Atmadeep Ghoshal

Civic and Responsible AI Lab
Department of Informatics, King's College London, London, UK
Email: atmadeep.ghoshal@kcl.ac.uk

Education

King's College London, UK *2024 - present*
PhD student in Computer Science
Advisors: Dr. Martim Brandao, Dr. Ruba Abu Salma & Dr. Sanjay Modgil

Jadavpur University, India *2020 - 2022*
MA in Sociology
GPA: 9.94/10
Advisor: Prof. Amites Mukhopadhyay

Jadavpur University, India *2017 - 2020*
BA in Sociology
GPA: 9.83/10

Technical Skills

Programming Languages: Python
Qualitative Research Methods: Ethnography, Participatory Design, Interviews, Survey, Horizon Scanning
Research Interests: Human Centred AI, AI Alignment, Participatory AI Design, Benchmarking and Evaluation Standards for LLMs in Global South, Constitutional AI approaches towards AI Safety, Red Teaming automation

Work Experience

King's College London *May 2025 - Present*
Graduate Research Assistant
Advisors: Dr. Martim Brandao, Dr. Caitlin Bentley
Responsibilities
- Worked on a **UK AISI** funded project about building a Human-AI collaboration framework for autonomous underwater robots
- Interviewed professionals from UK Army, UK Navy and other private bodies for understanding remote operators' general perspectives on maritime AI autonomy.
- Conducted quantitative user studies for gaining user feedback regarding specific AI enabled systems in maritime sector such as Guardian by Marine AI.
- Designed and evaluated results for an experiment related to degree of success in human-AI complementarity in AI assisted maritime operations.

King's College London *February 2025 - August 2025*
Graduate Research Assistant
Advisors: Dr. Martim Brandao, Prof. Claudia Aradu & Dr. John Liddicoat
Responsibilities
- Contributed to the developing of a machine learning system to assess potential risks of emerging AI technologies in the security domain using publicly available patent data from Lens.org
- Developed search strings and annotation protocols to identify and classify security-related AI patents according to risk taxonomies adapted from the EU AI Act framework
- Annotated 250+ patents following standardized protocols, assigning risk categories based on EU AI Act classifications to create training datasets for automated risk assessment
- Conducted statistical analysis of annotated patent data to understand distribution across risk categories and identify patterns in multi-category classifications

Current Projects

Benchmarking LLM-Controlled Robots in India *2025–Present*
- Developing a India-specific benchmark for **LLM-integrated robotic systems** operating in domestic and professional contexts.
- Constructing a **community-driven framework** incorporating anticipatory risks from LLM driven robots across caste, religion, gender, and spatial inequality.

- Jailbreaking robot specific LLMs (Jackal, Dolphin and Unitree Go), black and white-box LLMs as well as Indic LLMs using PAIR algorithms for measuring susceptibility to adversarial attacks in robot planning for Indic risk scenarios.
- Conducted participatory workshops and interviews to co-design mitigation algorithms following constitutional AI to prevent unsafe LLM-robot interactions.

Building IPV Prevention Datasets for Post-Training AI Companions *2026–Present*

- Extending ICML 2026 Spotlight work by constructing the first **preference modeling dataset** grounded in intimate partner violence (IPV) safety for AI companion systems.
- Synthesizing insights from **IPV survivor testimony, crisis intervention literature, and feminist HCI frameworks** to categorize companion response harms across coercive control, gaslighting, and isolation dimensions.
- Conducting participatory workshops and cognitive interviews with **IPV survivors, advocates, and frontline workers** to ground harm categorization in lived experience.
- Prototyping a **first-person account induction pipeline** to systematically surface survivor narratives as evaluation signal for companion model safety assessments.
- Developing a post-training dataset for harm-based reasoning to align AI companion responses with IPV prevention principles and trauma-informed care standards.

