Philosophy, AI, and Innovation Grad Seminar

Prof. Philipp Koralus, Brendan McCord

Time: Tuesday 4pm-6pm UK time (followed by drinks reception)
Date: Trinity Term 2025, weeks 1-8 (April 29-June 17)
Location: See venue adjustments for each week below

Description: The seminar will explore issues at the intersection of philosophy, AI, and technological innovation, co-taught by a philosopher and a technologist. The seminar will welcome a variety of visiting discussants from philosophy, computer science, and the technology industry throughout term. The focus will be on how a concern for human flourishing can be embedded in the global technology development pipeline, and on exploring how broader bridges can be built between philosophy and technology. The seminar is primarily aimed at philosophy graduate students and computer science graduate students but participants from other levels and areas are welcome. Topics include: truth-seeking AI, privacy, collective intelligence, decentralization in science and AI, and approaches to human autonomy. The seminar culminates in a clinic to facilitate grant applications for independent summer projects on the themes of the seminar.

Prerequisites: please email HAI Lab aiethics-hailab@philosophy.ox.ac.uk with a (very) brief explanation of your interest in the seminar to reserve a spot, and the subject line "TT Seminar". Space limited to maintain quality of discussion.

Week 1 (April 29) Philipp Koralus (HAI Lab) and Brendan McCord (Cosmos Institute). **Truth-seeking AI**

Readings:

- Mill, *On Liberty*, Ch. 2, "Of Liberty of Thought and Discussion" (excerpts)
- Plato, *Theaetetus*, excerpts (149A-152A; 189A-190A)
- Koralus (2025), "The Philosophic Turn for AI Agents: Replacing Centralized Digital Rhetoric with Decentralized Truth-Seeking". https://arxiv.org/abs/2504.18601
- Sarkar, "AI Should Challenge, Not Obey," (link)

Week 2 (May 6) Philipp Koralus and Jules Desai (HAI Lab). **The Inquiry Complex**
Readings:

- Koralus (2025), "The Philosophic Turn for AI Agents: Replacing Centralized Digital Rhetoric with Decentralized Truth-Seeking". https://arxiv.org/abs/2504.18601

Week 3 (May 13) Helen Nissenbaum (Cornell) and Carina Peng (Apple). **Privacy and the Future of AI**
VENUE: Riverside Lecture Theatre (St Catherines College)
Readings:

- H. Nissenbaum (2019) "Contextual Integrity Up and Down the Data Food Chain," *Theoretical Inquiries in Law* 20:1, 221-256.

- K. Martin, H. Nissenbaum, V. Shmatikov (2025) "No Cookies For You!: Evaluating The Promises Of Big Tech's 'Privacy-Enhancing' Techniques,"Georgetown Law Technology Review, 9 Geo. L. Tech. Rev. 1 (2025)

Week 4 (May 20) (May 20) Vincent Weisser (Prime Intellect). **Decentralization in Science & AI**
VENUE: Riverside Lecture Theatre (St Catherines College)

Readings:

- Polanyi, "Republic of Science" (link)
- INTELLECT–1: The First Decentralized Training of a 10B Parameter Model." (link)
- Accelerating Scientific Breakthroughs with an AI Co-Scientist (link)
- The AI Scientist: Toward Fully Automated Open-Ended Scientific Discovery (link)
- DeepSeek-R1: A Decentralized AI Research Platform. (link)

Week 5 (May 27) Ivan Vendrov (Midjourney). **Collective Intelligence**
VENUE: Mary Sunley (St Catherines College)

Readings:

- Hayek, "The Creative Powers of a Free Civilization"
- Stray, Vendrov, Nixon, Adler, Hadfield-Menell, "What are You Optimizing For? Aligning Recommender Systems with Human Values." (link)

Optional:

- Christiano, "What Failure Looks Like." (link)
- Jordan, "Dr. AI or: How I Learned to Stop Worrying and Love Economics." (link)

Week 6 (June 3) MH Tessler (Google Deep Mind), Chris Summerfield (Oxford and AI Security Institute). **The Habermas Machine**
VENUE: Riverside Lecture Theatre (St Catherines College)

Readings:

- Habermas, *The Structural Transformation of the Public Sphere* (short excerpt)
- Summerfield, et al., "AI Can Help Humans Find Common Ground in Democratic Deliberation." (link)

Optional:

- Summerfield, et al, "How Will Advanced AI Systems Impact Democracy?" (link)

Week 7 (June 10) Brendan McCord (Cosmos Institute) and Bethanie Drake-Maples (Stanford HAI). **AI and Human Autonomy**
VENUE: Riverside Lecture Theatre (St Catherines College)

Readings:

- *Truth of Face, Truth of Feeling, a short story by Ted Chiang. https://archive.org/details/ted-chiang-the-truth-of-fact-the-truth-of-feeling/page/16/mode/2up*

- Humboldt, *The Sphere and Duties of Government*, Ch. 2, “Of the Individual Man and the Highest Ends of his Existence”
- Tocqueville, *Democracy in America*, Volume 2, Part 4, Ch. 6, “What Kind of Despotism Democratic Nations Have to Fear”
- Maples, “Designing for Human Autonomy in an Age of AI” (presentation of research and framework for design)

Week 8 (June 17) Brendan McCord (Cosmos Institute), Philipp Koralus, HAI Lab team. **Project Clinic**.
VENUE: Riverside Lecture Theatre (St Catherines College)

Structured group discussion for those who plan to submit applications for summer funding