Philosophy, AI, and Innovation

Prof. Philipp Koralus, Brendan McCord (Visiting Fellow, St. Catherine's College)
Time and Date: M 4pm-6pm, Trinity Term, weeks 1-8 (April 22-June 10)
Location: Top floor, Porter's Lodge, St. Catherine's College

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 the technology industry throughout term, bringing insights from their time at places including Midjourney, Anthropic, Imbue, Google DeepMind, Story Protocol, ex/ante, and Stripe. The focus will be on how a concern for human flourishing can be embedded in the global technology development pipeline from the ground up, 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 areas are welcome. Prerequisites: please email philipp.koralus@philosophy.ox.ac.uk no later than April 15th with a (very) brief explanation of your interest in the seminar to reserve a spot. Space limited to maintain quality of discussion.

Week 1 (April 22): **Perspectives on AI and human flourishing.**

"We have a duty to be optimistic. Because the future is open, not predetermined and therefore cannot just be accepted: we are all responsible for what it holds. Thus it is our duty to fight for a better world." - David Deutsch

AI has opened a new continent, and humanity is setting foot on its shores. How we reaffirm principles of human flourishing in light of the coming innovations will be crucial. The following are guiding questions for the seminar. What is the conceptual structure of the problem of translating conceptions of human flourishing into AI technology? One important aspect of human flourishing in this regard is *freedom*–but what notion of freedom? How can the best notion of freedom be realised in practice in our technological age?

Readings:

1. Koralus, P. *Reason and Inquiry: The Erotetic Theory*, OUP, sections 1.3, 6.7, 6.8.
2. Hayek, F. *The Constitution of Liberty*, University of Chicago Press, Ch. 1, pp. 57-72.
3. Pettit, P. *Republicanism: A Theory of Freedom and Government*, Ch. 2, pp. 51-73.
4. Berlin, I. "Two Concepts of Liberty," in *Four Essays on Liberty*, sections I, II, and III

Week 2 (April 29): **AI, reasoning, and agency.** Visitor: Matt Boulos, Policy and Safety Lead, Imbue, San Francisco

When we say that a large language model (LLM) can "reason," we're saying a curious thing: instead of predicting an answer right away, the model predicts a logical sequence that leads to an answer. There are two arguments for doing this. The first is that, as an

empirical matter, LLMs are more likely to get to a correct answer when they reason. The second is that models taught to reason appear more likely to succeed at complex tasks, where reasoning can resolve ambiguity or steer the model toward a more sensible answer. Interestingly, we haven't really taught the model to reason in the colloquial sense; instead, we've trained it on logical sequences that we believe are more likely to bring us to correct answers. This understanding is liberating—it shows us that the model is not operating in some realm beyond our understanding and that we can deliberately shape reasoning sequences to get the results we want. The tricky part is knowing what we want. This introduces questions of human agency. Generally, we aim to empower individuals to achieve their goals, making reasoning in our models a means to this end. Achieving this, however, opens a can of worms around acceptable and unacceptable behaviour; the role of a company or developer in policing that behaviour; and the conditions society must furnish to maximise both positive and negative notions of liberty.

Readings:

1. Koralus, P. “Why AI still isn’t getting into Oxford, and why that shows it is astounding.” WIRED Italia, March 2024. (link to English version)
2. Huang, J. et al. “Large Language Models Cannot Self-Correct Reasoning Yet.” (link)
3. Turpin, M. et al. “Language Models Don't Always Say What They Think: Unfaithful Explanations in Chain-of-Thought Prompting.” (link)
4. Lazar, S. “Legitimacy, Authority, and Democratic Duties of Explanation.” (link)

Week 3 (May 6): **Blockchain, liberty, and political philosophy.** Visitor: Jason Zhao, Cofounder, Story Protocol, San Francisco

We live in an era of centralising tendencies, largely driven by technological capacity. States have more power and information at their disposal than ever before. AI is trained on terabytes of data, leading to algorithmic saturation of our lives across work, romance, and entertainment. We will begin with a theoretical analysis of the various problems posed to human freedom from within democracies themselves. Do blockchain technologies represent a potential technological antidote to these risks? This lecture will cover the visions of the early cypherpunks, as well as their contemporary adherents. From immutable money to networked states, we will discuss the divergent political philosophies behind blockchain, and whether it can live up to its promise of securing individual liberties via cryptography.

