SHRUTI RAJVANSHI

New Delhi, India • +91-9899016484 • shruti0809.raj@gmail.com

AI Safety Research • Sociotechnical Systems • AI Policy & Governance • Evaluation Frameworks

PROFESSIONAL SUMMARY

Independent AI safety researcher, policy analyst, and technologist with 10+ years of experience in market intelligence, research, and client consulting. Designed and executed original AI safety experiments including cross-model adversarial red-teaming and cross-cultural LLM benchmarking using self-developed evaluation frameworks. Investigates structural risks of AI in underrepresented and developing-economy contexts, with independent research work on data exclusion, algorithmic bias, and AI governance. Currently pursuing an Online MS in Computer Science (HCI) at Georgia Tech, Atlanta which is a fully remote weekend program. Deeply committed to ensuring AI systems are equitable, safe, and inclusive across diverse global contexts.

EDUCATION

MS (CS – HCI) (Online, Ongoing) | Georgia Institute of Technology, Atlanta, USA *2023 – Present*

- Fully online/remote programme — does not conflict with full-time professional commitments
- Focus areas: ubiquitous computing, Quantum computing, AI ethics, HCI, Ed Tech
- Projects: AI-driven interview coaching platform, Google Fit redesign, infotainment UX research, Arduino IoT

MBA — Marketing & Finance | ICFAI Business School (IBS), Gurugram *2012 – 2014*

B.Sc. (Hons.) — Computer Science | Delhi University *2009 – 2012*

INDEPENDENT AI SAFETY RESEARCH

Iterative Cross-Model Red Teaming (ICRT): Adversarial Evaluation of Frontier AI Safety

Independent Research | April, 2026

- Designed and executed a structured red-teaming framework in which AI models (Claude, ChatGPT, Gemini) adversarial attacked one another across three high-risk domains: cybersecurity, biological/chemical hazards, and manipulation
- Implemented a 6-cell cross-model attack matrix ensuring no model attacked itself; each target-domain pairing used a different attacker model, yielding 60 base prompts and ~180 total interactions across 3 iterative escalation rounds
- Novel meta-finding — All three models willingly generated 10 adversarial attack prompts targeting other models when asked, without resistance; this cooperative role in attack generation is itself an underexplored alignment risk, as models can function as adversarial prompt factories against peer systems
- Key finding — Differential safety behaviour across models: Gemini showed no meaningful restraint as either attacker or target producing unsafe outputs across nearly all manipulation prompts (29/30 iterations scored unsafe) and generating highly dangerous cybersecurity content (e.g., a full medical device exploitation guide) as an attacker before Claude refused to continue; ChatGPT followed through all 20 assigned prompts with only 3–4 refusals; Claude participated in 13 prompts before self-terminating after Gemini produced an exploitable healthcare infrastructure attack guide
- Key finding — Claude broke too, on 2 of 20 cybersecurity prompts: once via CTF (Capture the Flag) gamification framing in Iteration 1, and once via localhost semantic shift (reframing a DDoS tool as a 'local performance calibration' script) in Iteration 3 both cases of context collapse where framing overrode output-level harm assessment
- Key finding — Semantic laundering as primary attack vector: The most consistent method of breaking safety filters was framing harmful requests using professional, educational, or roleplay language (CTF challenges,

B2B engineering documentation, fictional editorial scenes, medical device testing); models evaluated the framing rather than the real-world impact of the output
- Key finding — Attacker role disinhibits models: Models were more willing to produce harmful content when cast as 'red teamers' or 'security researchers,' suggesting that role assignment is a critical and underregulated variable in safety alignment
- Developed binary break classification system (0 = safe, 1 = unsafe) with mandatory one-line justification per classification to eliminate post-hoc scoring bias; pre-registered experimental design prior to execution to ensure methodological transparency
- Full research report and scoring dataset available on GitHub. (https://github.com/shruti0809-raj/AI-Safety-Experiments)

