VARAD VISHWARUPE

Doctoral Scholar · Human-Centered Ethics of AI · Alignment & AI Safety

Oxford, UK · varad.vishwarupe@cs.ox.ac.uk · +44 7424 464146

Webpage: cs.ox.ac.uk/people/varad.vishwarupe · h-index = 12 · Citations = 700+

PhD researcher building deployment-time alignment frameworks and reproducibility standards for frontier AI safety.
Prior: Student researcher intern at Google DeepMind; Research scientist at Amazon Alexa AI; SDE at Microsoft Azure.

RESEARCH INTERESTS

Human-Centered AI · AI Safety · Alignment · Human–AI Interaction · Alignment Evaluation · Technical AI Governance ·

EDUCATION

DPhil (PhD), Computer Science – University of Oxford, UK **Oct 2024 – Sep 2027**
Inaugural Computer Science Scholar, Oxford Institute for Ethics in AI ·
Supervisors: Sir Nigel Shadbolt, Prof. Marina Jirotka, Prof. Ivan Flechais ·

M.Tech, Artificial Intelligence & Machine Learning – BITS Pilani, India **Sep 2022 – Aug 2024**
GPA 9.81/10 · Ranked 3rd of 122 · Bronze Medal · Best Dissertation Prize

M.Sc., Advanced Computer Science – University of Oxford, UK **Oct 2019 – Sep 2020**
Distinction (78.6%) · Department Merit Scholarship

B.Eng., Information Technology – MIT College of Engineering, Pune, India **Aug 2012 – Aug 2016**
Distinction (79.67%) · Departmental Rank 1st of 178 · Gold Medal

RESEARCH & INDUSTRY EXPERIENCE

Student Researcher Intern – Google DeepMind **Jul 2025 – Sep 2025**
Semantic stress-testing of RLHF-trained agentic models London, UK

- Developed semantic stress-testing methods for RLHF-trained models under paraphrase and inquiry-phrasing variation; characterised how small lexical shifts produce divergent reasoning paths in agentic settings.
- Contributed to alignment evaluation work on stability and value-preserving behaviour, examining how optimisation objectives shape implicit normative responses in deployed configurations.

Graduate Teaching Assistant – University of Oxford **Oct 2024 – Present**
Department of Computer Science Oxford, UK

- Teaching Assistant for Machine Learning, Probabilistic Inference (Neural Networks), Computer Architecture, Computer Vision, and Quantum & Probabilistic Computing.
- Mentored 8 undergraduate and MSc students across Oxford colleges; interviewed 18 candidates for undergraduate CS admissions at Keble College; marked the Mathematics Admissions Test (MAT).

Research Scientist II, Alexa AI – Amazon **Oct 2020 – Aug 2022**
Automatic Speech Recognition & Voice Services Hyderabad, India

- Led core ASR and wake-word detection research within Alexa Voice Services and Alexa Skills Kit; production work delivered a 12.7% increase in user satisfaction and 8% reduction in ASR error rates serving millions of users.
- Managed a six-person team of engineers and data scientists; shipped few-shot learning pipelines on Amazon SageMaker and AWS Lambda; awarded Impactful Team Player recognition.

Software Development Engineer I, Azure ML – Microsoft **Oct 2017 – Sep 2019**
ML-as-a-service infrastructure and Azure Stack HCI Hyderabad, India

- Built ML-as-a-service features within Azure Machine Learning Studio enabling scalable model deployment and integration of core and advanced ML models for enterprise users.
- Contributed to hybridisation of on-premises HCI infrastructure with Azure cloud services; deployment expanded Azure Stack HCI adoption by approximately 6%.

SELECTED PUBLICATIONS

Under Review

- **Vishwarupe, V.**, Flechais, I., Jirotka, M., Shadbolt, N. (2026). *NeurIPS Should Require Reproducibility Standards for Frontier AI Safety Claims*. **NeurIPS 2026 Position Paper Track.**
- **Vishwarupe, V.**, et al. (2026). *Position: Deployment-Relevant Alignment Cannot Be Inferred from Model Audits – Audit of 16 LLM Alignment Benchmarks (Cohen's κ = 0.87).* **NeurIPS 2026 Position Paper Track.**
- **Vishwarupe, V.**, et al. (2026). *Minimal Sufficient Benchmarks: Score Imputation for Cost-Efficient LLM Leaderboards.* **NeurIPS 2026.**
- **Vishwarupe, V.**, et al. (2026). *The Dual-Use Agency Gap: Missing Actors in Technical AI Governance Access Architectures.* **ICML 2026, TAIG Workshop**

- **Vishwarupe, V.**, et al. (2026). *Beyond the Artifact Layer: A Framing Note on the Locus of Governance for AI-Assisted Science.* **ICML 2026, TAIG Workshop.**
- **Vishwarupe, V.**, et al. (2026). *EU AI Act Article 12 and the Interaction-Trace Gap: Logging Human–AI Workflows for High-Risk AI Compliance,* **ICML 2026.**

