Symposium Proposal: Moral and Legal AI Alignment

Daniel Devatman Hromada, Bertram Lomfeld

30.11.2024

Description

The concept of AI alignment is frequently framed as a safety issue: how to mitigate biases in AI models or their training data. However, this perspective fails to account for a deeper truth: there is no such thing as a neutral AI system, just as there is no neutral human thinking. Both human and AI decisions are inherently shaped by prior experiences, which in turn are imbued with normative values. Thus, alignment is not an optional safety feature but a fundamental design question underlying the architecture of any AI system.

AI decisions, like human ones, are shaped by prior experiences and embedded normative values, challenging the notion of neutrality. In a pluralistic world, diverse and conflicting moral, legal, and cultural frameworks further complicate alignment, necessitating compliance with multiple norms and policies.

The symposium explores alignment as a reflexive, ongoing process, emphasizing dynamic human-AI interaction and adaptability to shifting ethical and legal landscapes. Sessions will address moral alignment—ensuring AI reflects diverse human value systems—and legal alignment, focusing on the technical and philosophical challenges of adapting AI to different legal frameworks.

Key themes include pluralism, reflexivity, and the integration of alignment with democratic processes. Symposium aims to advance responsible AI design, enabling systems to navigate diverse norms while contributing to collective societal deliberation. In particular, it aims to address the following themes:

- The situated nature of AI decisions and their embedded normative values.
- Alignment as compliance with diverse moral, legal, and institutional frameworks.
- The challenges posed by pluralism and diversity in moral and legal epistemologies.
- Reflexive, ongoing alignment as a procedural process of human-AI interaction.
- Relation between alignment and democratic processes of information aggregation.
- Evaluation of aligned AI systems.

The symposium intends to reflect on the theoretical, technical, and practical aspects of moral and legal AI alignment. Thus, it will be relevant to researchers in AI ethics, law, philosophy, and computational design, as well as practitioners aiming to build responsible, adaptable AI systems.

Potential invited Speakers

We could think of the following invited speakers, whose expertise aligns with the symposium themes:

- Prof. John Tasioulas, Director of Institute for Ethics in AI, Oxford University.
- Prof. Anca Dragan, Director, AI Safety and Alignment, Google DeepMind and UC Berkeley.
- Prof. Giovanni Sartor, Law and AI, European University Institute, Bologna University and European Research Council.

Further invitations may be issued depending on availability.

Organization

Format and Submission

The symposium is intended to run for one day. This length allows sufficient time to address the theoretical, technical, and practical components of moral and legal alignment while maintaining a focused and engaging schedule for participants.

The symposium will feature a variety of formats, including:

- Invited talks by leading experts.
- Contributed presentations of research papers.
- Interactive panel debates or fishbowl discussions to foster audience engagement.
- Hands-on workshop on moral alignment of mid-sized open-source language models.

The event is not intended as a sequel to any previous conference symposium.

Call for papers

Immediately after acceptation of this symposium proposal, a dedicated web-site will be created as well as recommended LaTex template.

Submissions will be accepted in the form of extended abstracts (2-4 pages) or full papers (6-12 pages). A program committee will review the submissions and rank the submissions according to list of pre-defined objective criteria.

Organizers

- Prof. Daniel Hromada, Berlin University of the Arts, Faculty of Design.
- Prof. Bertram Lomfeld, Free University Berlin, Department of Law and Department of Philosophy.

In spite of being primarily active in disciplines of AI education (Hromada) and Law (Lomfeld), both organizers have a peer-reviewed record of publications addressing notions of “alignment” - labeled as “central problem of roboethics” [3] and morally [4] resp. legally [2, 1] reasoning machines.

Organizational Experience

The organizers have experience in planning and executing academic research meetings, including:

- Organizing international workshops at international conferences or as genuine events, e.g. AI & Law Workshop Series, London (KCL) 2022 and Berlin (FUB) 2023.
- Organizing international workshops at international conferences or as genuine events, e.g. AI & Law Workshop Series, London (KCL) 2022 and Berlin (FUB) 2023; Symposium on Artificial Teacher Avatars (UdK Berin, 2024); Symposium on Reading Acquisition in Era of AI (Einstein Center Digital Future Berlin, 2024).
- Serving on program committees for AI-related events, e.g. Artificial Intelligence in Education (AIED), Durham (2022).
- Hosting panel discussions and facilitating interactive formats, e.g. Transatlantic Seminar on Consumer Law, Technology and Inequality 2022 (Yale, EUI, MPI Hamburg and Free University Berlin), board-room dialogue on Human-Machine Peer Learning (OEB Berlin 2022).

This experience ensures that the symposium will be well-structured and effectively managed to maximize impact.

Program Committee

Immediately after this proposal will be accepted as Symposium at IACAP / AISB 2025, organizers will contact persons affiliated to following institutions and initiatives:

- Center for AI Safery, UC Berkeley
- Mila - Quebec Artificial Intelligence Institute
- University of Bamberg
- Einstein Center Digital Future
- Athena - AI Alignment Mentorship for Women program

as potential program committee members.

References

[1] Christoph Benzmüller, David Fuenmayor, and Bertram Lomfeld. "Modelling Value-Oriented Legal Reasoning in LogiKEy". In: *Logics* 2.1 (2024), pp. 31–78.

[2] Christoph Benzmüller and Bertram Lomfeld. "Reasonable machines: A research manifesto". In: *KI 2020: Advances in Artificial Intelligence: 43rd German Conference on AI, Bamberg, Germany, September 21–25, 2020, Proceedings 43*. Springer. 2020, pp. 251–258.

[3] Daniel Devatman Hromada. "The Central Problem of Roboethics: from Definition Towards Solution". In: *Proceedings of the 1st International Conference of the International Association of Computing and Philosophy*. Aarhus, Denmark: Verlagshaus Monsenstein Und Vannerdat, 2011.

[4] Daniel Devatman Hromada and Ilaria Gaudiello. "Introduction to Moral Induction Model and its Deployment in Artificial Agents". In: *Sociable Robots and the Future of Social Relations*. IOS Press, 2014, pp. 209–216.