Research Proposal

November 1st, 2024

Bridging AI & Human Cognition for Public Interest: A Human-Centric Co-Governance Framework for Auditing Fairness and Preserving Trust in Digital Ecosystems

Meriem Mehri[1]

[1]Department of Computer Science, Faculty of Engineering Sciences, University College London

***

As AI technologies continue to reshape societal, economic, and technological landscapes, there is a crucial need to develop governance frameworks that prioritize transparency, fairness, and ethical alignment with human values. This proposal seeks to introduce a co-governance model that integrates artificial intelligence (AI) and human cognitive processes to drive responsible and adaptive AI governance across rapidly evolving digital ecosystems. Rather than relying solely on existing frameworks and grading rubrics, which – while beneficial – are insufficient for addressing the full spectrum of risks posed by frontier AI systems, this framework envisions a deeper integration of human factors.

Combining multi-agent reinforcement learning (MARL) for collaborative decision-making with brain/cognitive-inspired fairness auditing and differential privacy, this research proposes a technically rigorous, privacy-preserving governance model grounded in principles of AI alignment. Moving beyond conceptual models, this framework introduces psychometric factors to facilitate human involvement in governance oversight, in order to ensure that AI systems not only meet technical safety requirements but also align with the broader public interest in a sustainable and equitable digital future.

Existing AI safety rubrics and assessment schemes provide useful starting points for assessing governance practices but often lack the granularity needed to account for human-AI interaction and alignment with fundamental rights. This proposal builds upon Human-Centered AI (HCAI) literature to craft a governance model that is actionable for policymaking and that positions AI as an ally in advancing public welfare. By embedding socio-technical and cognitive insights into AI frameworks, it envisions a shift from compliance-based algorithmic control to governance approaches deeply rooted in democratic values and ethical integrity.

Potential research articulations:

- How can we establish governance structures that grant users greater transparency and agency over adaptive, high-stakes AI systems, and ensure these systems align with democratic values and public accountability?
- What concrete, human-centered mechanisms can be developed to enable meaningful user control and influence within complex AI decision-making processes, while addressing both technical and behavioral dimensions of human-AI interactions?
- How can a co-governance model balance the adaptive capabilities of AI with safeguards that prevent erosion of user autonomy, and create clear, enforceable standards for accountability in AI deployment across diverse societal contexts?