Accepted

- **Vishwarupe, V.**, et al. (2026). *From Dual-Use Awareness to Dual-Use Agency: Epistemic Distance and the Structural Limits of Ethical Responsibility in AI Research.* **ICLR 2026.**
- **Vishwarupe, V.**, et al. (2026). *Pluralism as a Safety Property: Auditing RLHF-Tuned LLMs in Labour Contexts.* **FAccT 2026.**
- **Vishwarupe, V.**, Flechais, I., Jirotka, M., Shadbolt, N. (2026). *"To LLM, or Not to LLM": How Designers and Developers Navigate LLMs as Tools or Teammates.* **CHI 2026,** Barcelona. ACM. doi:10.1145/3772363.3798953
- **Vishwarupe, V.**, Jirotka, M., Shadbolt, N., Flechais, I. (2026). *The Collaboration Gap in Human–AI Work: Grounding and Repair Conditions for Stable Collaboration.* **ECSCW 2026**, Germany.
- **Vishwarupe, V.**, Flechais, I., Jirotka, M., Shadbolt, N. (2026). *From Rights to Rites: Expectations Management in Smart Home AI.* **HCI International (HCII)** 2026, Montreal. Springer.

Previously

- **Vishwarupe, V.**, Kuklani, R., et al. (2025). *Semantic Jailbreaks and RLHF Limitations in LLMs: A Taxonomy, Failure Trace, and Mitigation Strategy.* **International Journal of Computer Applications**, 187, 187–196.
- **Vishwarupe, V.**, et al. (2025). *Breaking the Black Box: Securing and Auditing Edge-Deployed LLMs via Shard Traceability.* **International Journal of Computer Applications**, 187(27), 44–49.
- **Vishwarupe, V.**, Joshi, P., et al. (2022). *Explainable AI and Interpretable Machine Learning: A Case Study in Perspective.* **ACM ISCSI 2022,** Portugal. **[Best Paper Award]**
- **Vishwarupe, V.**, Bedekar, M., Pande, M., et al. (2022). *Bringing Humans at the Epicentre of Artificial Intelligence: A Confluence of AI, HCI, and Human-Centred Computing.* Procedia Computer Science, **Elsevier.**

SELECTED RESEARCH PROJECTS

Repair-Oriented Alignment Evaluation: Dual-coded audit (Cohen's κ = 0.87) of 16 leading LLM alignment benchmarks against an eight-dimension interactional rubric; verification support absent across the ecosystem, process steerability nearly absent, interactional tier fragmented.

Reproducibility Standards for Frontier AI Safety Claims: Three-tier disclosure framework (T1 public / T2 controlled / T3 restricted) with claim inventory, mandatory scope statements, and graduated sanctions, modelled on ACM artefact-review badging.

Virtuous Intelligence Framework (VIF): Deployment-time alignment framework operationalising three mechanisms – scoping, signalling, and repair – derived from grounded-theory analysis of human–agent interaction breakdowns; prototype Live VIF Monitor wired to a frontier LLM API for controlled validation.

Minimal Sufficient Benchmarks: Treated leaderboard maintenance as a score imputation problem; exhaustively searched all 62 benchmark subsets across two leaderboards (HF Open LLM v2 and LiveBench); identified Instruction-Following as the only orthogonal axis to general capability and a two-stage hybrid pipeline saving ~49% of evaluation cost at 83% true-top-100 recall.

SKILLS

AI Safety & Alignment. Alignment evaluation design, RLHF, Constitutional AI principles, red-teaming, jailbreak analysis, semantic stress-testing, repair-oriented benchmarks, scalable oversight; ARENA programme and BlueDot Impact courses.

Programming. Python (fluent, production-grade), C++, R, SQL, JavaScript (React, Node.js), HTML/CSS.

ML Frameworks & Infrastructure. PyTorch, TensorFlow, JAX, Haiku, Scikit-learn, Hugging Face, Pandas, OpenCV, AWS (EC2, Lambda, SageMaker), Azure ML.

Empirical & HCI Methods. Constructivist grounded theory, semi-structured interview design, dual-coded thematic analysis, mixed-methods study design, RCT design, NVivo.

HONOURS & SERVICE

- Inaugural Computer Science Scholar, Oxford Institute for Ethics in AI – first CS researcher in the Institute's history.
- Funded PhD admits from Oxford, Cambridge, UCL, and Harvard (acceptance rates < 10-15%).
- Top 20 Young Scientists (out of 835 applicants) – Global Partnership on AI (GPAI) Summit 2023, New Delhi; presented research on AI to the Honourable Prime Minister of India.
- Young Researcher, ACM Heidelberg Laureate Forum 2023 (200 selected globally).
- Best Paper Awards: ACM ISCSI 2022 (Portugal); ACM CIIS 2022 (China); Springer ICT4SD 2018 (London).
- Conference Reviewer: NeurIPS, ICML, ACM SIGCHI, ECCV, SIGIR, ACM FAccT, ICLR (2022–2026).
- Peer Reviewer: WIREs Data Mining and Knowledge Discovery (Wiley); International Journal of Intelligent Systems (UK).
- India's Top 30 under 30: Excellence in Technology Awardee 2023.
- Granted Patents: Content-relevance and priority-filtered text-message notifications (IN201621024128), AI-driven intelligent payment alerts (ZA202208837B), Intelligent hybrid Twitter spam classifier (ZA202108302B).