Readings:

1. Mansfield, H. *Tocqueville: A Very Short Introduction*, OUP, Ch. 4, pp. 57-83.
2. Nakamoto, S. “Bitcoin whitepaper.” (link)
3. Balaji, S. “*The Network State* in One Essay.” (link)
4. Hayek, F. “The Use of Knowledge in Society.” (link)

Week 4 (May 13): **A model for parity and its relevance to AI**. Visitors: Kit Fine, New York University; Ruth Chang, University of Oxford; Nick Hawes, University of Oxford

We propose a model of parity in terms of approximate differences and approximate quotients and show how it can be used to facilitate communication between AI and its users when hard choices need to be made.

Readings:

1. Chang, R. (2017). "Hard Choices."

Week 5 (May 20): **Public deliberation and the correction of collective errors.**

For Publius of the Federalist Papers, public deliberation requires that "the cool and deliberate sense of the community ought, in all governments, and actually will, in all free governments, ultimately prevail over the views of its rulers." In this session, we will address the fate of this high standard for free government in the works of those who embraced and extended its vision. Then, we will examine the fate of public deliberation under AI, assessing both a critical approach and a more positive vision for its potential renaissance.

Readings:

1. Polanyi, M. "Republic of Science: Its Political and Economy Theory." (link)
2. Mill, J.S. *On Liberty*, EconLib, Ch. 2. (link)
3. Coeckelbergh, M. "Democracy, epistemic agency, and AI: political epistemology in times of artificial intelligence." (link)

Week 6 (May 27): **Recommender systems, human agency, and collective intelligence.**
Visitor: Ivan Vendrov, Researcher, Midjourney, NYC

As a social species, key aspects of human flourishing and human intelligence depend upon our relation to the collective. How can we build technologies that help human groups to communicate and act together? Why has there been so little progress in useful collective technologies? How does this relate to the Silicon Valley ethic of empowering individuals by maximising optionality, or the market dynamics that lead to each software product trying to capture as much human attention as possible? Language models seem poised to be more important than the printing press in changing the information architecture of society, with major consequences for politics, religion, and human flourishing.

There will be challenging trade-offs between markets, bureaucracy, and democracy as ways of organising human collective action. The question is how technological progress will likely influence these tradeoffs, and what we can do about it.

Readings:

1. Stray, J. Vendrov, I, et al. "What are you optimizing for? Aligning Recommender Systems with Human Values." (link)
2. Christiano, P. "What failure looks like." (link)

3. Jordan, M.I. “Dr. AI or: How I Learned to Stop Worrying and Love Economics.” (link)

Week 7 (June 3): **A framework for thinking about the challenges and opportunities associated with human learning and AI tutors.** Visitor: Michael Strong, Founder, Socratic Experience, Austin

If we recognise the extent to which human beings are genetically programmed to be cultural creatures, that it is the water in which we swim, so to speak, then issues associated with AI and learning become much different and more interesting than is usually recognized. Currently, most of edtech, including AI-powered edtech, is focused narrowly on the pedagogy of instruction, e.g. "How does the learning bot help respond to a question on the mathematics curriculum?" If we care about using AI to educate a higher percentage of humans globally at a lower cost, then we must consider the cultural embeddedness of learning and how to optimise learning for differing cultural contexts, broadly construed. This line of inquiry can dovetail with the role of questions in creating a culture of learning, either in a human classroom/school or in a strictly AI environment.

Readings

1. Henrich, J. "A Cultural Species: Why a theory of culture is required to build a science of human behavior," pp. 25-30. (link)
2. Karlsson, H. "AI tutors will be held back by culture." (link)
3. Gatto, J.T. "The Seven Lesson Schoolteacher." (link)

Week 8 (June 10): **Part 1: Agentic tech and countering digital authoritarianism.** Visitor: Zoe Weinberg, Founder & Managing Partner, ex/ante, NYC

We will examine the technological developments (proprietary software, closed ecosystems, cloud computing, network effects) that have progressively stripped individuals of agency in their interactions with technology, and the authoritarian consequences of those developments (censorship, surveillance, disinformation). Finally, this session will look at “ways out” of the status quo, including the “agentic tech” movement, which seeks to prioritise human agency in future technological innovation.

