Rishav Punatar

AI SAFETY · EVALUATION · PUBLIC COMMUNICATION

rishavpunatar@gmail.com

PROFILE

AI safety practitioner with a humanities background, working at the intersection of frontier model evaluation, child safety policy, and public communication. Leads the UK government's first national safety benchmark for AI systems used by children, and routinely translates technical safety findings for non-specialist audiences: government officials up to Secretary of State level, classroom teachers, and national press.

AI SAFETY EXPERIENCE

Government Incubator for Artificial Intelligence (i.AI)

Mar 2025 - Present

Technical Program Manager, Safety Evaluations · London, UK

No. 10-founded AI start-up transforming UK public services

- Lead the Sovereign Education Benchmark, the UK's first national safety evaluation of AI tutoring systems used by children. Own the programme end to end, from pitching it to the Prime Minister's AI Adviser through to technical delivery.
- Translate technical evaluation findings for audiences with widely differing expertise: senior government officials, safeguarding professionals and classroom teachers, and national press. Much of the work is explaining why abstract model behaviours matter in concrete settings, and what should be done about them.
- Built a novel harm taxonomy for child-AI interactions grounded in KCSIE statutory guidance and the Online Safety Act, translating legal and safeguarding frameworks into evaluation criteria a machine learning team can actually implement.
- Designed and ran adversarial evaluations using Anthropic's Petri and Bloom frameworks and AISI's Inspect, generating thousands of synthetic multi-turn scenarios across child safety risks, and built an LLM-as-judge adjudication pipeline calibrated against 100+ expert human annotators.
- Work across the AI Safety Institute, the Department for Education, frontier labs (Anthropic, Google DeepMind), and civil society organisations (NSPCC, Internet Watch Foundation, 5Rights, the Children's Commissioner's Office), giving a practical map of who is doing what in UK AI safety and where people outside the field can contribute.

RESEARCH, WRITING & FELLOWSHIPS

- "Activation Oracles Improve Deception Detection" (2026). Capstone research paper on interpretability-based methods for detecting model deception. Co-authored with Kai Rawal, Zach Yahn and Ishan Kavathekar; mentored by Lorenzo Pacchiardi.
- Inaugural cohort, International Programme on AI Evaluation (ValgrAI, 2026). The world's first academic programme dedicated to AI evaluation. 40 participants selected globally; faculty drawn from Cambridge, Stanford, Princeton, the UK AI Safety Institute, and Google DeepMind.
- AI Policy Fellow, Imperial College London (2026). Research and writing at the boundary between technical AI systems and the policymakers who have to legislate for them.
- BlueDot Impact Technical AI Safety (2025). Independent red teaming and interpretability work alongside.

EARLIER CAREER

Publicis Sapient

Oct 2020 - Feb 2025

Agile Programme Manager, CoreAI · London, UK

- Delivered a £40m generative AI platform across three continents. Built AI governance frameworks covering model testing, bias detection, and drift monitoring, aligned to the EU AI Act and NIST standards.

Product Manager, Lloyds Bank (2020-2022). B2B API marketplace spanning 500+ APIs.

EDUCATION

University of Warwick · BA Politics, Philosophy and Economics (2:1). Year abroad: University of Bergen, Norway. 2016-2020

University of Oxford, Saïd Business School · Artificial Intelligence Programme (Merit) 2024

INTERESTS

Indian classical vocalist · Filmmaker at the British Museum · Volunteer with Beyond Equality, the Refugee Council, and Shishukunj