ABDULLAH X

AI Safety and Governance Researcher | Founder, Project AWARE | Founder, Zephara AI
New Delhi, India | +91 82879 59421 | abdullah@projectawareai.org
linkedin.com/in/abdullah-x-22a2211b7 | github.com/abdullah-x-bd | projectawareai.org

PROFESSIONAL PROFILE

AI safety researcher, institution builder and technology policy practitioner working at the intersection of technical evaluation, frontier AI governance and international cooperation. Founder of Project AWARE, a New Delhi-based think tank with an approximately 55-person researcher and fellow network, and Zephara AI, an applied legal technology venture. Research and institutional work focus on AI assurance, executable safeguards, multi-agent systems, deployment governance, machine unlearning and public-sector oversight.

LEADERSHIP AND INSTITUTIONAL APPOINTMENTS

Project AWARE | Founder and President | New Delhi, India | Nov 2024 - Present

- Founded and lead a New Delhi-based think tank working on institutional AI safety through research, diplomacy and capacity building; built a distributed network of approximately 25 researchers and 30 fellows.
- Established research and delivery programmes in technical AI safety, country governance, institutional and legal research, strategic advisory, international engagement and university capacity building.
- Direct the AI Sovereignty in an Interdependent World programme, mapping dependence on foreign frontier models, compute, cloud and deployment routes, and identifying practical governance leverage for countries without domestic frontier developers.
- Lead comparative country governance research spanning India, Malaysia, Colombia, Australia, South Africa and France, alongside smaller-state research, using a shared strategic-node and deployment-pathway framework.
- Lead institutional and diplomatic cooperation on AI assurance, procurement, deployment safeguards and sovereignty; the August 2026 flagship convening briefed 13 states and generated seven bilateral follow-up pathways.
- Developed and submitted to the Ministry of External Affairs (NEST) a concept note and technical annex for a jointly governed multinational frontier AI capability, covering model and compute access, financing, export controls, security and oversight.
- Built a selective international engagement fellowship and distributed research network across major policy centres in Europe, the Americas and Asia-Pacific; develop research partnerships, training programmes and institutional advisory capacity.

Arcadia Impact | Research Associate, AI Governance Taskforce | Nov 2026 - Present

- Research associate on Risk Escalation Indicators and Thresholds for Multi-Agent AI, working with Mike Harre's project team.

Zephara AI | Founder and Chief Executive Officer | Noida, India | Aug 2023 - Present

- Founded an AI legal technology venture for Indian criminal-justice workflows; built an NLP/LLM-based MVP supporting legal drafting, case management and language access.
- Led a distributed team of more than 50 researchers, engineers and business staff across development, operations, product strategy and partnerships; established a 300-lawyer pilot network for product validation.
- Selected for the DoT/TCOE/COAI ASPIRE programme at India Mobile Congress in both 2024 and 2025.

Department of Telecommunications, Government of India | Member, National Working Group on Artificial Intelligence | New Delhi, India | Jul 2026 - Present

- Contribute technical and policy perspectives to national work on AI standards and international standardisation, including responsible deployment, safety assurance, ITU-T processes and increasingly autonomous telecom systems.

CoRE-AI, The Dialogue | Member | New Delhi, India | Sep 2024 - Present

- Participate in a cross-sector AI safety and governance coalition, supporting discussion of institutional safeguards, public-interest deployment and technology policy.

MAP-AI Global Working Group on Safe & Trusted AI | Selected Organisational Representative, Project AWARE | International | Oct 2026 - Present

- Project AWARE was selected for the international working group in October 2026; represented as organisational selection rather than completed joint outputs.

TEACHING APPOINTMENTS

IILM University | Guest Faculty, Digital Transformation II | Greater Noida, India | Aug 2025 - Present

- Teach Digital Transformation II to second-year MBA students using real-world technology cases and structured practitioner problems; design assessments, marking rubrics, mini-projects, capstone guidance and weekly learning support.

Kautilya School of Public Policy | Incoming Guest Faculty | Hyderabad, India | From Jan 2027

- Developed the 30-contact-hour Master of Public Policy course AI Sovereignty in an Interdependent World, covering deployment routes, public law, procurement, institutional AI safety, comparative governance and international cooperation.
- Designed applied exercises in deployment mapping and institutional readiness, with a final state-strategy simulation.

SELECTED TECHNICAL RESEARCH AND PROGRAMMES

POLIS: Algorithmic Institutions for Multi-Agent AI Safety | Research Lead | Project AWARE | 2026 - Present

- Designed and ran a frozen 5,280-episode study across seven live model endpoints, comparing rule communication, trusted authority state, executable enforcement and recovery in multi-agent settings.
- In 96 matched transformation trials, mutable local guards permitted 22 policy-laundering violations versus zero under immutable provenance enforcement (exact paired p = 4.77e-7); identified divergent causal mechanisms beneath identical final safety rates.

