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 advanced AI risks. 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 - July 2023-Sept 2023 - Cambridge, UK - erafellowship.org

Led self-directed and unreleased technical research aimed at reducing risks from advanced language model systems in multi-agent scenarios. Skills: Research, AI safety/alignment/governance, prompt engineering, machine learning.

Gameplay Engineering Intern, Respawn Entertainment - June 2022-Sept 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 - June 2021-Sept 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 - June 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 - June 2019-Sept 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:

Welfare Diplomacy: Benchmarking Autonomous LLM Cooperation - June 2023-Sept 2023 - github.com/mukobi/welfare-diplomacy

Lead author for multi-agent autonomous language model evaluations in a novel general-sum variant of the game Diplomacy designed to better incentivize and measure AI cooperation and evaluated with custom zero-shot scaffolding on various LLMs.

SuperHF - Supervised Program for Alignment Research (SPAR) - Jan 2023-Sept 2023 - github.com/openfeedback/superhf

Lead author for a small research project to develop alternatives to reinforcement learning from human feedback (RLHF) which use supervised learning instead of RL, then evaluating and improving the safety of such methods.

Rogue Starfighter VR - Personal Project - Feb 2020-Mar 2020 - gameplay video - github.com/mukobi/Rogue-Starfighter-VR

A virtual reality Star Wars X-wing flight simulator fan game behind the controls of a fully interactive T-65B X-wing starfighter.

Skills:

Artificial Intelligence - software.gabrielmukobi.com/ai

AI safety, machine learning, deep learning, fine-tuning foundation models, NLP, interpretability, prompt engineering, AI/ML research, reinforcement learning. Languages: Python, PyTorch.

Software Engineering- software.gabrielmukobi.com

Agile development, product management, documentation, unit testing, bug reporting, code review, data structures, algorithms, CI/CD, debugging, IDEs, command line, Linux, Git, Perforce, GitHub, GitLab. Languages: C++, C#, C, Python, Java.

Game Development - software.gabrielmukobi.com/games

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

Web Development - software.gabrielmukobi.com/web

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

Education:

Stanford University - M.S. Computer Science - Sept 2023-June 2024, B.S. Computer Science - Sept 2018-June 2023 - GPA: 3.995

Coursework in AI, Computer Graphics, Computer Systems, Theory, and Algorithms. Stanford AI Alignment Founder and President, past leadership in Stanford Effective Altruism, Stanford XR, Stanford AltPro, and People for Animal Welfare.

Interests:

Photography, digital 3D art, filmmaking, music, video and tabletop gaming, fantasy, and science-fiction.