May 2024 - [Paper](#), [Thread](#)

Co-first author. How society might adapt to lessen possible future risks from advanced AI systems.

Skills:

Artificial Intelligence - software.gabrielmukobi.com/ai

AI safety, NLP, AI governance, evaluations, ML, DL, foundation models, research mentorship. Languages: Python.

Software Engineering - software.gabrielmukobi.com

Product management, documentation, testing, bug reporting, code review, CS, VCS, [GitHub](#), [GitLab](#). Languages: Python, C++, C#.

Web Development - software.gabrielmukobi.com/web

Full-stack, web design, cloud computing, databases, Docker containerization. Languages: JavaScript, Node.js, Python, HTML.

Game Development - software.gabrielmukobi.com/games

Unreal Engine, Unity, gameplay programming, tools, virtual reality, 3D modeling, computer graphics. Languages: C++, C#, Python.

Education:

University of California, Berkeley - Ph.D. Computer Science - Aug 2024–Future

Ph.D. student researching technical AI governance. Advised by [Jacob Steinhardt](#) and [Dawn Song](#).

Stanford University - M.S. Computer Science - Sep 2023–Mar 2024, B.S. Computer Science - Sep 2018–Dec 2023 - Cum GPA: 4.01

[Stanford AI Alignment](#) Founder and President 2022–24. Coursework in AI/ML, Computer Systems, Graphics, Algorithms, and Theory.