SafeguardShift: Safety Under Institutional Safeguard Loss | Research Lead | Project AWARE | 2026

- Built a benchmark of 72 decision tasks spanning six operational domains and six safeguard classes, collecting 3,240 agent trajectories; measured continued action, hidden prerequisites, workaround attempts and recovery when safeguards disappear.
- Separated attempted unsafe behaviour from executed harm, demonstrating how environmental blocking can obscure unsafe agent intent.

State-Exact Trace-Preserving Deletion | Independent and Project AWARE Research | 2025 - 2026

- Developed a deterministic training-replay approach to machine unlearning that physically removes deleted data while preserving execution traces; demonstrated exact model and optimizer replay for Pythia 160M, Pythia 2.8B and Llama 3.2 1B.
- Implemented audit-gated pathways for exact replay, short-horizon reversion, adapter deletion and approximate deletion methods.

Status Association and Decision Leakage | BlueDot-Supported Research | 2026

- Conducted a frozen eight-model, 8,256-response audit using Chilean surname probes to distinguish latent socioeconomic status association from consequential decisions in hiring, education, fellowships and legal aid.
- Found that detectable status association did not reliably predict decision leakage, motivating direct association-to-action evaluation.

Markovian Circuit Tracing | Independent Research | 2026

- Constructed a transformer interpretability benchmark with known Hidden Markov Model states, Bayesian belief targets and forced-state counterfactuals; tested recovered latent representations through activation patching and negative controls.

PipelineProof | Independent Research | 2026

- Built an execution-based benchmark and reinforcement-learning environment for silent methodological failures in machine-learning pipelines, spanning six defect families with adversarial attacks, legitimate repairs and local/container parity checks.

PDRS and FinSpace | Independent Technical Research | 2026

- Implemented Python, C11 and Rust runtimes for finite dependent schemas and exact integer-addressable domains; built FinSpace on this foundation for deterministic financial scenario enumeration, reconstruction and allocation.

Policy Research and Work in Development

- AIPEX (2026): Developed a policy proposal for structured access to frontier AI systems in designated high-risk professions, subject to institutional assurance and oversight; not an adopted programme.
- Constitutional AI and Conflicting Principles (2026): Developed a formal research design for model response selection under competing principles and planned empirical tests; results not yet established.
- Autonomous Telecom Network Assurance (2026): Developed an initial policy and expert-consultation concept on evaluation and accountability for increasingly autonomous networks; convening not yet held.
- Additional technical artifacts include Probe Scent and the BlueDot Technical AI Safety Puzzle; earlier CrisisBench work was incorporated into SafeguardShift.

PUBLICATIONS, PREPRINTS AND RESEARCH MANUSCRIPTS

Published Papers, Public Preprints and Policy Research

- **Multi-Agent AI Safety as an Institutional Design Problem.** Abdullah X. arXiv:2608.09828 (2026).
- **Status Association Does Not Reliably Predict Decision Leakage.** Abdullah X. arXiv:2608.10089 (2026).
- **AI Agents Keep Acting After Safeguards Disappear.** Abdullah X. SSRN:7271402 (2026).
- **Markovian Circuit Tracing: Ground-Truth Tests of Transformer State Abstractions.** Abdullah X. arXiv:2605.20824 (2026); revised manuscript under peer review at Neural Processing Letters. Original preprint title: Markovian Circuit Tracing for Transformer State Dynamic.
- **Unlearning at Scale: State-Exact Trace-Preserving Deletion in Billion-Parameter Language Models.** Abdullah X. arXiv:2508.12220 (2025; revised 2026); revised manuscript under review at TMLR.
- **Governing Foreign Generative AI Through Deployment Routes in the Global South.** Abdullah X. SSRN:6508277 (2026).
- **Beyond Universal Metrics: A Global South-Centered Framework for Ethical AI Development.** Abdullah X and Fiza Siddiqui. SSRN:5792282 (2025; revised 2026); under peer review at Humanities and Social Sciences Communications.
- **Stabilising and Reconciling India's Composite School Grant.** Abdullah X. Empirical policy brief based on eight years of UDISE+ school-level data (2026).