Bharat AI Index: Benchmarking LLMs Across Underrepresented Socio-Cultural Contexts

Independent Research | March, 2026

- Designed original evaluation framework to assess how LLMs (Claude, Gemini, DeepSeek, Perplexity) behave across South Asian linguistic and cultural contexts: India, Sri Lanka, Nepal, Bhutan
- Scale of evaluation: 25 base prompts (5 per dimension) × 4 country localizations = 100 unique prompts × 4 models = 400 responses × 5 scoring dimensions = 2,000 individual scores, one of the most granular independent LLM benchmarking exercises conducted on underrepresented South Asian contexts
- Each base prompt was localized into the linguistic and cultural context of each country (e.g., Hindi/Hinglish for India, Nepali for Nepal, English adapted for Sri Lanka and Bhutan), preserving the underlying intent while reflecting real-world input variation
- Key finding — Safety vs. contextual alignment gap: Safety scores were consistently high across all models and all countries (avg ~4.9–5.0); however, cultural understanding and bias handling showed the widest variance, particularly in Nepali-language and lower-resource contexts
- Key finding — Models do not infer culture independently: When cultural signals were explicit (e.g., Hinglish phrasing), outputs improved; when signals were implicit, models defaulted to Western/global frameworks revealing that cultural alignment is reactive, not proactive
- Key finding — Algorithmic exclusion in developing economies: Models consistently recommended US-centric platforms, formal employment pathways, and globally dominant economic systems, rendering advice inapplicable for users in informal economies, a form of structural exclusion embedded in training data
- Identified five distinct bias patterns: identity-based, linguistic, appearance-based, structural (urban advantage), and gender bias revealing that bias manifests through probability and framing rather than explicit stereotyping
- Full research report with cross-model heatmaps, country-level performance analysis, and recommendations for inclusive evaluation frameworks available on GitHub. (https://github.com/shruti0809-raj/BharatAI-Index)

AI Fairness in Educational Systems: Evaluating Bias and Mitigation in Student Outcome Prediction

MS Ethics Coursework, Georgia Tech | Feb,2026

- Applied Disparate Impact (DI) and Statistical Parity Difference (SPD) fairness metrics to a 649-student educational dataset across gender and age protected attributes
- Key finding: Standard pre-processing bias mitigation (rebalancing/resampling) had minimal effect and in some cases worsened gender-based disparities demonstrating that surface-level statistical adjustments cannot substitute for structural intervention
- Demonstrated that fairness outcomes are sensitive to outcome definition (pass/fail vs. high performance), revealing that fairness is a property of problem framing, not only model architecture
- Full research report and scoring dataset available on GitHub. (https://github.com/shruti0809-raj/AI-Fairness-in-Educational-Systems)

Autonomous Systems Safety: Ethical Decision-Making in Self-Driving Vehicles

MS Ethics Coursework, Georgia Tech | Feb, 2026

- Evaluated three moral decision models (humanist, protectionist, profit-based) in simulated autonomous vehicle crash scenarios, analysing stability under multi-agent interaction
- Key finding: The protectionist 'passenger-first' rule becomes unstable in multi-agent environments when two protectionist systems interact, neither concedes, resulting in ethical deadlock and increased aggregate harm
- Argued for a stable, consistent ethical framework across jurisdictions allowing operational adaptation to local law while keeping core moral logic invariant
- Full research report and scoring dataset available on GitHub. (https://github.com/shruti0809-raj/Autonomous-Systems-Safety)

