The Moral and Political Philosophy of AI

Description

This seminar seeks to examine key moral and political issues raised by the development and deployment of artificial intelligence. These include the challenge of aligning AI with human preferences, values, and norms; the problem of algorithmic power and the permissibility of collecting and analyzing people's data to deliver interventions that shape their behavior; whether creating and distributing deepfakes is inherently wrong; the changing nature of employment and the desirability of a post-work world; and the conditions under which AI systems would themselves have well-being or qualify as moral patients.

Proposed format

This course will be offered as a graduate seminar in the philosophy department. We will meet weekly for two hours to discuss recent articles and manuscripts on the moral and political philosophy of AI. The seminar is primarily aimed at graduate students and faculty members in the philosophy department, but graduate students and faculty members from related fields in the humanities and sciences are welcome to participate in the entire seminar or in particular sessions, as well. The seminar will end with an internal mini-conference, where participants (students, faculty members, and instructors) will have an opportunity to receive feedback on their work in progress related to the moral and political philosophy of AI.

Proposed time

Spring 2024

Instructors

Adriano Mannino is a Postdoctoral Fellow at the Kavli Center for Ethics, Science, and the Public. He holds a PhD in philosophy from LMU Munich.

Michal Masny is a Postdoctoral Fellow at the Kavli Center for Ethics, Science, and the Public. He holds a PhD in philosophy from Princeton University.

Week 1. Introduction to AI

An opinionated introduction to the ambitions, methods, and challenges of artificial intelligence research – ideally delivered together with someone involved in technical AI research, and followed by an extended Q&A session.

- Stuart Russell. (2019). *Human Compatible: Artificial Intelligence and the Problem of Control.* (excerpts)

Week 2. Preference Learning and the Alignment Problem

What does it mean to "align" AI systems with human preferences, values, and norms? What are the main conceptual-cum-normative options on offer to define the problem?

- Iason Gabriel. (2020). Artificial Intelligence, Values, and Alignment. *Minds and Machines*. (27 pages)
- Richard Ngo et al. (2022). The Alignment Problem from a Deep Learning Perspective. *arXiv:2209.00626*. (21 pages)

Week 3. Decision Delegation to AI: The "Moral Proxy" Problem

May we recruit AI systems to do our "moral dirty work"? In particular, should nonconsequentialists opt for a consequentialist programming of artificial "moral proxies"?

- Johann Frick. (2015). Contractualism and Social Risk. *Philosophy and Public Affairs*. (excerpts: Automatic Experiment)
- Brian Talbot et al. (2017). When Robots Should Do the Wrong Thing. In Patrick Lin et al. (eds.), *Robot Ethics 2.0: From Autonomous Cars to Artificial Intelligence*. (12 pages)
- Todd Karhu & Tomi Francis. (ms). Getting Machines to Do Your Dirty Work.

Week 4. Deontological AI

If deontologists should opt for a deontological programming of artificial "moral proxies," what form should the respective goal function take? Should consequentialists endorse a deontological programming on instrumental safety grounds, too?

- William D'Alessandro. (ms). Is Deontological AI Safe?
- Mitchell Barrington. (ms). Absolutist AI.

Week 5. Decision Theory for Advanced Artificial Agents

Will "decision-theoretic pressures" lead advanced AI systems to converge on highly risky dispositions to accumulate resources and power?

- Dmitri Gallow. (ms). Instrumental Convergence?
- Elliott Thornley. (ms). There Are No Coherence Theorems.
- Adam Bales. (2023): Will AI Avoid Exploitation? AGI and Expected Utility Theory. *Philosophical Studies*.

Week 6. Automated Influence and Algorithmic Power I

Is it permissible to use AI systems to collect, integrate, and analyze people's data in order to deliver targeted interventions that shape their behavior?

- Claire Benn & Seth Lazar. (2022). What's Wrong with Automated Influence? *Canadian Journal of Philosophy*. (24 pages)
- Mathias Risse. (2023). *Political Theory for the Digital Age: Where AI Might Take Us.* (excerpts)

Week 7. Automated Influence and Algorithmic Power II

To what extent does algorithmic power threaten fair procedure, equal opportunity, and political equality more broadly? How do phenomena such as filtering, hypernudging, and microtargeting affect democratic communities and their deliberative potential?

- Kathleen Creel & Deborah Hellmann. (2022). The Algorithmic Leviathan: Arbitrariness, Fairness, and Opportunity in Algorithmic Decision-Making Systems. *Virginia Public Law and Legal Theory Research Paper*. (32 pages)
- Thomas Christiano. (2022). Algorithms, Manipulation, and Democracy. *Canadian Journal of Philosophy*. (16 pages).

Week 8. Deepfakes I

What is the evidential status of audio and video recordings, and what role do they play in our testimonial practices? Will the proliferation of deepfakes cause an epistemic apocalypse?

- Regina Rini. (2020). Deepfakes and the Epistemic Backstop. *Philosophers' Imprint*. (16 pages)
- Joshua Habgood-Coote. (2023). Deepfakes and the Epistemic Apocalypse. *Synthese*. (23 pages)

Week 9. Deepfakes II

Is creating or distributing deepfakes of other people inherently objectionable?

- Robert Chesney & Danielle Citron. (2019). Deep Fakes: A Looming Challenge for Privacy, Democracy, and National Security. *California Law Review*. (selections)
- Adrienne de Ruiter. (2021). The Distinct Wrong of Deepfakes. *Philosophy & Technology*. (22 pages)

- Daniel Story and Ryan Jenkins. (2023). Deepfake Pornography and the Ethics of Non-Veridical Representations. *Philosophy & Technology*. (22 pages)

Week 10. The Future of Work I

Should justice be concerned with the distribution of sleep? Could the reliance on AI tools in the workplace and economic institutions lead to alienation?

- Jonathan White. (2021). Circadian Justice. *Journal of Political Philosophy*. (25 pages)
- Kate Vredenburgh. (2022). Freedom at Work: Understanding, Alienation, and the AI-Driven Workplace. *Canadian Journal of Philosophy*. (15 pages)

Week 11. The Future of Work II

Which jobs are most likely to be automated? Are there distinctive goods of work? Would the technological elimination of work be desirable?

- Carl Frey and Michael Osborne. (2017). The future of employment: How susceptible are jobs to computerization? *Technological Forecasting & Social Change*. (16 pages)
- Anca Gheaus & Lisa Herzog. (2016). The Goods of Work (Other Than Money!). *Journal of Social Philosophy*. (20 pages)
- John Danaher. (2019). Chapter 3: Why you should hate your job. In his *Automation and Utopia: Human Flourishing in a World without Work*. Harvard University Press. (34 pages)

Week 12. Moral Patiency and Well-being I

Under what conditions would an artificial intelligence system have well-being or moral status?

- Simon Goldstein and Cameron Domenico Kirk-Giannini. (ms). AI Wellbeing.

Week 13. Moral Patiency and Well-being II

Could and should we create AI "super-beneficiaries" who may be able to enjoy far greater levels of well-being than we do? If we ever create them, what would be the best and most just allocation of resources between humans and AIs?

- Carl Shulman and Nick Bostrom. (2021). Sharing the World with Digital Minds. In Steve Clarke, Hazem Zohny, and Julian Savulescu (eds.), *Rethinking Moral Status*. (21 pages)
- Frank Hong. (ms). Group Prioritarianism: Why AI Should Not Replace Humanity.

Week 14. Mini-Conference

Seminar participants (graduate students, faculty members, and instructors) will have an opportunity to receive feedback on their pertinent work in progress.