Accepted and Under-Review Manuscripts

- **AI Copyright Must Show the Human Hand.** Abdullah X. Accepted, Delhi Law Review (2026).
- **When Exception Becomes Architecture in Indian AI Governance.** Abdullah X. Accepted, Journal of Global Legal Review (2026).
- **Improved prediction of individual respondents does not reliably imply improved synthetic population fidelity.** Abdullah X. Under review, Discover Artificial Intelligence (2026); survey-weighted study of 1,000 Indian respondents.
- **Finite Evidence Cannot Certify Exact Chance.** Abdullah X and Fiza Siddiqui. Under review, Journal for General Philosophy of Science (2026).
- **Can Public Credit Pay for Itself? Sovereign Offset Financing and the Fiscal Architecture of Atmanirbhar Bharat.** Abdullah X. Policy-economics manuscript prepared for the XXI International Conference on Public Policy and Management (2026).

Public Commentary

- **Regulate or Enable? Mixed Signals on AI Infrastructure Cause for Concern.** Abdullah X. ET Edge Insights, 21 Oct 2025.

SELECTED SPEAKING, MAJOR CONVENINGS AND MEDIA

- **NDTV Profit, The Big Question | Invited Television Panellist | 2026.** Discussed the White House AI Accord, independent oversight of frontier AI developers and AI assurance capacity in India.
- **National Law University Delhi** | Panellist | 30 Oct 2026. Indo-German Conference on AI in Law: Navigating Regulation, Ethics, and Accountability; Theme 2 on digital courts, judicial automation and innovation in justice delivery.
- **O.P. Jindal Global University | Confirmed Invited Speaker | 27 Nov 2026 (upcoming).** Session: Comparative Approaches to AI Governance, International Convention on Constitutionalism in the Age of Artificial Intelligence, Sonipat, India.
- **Universidad Carlos III de Madrid | Invited Panellist | 2 Oct 2026.** India Week session Ethics as the Red Button of AI, online; addressed ethical governance and institutional consequences of advanced AI.
- **NASSCOM and RBB Economics | Invited Speaker | Apr 2026.** Spoke on implementation of the AI Chakras and competitive access to AI resources.
- **AI Sovereignty in an Interdependent World | Convener | 21 Aug 2026.** Conceived and co-convened Project AWARE's closed-door diplomatic roundtable with Observer Research Foundation and The Dialogue, New Delhi.
- **India AI Impact Summit | Convener and Moderator | 16 Feb 2026.** Led Project AWARE's main-summit roundtable From Guidelines to Ground: Institutional AI Safety for India and the Global South, Bharat Mandapam, New Delhi.
- **Jindal Global Business School | Invited Guest Lecturer | Feb 2024.** Delivered Balancing Innovation with Responsibility in Leadership.

SELECTED GOVERNMENT AND DIPLOMATIC ENGAGEMENTS

- **Ministry of External Affairs, Government of India | 2026.** Developed and presented a proposal for a six-country shared frontier AI capability, supported by a technical and institutional annex and financial model. Held discussions with the NEST division on the proposed governance structure and intergovernmental cooperation.
- **Netherlands and Prime Minister's Office of Liechtenstein | 2026.** Advanced country-specific strategic advisory work on how states can exercise leverage over foreign frontier AI developers. Discussions with the Dutch Deputy Head of Mission explored the Netherlands' position in the AI supply chain, while Liechtenstein work examined practical governance options for smaller states.
- **United States and Brazil | 2026.** Developed two Track 1.5 AI assurance dialogue initiatives through diplomatic engagement, examining independent frontier-model evaluations and reciprocal assurance arrangements. Secured preliminary diplomatic interest and advanced preparations for bilateral convenings.
- **France | 2025–2026.** Selected for a closed-door AI policy consultation with the French Ambassador and the French President's Special Envoy for AI. Subsequently developed discussions with French counterparts on a governance framework for sustainable AI data centres.
- **El Salvador and SICA | 2026.** Held consultations with El Salvador's diplomatic mission on regional AI cooperation, including a proposed SICA workshop on sovereign-language AI and opportunities for smaller states to participate in international AI assurance arrangements.

FELLOWSHIPS AND SELECTIVE PROGRAMMES

Policy Agora, Esya Centre and Observer Research Foundation | Applied Policy Fellow | New Delhi, India | Jul - Aug 2026

- Completed an eight-week applied policy fellowship focused on regulatory research, geopolitics, practitioner-led policy writing and stakeholder engagement.

Young Leaders for Active Citizenship (YLAC) | Policy-in-Action Fellow | New Delhi, India | Oct 2025 - Feb 2026

- Researched Scheduled Caste representation and educational inequality in Tamil Nadu for Lok Sabha member Sasikanth Senthil; developed district and subgroup comparisons and an implementation-focused policy memo.

BlueDot Impact | Technical AI Safety Research Project and Course | Remote | May - Jun 2026

- Completed a 30-hour technical AI safety course and an original research project on status associations in model decisions; BlueDot supported subsequent expansion to the eight-model confirmatory study.

