# Gabriel Mukobi

Web: gabrielmukobi.com | Email: gmukobi@cs.stanford.edu | Mobile: 360.525.7299 | GitHub: mukobi | LinkedIn: gabrielmukobi

## Summary:

Researcher, engineer, and student 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 AI Safety Research Fellow, Existential Risk Alliance** - Jul 2023-Sep 2023 - Cambridge, UK - erafellowship.org

Led self-directed technical research benchmarking cooperative AI capabilities with language model agents in multi-agent environments; also a Krueger AI Safety Lab Intern. Skills: Research, AI safety/governance, prompt engineering, machine learning.

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

Engineered core gameplay and AI features as a Software Engineering Intern on the gameplay team of Respawn's unreleased Star Wars first-person shooter title. Skills: Unreal Engine 5, C++.

**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 Project L, Riot Games' unreleased fighting game set in the League of Legends universe. Skills: Unreal Engine 4, C++.

**Research Programmer Intern and Tools Programmer Intern, Epic Games** - Jun 2020-Jan 2021 - Remote - unrealengine.com

Created deep reinforcement learning samples in Unreal Engine and a plugin to facilitate the use of ML in UE4. Engineered tools to predict LED wall moiré and other issues to improve Unreal virtual production shoots. Skills: RL, Unreal Engine 4, C++, Python.

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

Developed several full-stack open-source applications as proof of portability for Knative, an open-source platform for serverless containerized workloads. Skills: Python, Flask, MongoDB, CI/CD, testing, OAuth 2.0, Node.js, Express.js, TypeScript, PostgreSQL.

## 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 (spotlight) to the MASEC NeurIPS 2023 workshop, in review at ACM FAccT 2024.

**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.

**Towards Societal AI Resilience** - Jan 2024-Mar 2024 - Forthcoming

First author; work in progress. Research direction and policy proposal for bolstering society's capacity to withstand, recover from, and adapt to risks from advanced AI systems. Work done through the Astra Fellowship with mentorship by Lennart Heim.

## Skills:

**Artificial Intelligence** - software.gabrielmukobi.com/ai

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

**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 modelling, computer graphics. Languages: C++, C#, Python.

## Education:

**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.

## Interests:

Photography, digital 3D art, filmmaking, music, video and tabletop games, animal welfare, virtual reality, fantasy, and science fiction.