**Why do you care about AI existential safety?\***

The harms of unchecked AI development are felt disproportionately in the global majority, whom my work aims to serve. My postdoctoral research focuses on fostering digital trust and designing interventions to counter misinformation, limit harmful content, and incentivize truthful information production—critical steps in mitigating societal harms caused by AI systems. I co-founded SimPPL, a nonprofit rebuilding trust on the social internet across six countries through responsible computing tools, including projects addressing misinformation and platform governance. Together, we have taken down botnets using AI to spread digital hate among over 100M users, winning awards from Google, Mozilla, Wikimedia, and others. My Ph.D. research involved causal inference to evaluate interventions against online misinformation, including simulations of influence operations on social networks. These efforts directly address risks stemming from AI misuse, such as disinformation amplification and societal destabilization. By contributing expertise in behavioral economics, causal inference, and platform governance, I aim to ensure AI systems are aligned with humanity’s long-term safety and benefit.

**Please give at least one example of your research interests related to AI existential safety.\***

One of my key research interests related to AI existential safety is ensuring that AI systems are robust, culturally adaptive, and aligned with human values to mitigate risks like misinformation, hallucinations, and inequitable outcomes. We know that the erosion of digital trust has a direct impact on health literacy in vulnerable communities in India and Bangladesh. A recent initiative we spun out from SimPPL, Sakhi, delivers maternal health literacy in India via WhatsApp based conversational AI that is deeply integrated with local frontline health workers, demonstrating how AI can be harnessed for equitable social good while promoting economic prosperity for ASHA workers. It is an AI-powered maternal healthcare platform that integrates local language models (LLMs), behavioral science, and real-time analytics to transform maternal care delivery in India. Unlike general-purpose LLMs prone to hallucinations, Sakhi employs regionally adapted LLMs with expert-curated knowledge bases and user-informed guardrails with real-time integrations for frontline workers to answer user questions, to ensure accurate, culturally relevant information is passed on.The platform also features a behavioral analytics dashboard that uses geo-tagged data to prioritize care for at-risk women during critical periods of pregnancy in order to integrate with state-wide deployment. By reducing monitoring costs by 80% and enabling timely interventions, Sakhi demonstrates how AI can be safely deployed for social good. In fact our work has led to interest from OpenAI and Agency Fund to invite us to deliver a talk to their cohort of GenAI grantees building chatbots for healthcare in India.

In my postdoctoral research I investigate how we can ensure AI systems are transparent, aligned with human values, and resilient against misuse or unintended consequences in the ecommerce marketplaces using behavioral experiments with online audiences recruited via MTurk and Prolific to create a unique human-AI research apparatus that enables reproducibility in an interactive experimental setting. Our paper explores the use of generative AI (GenAI) in digital marketplaces and its implications for AI safety, focusing on mitigating existential risks posed by agentic AI systems. It examines how AI agents, particularly those powered by large language models (LLMs), can exacerbate deceptive practices in e-commerce, such as false advertising and reputation manipulation, which undermine trust and consumer welfare. We then develop an economic intervention using decentralized peer juries to combat the influence of misleading claims produced by agentic sellers in online markets. My presentation of this work at Stanford received invitations to present it to Google's Monetized Policy team and we are actively working on field experiments on Bluesky using their decentralized "feeds" of content.