https://airtable.com/app4P2nC4ke1jkdUW/pagQ2rFSTHxiUUGxZ/form

info@publicbenefitinnovationfund.org

# Tell us about your current projects

### What are you working on? What are you excited about?

SimPPL (https://simppl.org) is a tech nonprofit rebuilding digital trust, working in 7 countries with newsrooms, civil society, and developmental agencies. We work on the value chain of online information, understanding the actors that promote claims, networks that amplify claims, measure its influence on people, and design interventions that counter such influence. Our work has contributed to the takedowns of adversarial threats by Meta, investigations by X/Twitter, awards from Google and Wikimedia, and multiple deployed products with hundreds of users across the world. Key to our work is training young innovators in India, US, and now in other countries through our award-winning responsible computing fellowship programs where we develop an idea right up through its commercialization to teach student teams to deploy AI Products in the field, for public benefit: https://simppl.org/fellowships. This contributed to the launch of an MIT and UNICEF-incubated tech venture that is receiving awards from UNDP, Commonwealth Youth Program, Amazon, and is in talks for a partnership to deploy a state-wide health intervention program with a large tech company in India.

I am excited to continue innovating in the educational space, supporting capacity-building for responsible computing education in CS programs at Tier II and Tier III engineering colleges like mine, back home in India; and now in the US, through initiatives like the NYU AI School. We won Google's prestigious exploreCSR award twice and Mozilla's inaugural Responsible Computing Challenge. With our suite of products for journalists, newsrooms, fact-checkers, and civil society, we are building public interest technology to make reliable information accessible to the global majority, in their language, and on their devices.

# Tell us about any big ideas, gaps, or opportunities you see

### Are there moonshot ideas, tools, or initiatives that should exist? This could be your work – or something you wish someone else would build. If these ideas succeed, how would the world be different?

Integrating AI into physical healthcare networks: Building behavioral analysis engines for integrating conversational health AI systems like Sakhi (https://sakhi-health.com) into state healthcare machinery as a public interest technology. This is so that key metrics from healthcare conversations could be tracked in alignment with state funding priorities in countries like India, Bangladesh where there is a cadre of frontline health workers that are prepared to be tech enablers for the beneficiaries.

Advancing consensus building through democratic dialogues: https://deliberation.io, papers on AI and the future of the digital public square (https://arxiv.org/abs/2412.09988) and the Habermas machine results (https://arstechnica.com/ai/2024/10/googles-deepmind-is-building-an-ai-to-keep-us-from-hating-each-other/) indicate AI can help with consensus building addressing the challenges that decentralized orgs. often face wrt public participation and consensus. It is the future of what will impact democratic speech directly. I wrote an (unfunded) proposal with Deutsche Welle to partner on local deployments of deliberative dialogues technology since trust in public broadcasters was still three times higher in Europe than trust in privately held media orgs. and newsrooms was.

High-fidelity red teaming using LLM-simulated humans to evaluate the harms arising for users, from platform trust and safety choices. Advancing work that facebook was doing internally with web-enabled simulations https://research.facebook.com/publications/wes-agent-based-user-interaction-simulation-on-real-infrastructure except in the interest of public audits of social platforms. Create a "healthy platform benchmark" to aid regulators and policymakers better understand the impact of social media policies. Past efforts include Salesforce AI Economist (didn't work because of low-fidelity and high cost of simulation), current ones include AgentTorch, Oasis, and other multiagent frameworks but lack good validation of real human behavior expressed by LLM agents--and there's a running debate on that. I worked on this in early 2021 at Oxford with Meta researchers: https://misinfocon.com/estimating-harms-from-coordinated-behavior-on-social-networks-897ce7a5447c