UK-India AI Collaborative Innovation Programme | Selected Participant | 2026

- One of 14 selected Indian participants; developed SourceChain AI, a dependency-discovery and assurance concept selected for demonstration at the India AI Impact Summit.

UK-India AIxcelerate | Selected Participant | 2025

- Contributed to Kredible, a local-language borrower journey comparing real-time bank offers; team placed among the top groups.

The Dialogue by GDi Partners | Selected Policy Debate Judge | 2025

- Served as a remunerated adjudicator for a national policy debate programme.

EARLIER PROFESSIONAL AND NONPROFIT EXPERIENCE

Complidata | Consultant, then AI Engineer | Tervuren, Belgium | Jul 2022 - Oct 2023

- Developed Java/Python automation, cloud-based AI and document intelligence workflows on Google Cloud; worked on infrastructure processing more than 200,000 documents.
- Built anti-money-laundering and sanctions-screening systems involving international lists; client systems included work for Citi and Credit Suisse.

180 Degrees Consulting | Strategy Consultant | Leuven, Belgium | Feb - May 2023

- Designed a mathematical model and franchise playbook for off-grid hydrogen cooking and cooling in Uganda; contributed to business-model research for the Impact Licensing Initiative.

Academics for Development, Brightminds | Brightminds Fellow | Leuven, Belgium | Feb - May 2023

- Participated in interdisciplinary student social-impact problem solving and project work.

Undivided vzw | Volunteer, Board Member and Acting Chief Financial Officer | Leuven, Belgium | Jan 2021 - May 2022

- Helped establish an independent Belgian nonprofit; served on its board with responsibility for budgeting, donor reporting and financial controls.

DecolonizeKUL | Co-Founder | Leuven, Belgium | May - Nov 2022

- Co-founded a student initiative at KU Leuven and supported early strategy, outreach and operations.

IMSAL | Volunteer | Leuven, Belgium | Feb 2021 - Jan 2024

- Supported student and community events and associated coordination activities.

McDonald's | Crew Member | Leuven, Belgium | Sep - Dec 2021

- Worked part time in high-volume customer service and team operations while studying.

EDUCATION

Aligarh Muslim University | Bachelor of Arts in Economics, Minor in Political Science | Aligarh, India | Expected Dec 2026

- Concentration in international, public and development economics, advanced economic theory and applied analysis.

Takshashila Institution | Graduate Certificate in Public Policy, Technology and Policy | Bengaluru, India | Dec 2025

- Training in economic reasoning, policy formulation, AI policy, information economics and technology geopolitics.

KU Leuven, Groep-T | Electronics and ICT Engineering Coursework (Degree Not Completed) | Leuven, Belgium | 2020 - 2023

- Completed undergraduate technical coursework in computing, software engineering, networks, data science, engineering economics and enterprise management; programme left in Dec 2023.

Bhartiya Vidya Bhavan's Mehta Vidyalaya | CBSE Senior Secondary, Physics, Chemistry and Mathematics | New Delhi, India | Jul 2020

- CGPA: 9.2/10.

SELECTED RECOGNITION AND CREDENTIALS

- siliconindia: Top 10 Legal Tech Startups of 2023 (Zephara AI).
- Global Business Awards 2024: recognition for AI-driven legal solutions (Zephara AI).
- CodeSignal Industry Assessment: 600/600 (May 2026).
- TOEFL iBT: 110/120 (Feb 2020). SAT Subject Test, Mathematics: 740/800 (Dec 2019).

TECHNICAL, RESEARCH AND POLICY SKILLS

AI safety and research: frontier-model evaluation; agentic and multi-agent safety; executable institutional safeguards; mechanistic interpretability; machine unlearning; bias and fairness audits; synthetic-population validation; causal interventions; benchmark design; reproducible experimentation; statistical and equivalence testing.

Software and ML: Python; Java; Kotlin; C++; C11; Rust; SQL/PostgreSQL; Linux; Bash; Git; PyTorch; TensorFlow; JAX; Hugging Face; embeddings and RAG; instruction tuning; PEFT/LoRA/QLoRA; GCP; Vertex AI; BigQuery; Document AI; Docker; Kubernetes; CI/CD; MLflow; Weights & Biases; FAISS; Milvus; Pinecone; Airflow; dbt; REST/gRPC.

Governance and leadership: frontier AI governance; deployment oversight; independent assurance; procurement and contracting; AI standards; public policy; regulatory analysis; GDPR; India's DPDP framework; NIST AI RMF; ISO/IEC 42001; ITU-T standardisation; institutional development; policy research; research supervision; programme management; curriculum design; international stakeholder engagement.

LANGUAGES

English, Hindi and Urdu: native or bilingual proficiency. French, Arabic and Dutch: limited working proficiency.