Automated Constitutional Alignment for High-Stakes AI Deployment *2026–Present*

- Developing an **empirically derived constitutional AI pipeline** that automatically generates harm taxonomies from real-world complaint data, replacing hand-authored constitutions with bottom-up, context-grounded alternatives.
- Designing a **persona generation framework** using agglomerative clustering over domain-specific corpora to synthesize stakeholder voices across regulatory, institutional, and affected-community dimensions.
- Implementing a **Habermas Machine** to facilitate structured deliberation across synthesized personas, producing consensus-driven constitutional principles for post-training alignment in high-stakes sociotechnical domains.
- Evaluating pipeline generalizability across deployment contexts where standard RLHF and hand-written constitutions systematically fail to capture local normative structures.

Publications

Conference Publications

1. **"It looks useful, works just fine, but will it replace me?" Understanding Special Educators' Perception of Social Robots for Autism Care in India**
   Ashwini B, **Atmadeep Ghoshal**, Krishnaveni Achary, Venkata Ratnadeep Suri and Jainendra Shukla
   Conference on Human Factors in Computing Systems (CHI), 2024
   Best Paper Honourable Mention Award

2. **Value Alignment in the Global South: A Multidimensional Approach to Norm Elicitation in Indian Contexts**
   Atmadeep Ghoshal, Martim Brandao and Ruba Abu Salma
   Workshop on Bidirectional Human-AI Alignment at the International Conference on Learning Representations (ICLR), 2025

3. **Embodied AI at the Margins: Postcolonial Ethics for Intelligent Robotic Systems**
   Atmadeep Ghoshal, Martim Brandao, Ruba Abu Salma and Sanjay Modgil
   AAAI/ACM Conference on Artificial Intelligence and Ethics in Society (AIES), 2025

4. **An Ethnography of Restaurant Robots in Japan: Promises, Perceptions and Impacts**
   Martim Brandao, Ana Sharko, **Atmadeep Ghoshal**, Zoe Evans, Wenxi Wu and Brain Tshuma
   Annual IEEE/ACM International Conference on Human-Robot Interaction (HRI), 202

5. **From the Field to the Algorithm: Understanding Indian Ethnographers' Perspectives on Responsible AI**
   Anasmita Ghoshal &, **Atmadeep Ghoshal**
   Conference on Human Factors in Computing Systems (CHI), 2026

6. **Uncovering Blindspots for Systemic Safety: Relational Accountability in Maritime Autonomous Systems**
   Atmadeep Ghoshal, Caitlin Bentley, Gordon Meadow, Martim Brandao, David Wavell, Jonatan Scharff Willners, Saumya Srivastava, Ethan Woolf Monino, Kimberly Tam, Henry Duffy & Anasmita Ghoshal
   ACM Conference on Fairness, Accountability, and Transparency 2026 (Accepted)

7. **Position: Responsible AI for AI companions must actively combat violence toward intimate partners**
   Atmadeep Ghoshal, Anasmita Ghoshal, Volodymyr Shevchenko, Ashwini B, Arshia Dutta, Ruba Abu-Salma and Martim Brandao
   International Conference on Machine Learning (ICML) 2026
   Spotlight

Selected Awards and Honors

- ICLR Bi-Align Travel Award by Prolific and Layer 6 AI for attending ICLR'25 *2025*
- King's College London Global Research Grant for **doctoral studies** *2024*
- King's College London NMES Faculty International PhD Studentship for **doctoral studies** *2024-2028*
- **Gary Marsden Travel Award** for attending **CHI '24** *2024*
- **CDNM Predoctoral Fellowship** Award for investigating the use of social robots in the Global South *2022-23*
- **University Gold Medal** for standing first class first in the postgraduate sociology program *2022*
- **University Gold Medal** for standing first class first in the undergraduate sociology program *2020*

Talks

- **Rethinking Responsible AI: Current Approaches, Knowledge Constructions and Why We Need a Decolonial Lens**
  - Microsoft Research Cambridge hosted by **Dr.Advait Sarkar** from the Calc Intelligence Group *Oct 2024*
  - Nokia Bell Labs Cambridge hosted by **Dr.Daniele Quercia** from the Social & Responsible AI Group *Feb 2025*
  - FAccT'25 DC hosted by ACM Conference on Fairness, Accountability, and Transparency *June 2025*