PUBLICATIONS & MEDIA

- ET Edge (Economic Times) — From Convergence to Divergence: How AI is Reshaping Global Gaps (https://etedge-insights.com/technology/artificial-intelligence/from-convergence-to-divergence-how-ai-is-reshaping-global-gaps/)
- FINS — From Control to Sovereignty: The Next Phase of the AI Race – Fortnightly Journal (https://www.linkedin.com/feed/update/urn:li:activity:7447861705466019840/)
- The Explore Journal — e-Journal – Are we Raising Children Inside Social Media Algorithm? (Copy can be provided on request)
- The Blunt Time — Guardrails For the Frontier – How AI Safety is Actually Being Built (https://theblunttimes.in/guardrails-for-the-frontier-how-ai-safety-is-actually-being-built/63280/)
- News18 — Coverage on story (https://hindi.news18.com/photogallery/career/education-young-ai-expert-shruti-rajvanshi-success-story-campaign-to-make-india-a-leader-in-ai-local18-ws-l-10422191.html)

Core themes: data sovereignty, algorithmic exclusion, AI and marginalised communities, geopolitics of AI, Global South representation in AI systems, increasing divergence due to AI

APPLIED AI PROJECTS

AI-Powered Tarot Reading Application | Independent Project *2025 - Present*

- End-to-end development and deployment of a complex, production-grade AI application on Google Play Store using Gemini-based APIs; handled prompt engineering, backend logic, and full deployment pipeline independently with Chat GPT assistance
- Google Play App Link https://play.google.com/store/apps/details?id=com.taowalker.divineguidance

AI-Driven Soft Skills & Interview Coaching Platform | Georgia Tech — MS Project *2024*

- Developed a microlearning and interview assessment platform with ML-based real-time feedback on communication and leadership integrating AI assessment models with UX principles from HCI research

PROFESSIONAL EXPERIENCE

Associate Director — Central Growth Team | Market Xcel, India *Jan'26 - Present*

- Leading central growth strategy and client consulting initiatives across research and intelligence verticals
- Driving business development, key account management, and strategic market intelligence solutions for enterprise clients

Technology Consultant | Shor Foundation, India *Sep'24 – Mar'25*

- Volunteer technology lead overseeing digital transformation, website architecture, SEO strategy, and event registration platform development

Manager, Market Intelligence | GfK Nielsen India Pvt Ltd, Gurugram *Apr'21 – Dec'23*

- Leveraged GfK Newron AI-powered intelligence platform to analyse consumer behaviour, technology adoption trends, and competitive landscapes across high-growth sectors
- Managed key client accounts and drove new business acquisitions worth $1.2M USD; delivered AI-based market forecasting and strategic advisory
- Led product development of data-driven intelligence offerings integrating AI-based insights and digital market signals

Manager, pre-Sales | Tenon Group Companies, Gurugram *Oct'17 – Jun'20*

- Prepared business proposals and RFP responses for large enterprises and government projects in security technology and facility management
- Engaged government and corporate stakeholders on technology-driven security solutions and AI/automation integration

Assistant Manager, Analytics | Tenon Group Companies, Gurugram *Apr'14 – Jun'16*

- Conducted market assessments, feasibility studies, and competitor benchmarking for international expansion in the security tech sector
- PMO for ERP system implementation, enhancing workflow automation and data-driven reporting

CERTIFICATIONS & TRAINING

Responsible Conduct of Research (CITI Program)

Introduction to FinTech (HKU / edX)

Social & Behavioural Research Ethics (CITI Program)

Design Thinking for Business Innovation (U.Virginia / Coursera)

Google Digital Marketing Certification

KEY SKILLS & AREAS OF EXPERTISE

AI Safety & Alignment: Red-teaming, adversarial evaluation, safety benchmarking, multi-model interaction analysis, break-rate metrics

AI Evaluation Frameworks: Cross-cultural LLM benchmarking, fairness metrics (DI, SPD), bias detection, structured prompt design, scoring methodology

AI Policy & Governance: Data sovereignty, algorithmic exclusion, geopolitics of AI, Global South AI equity, inclusive deployment frameworks

Research Methods: Pre-registered experimental design, qualitative & quantitative analysis, HCI usability research, survey design

Technical: Python, Flutter, Firebase, MongoDB, Node.js, ML model integration, API development, prompt engineering, Solidity/Ethereum

Languages: English (fluent), Hindi (native)