Part 2: The power of open-ended systems. Visitor: Alex Komoroske, founder, stealth startup; former Head of Corporate Strategy, Stripe, San Francisco

Systems that are alive escape the control of their creator. This makes them open-ended: they can achieve outcomes that none of the participants ever imagined. They harness a core asymmetry that means that as long as any member of the swarm has the right idea, the entire swarm can benefit. However, they are not without their downsides. For example, decentralisation makes coherent outcomes extremely difficult to coordinate. We will consider the fundamental trade-offs of decentralisation and where the cost is worth it, based on extensive real-world experience in open-ended software ecosystems.

Readings

1. Weinberg, Z. “Agentic tech to counter digital authoritarianism.” (link)

2. Estrin, J. “Authoritarian Technology: Attention!” (link)
3. Nissenbaum, H. “Privacy as contextual integrity.” (link)
4. Tang, A., Weyl, G. *Plurality*, Parts 1 and 2 (Preface & Introduction). (link)
5. Komorske, A. “The Meaning of Open.” (link)
6. Brander, G. “Aggregators aren’t open-ended.” (link)

Biographies of Instructors

Philipp Koralus is the Fulford Clarendon Professor of Philosophy and Cognitive Science at the University of Oxford. He studies the human capacity for reasoning and decision-making and how it relates to artificial agents and large language models like GPT. His recent book *Reason and Inquiry* presents a theory of this capacity and its two-faced nature: On the one hand, we are subject to systematic fallacies and framing effects, empirically documented in psychology and behavioural economics. On the other hand, we largely get things right and are capable of incredible feats of rationality. Philipp is a Senior Research Associate at the Institute for Ethics in AI and a Fulford Fellow at St. Catherine’s College.

Brendan McCord is the founder and Chair of the Cosmos Institute and a Visiting Fellow at St. Catherine's College at the University of Oxford. In the private sector, Mr. McCord was the founding CEO of two AI startups that were acquired for $400 million, founder of a quantitative hedge fund, President of an AI lab, and executive responsible for data and AI as President of a large company. In the public sector, Mr. McCord was the principal founder of the first applied AI organization for the U.S. Department of Defense and the author of its first AI strategy. He holds an SB from MIT and MBA from Harvard and is a founding Advisory Board Member of *Harvard Data Science Review*.

Biographies of Visitors

Matt Boulos leads policy and safety at Imbue, an independent research lab developing AI agents that solve problems in the real world. He's a lawyer, computer scientist, and the founder of Canada's first anonymous online legal service as well as a YC-backed startup. At Imbue, he works to ensure a fair and broad societal distribution of AI’s benefits. Matt earned a Bachelor’s in Computer Science and International Relations from the University of Toronto and a JD from Harvard Law School.

Jason Zhao is the co-founder of Story Protocol, a blockchain-based open IP infrastructure based in San Francisco. He is a former Product Manager at DeepMind and angel investor in frontier tech startups. He founded Stanford Rewired and taught at Stanford's d.school as a Design Fellow. He has lectured at Stanford GSB, UATX, and Google. He is an inaugural John Stuart Mill Fellow at the Mercatus Center. Jason received a Bachelor’s in Philosophy and a Master’s in Computer Science at Stanford University.

Kit Fine is University Professor and Silver Professor of Philosophy and Mathematics at New York University. His specialization is in Metaphysics, Logic, and Philosophy of Language. He is a fellow of the American Academy of Arts and Sciences, and a corresponding fellow of the British Academy. In addition to his primary areas of research, he has written papers on ancient philosophy, linguistics, computer science, and economic theory.

Ruth Chang is Chair of Jurisprudence at the University of Oxford and a Professorial Fellow at University College, Oxford. Before that, she was a Professor of Philosophy at Rutgers University, New Brunswick, NJ. Chang's research concerns the nature of normativity, the structure of values and reasons, practical reason, agency, rationality, population ethics, love, commitment, decision-making, and the self.

Nick Hawes is Professor of AI and Robotics in the Oxford Robotics Institute, part of the Department of Engineering Science at the University of Oxford. Nick has a background in AI for autonomous systems, particularly mission planning for autonomous robots operating in everyday environments. He has experience leading teams to deploy autonomous robots into application domains including care, logistics, and security, and to integrate multiple AI subsystems, from computer vision to learning from demonstration and dialogue.

Ivan Vendrov is leading the collective intelligence team at Midjourney, which is building AI tools to help people better understand and coordinate with each other. Previously, he was a member of the technical staff at Anthropic, working on the safe deployment of advanced AI systems. Prior to Anthropic, he was the founder and CTO of Omni and a researcher at Google Research and the University of Toronto. Ivan received his Bachelor's (double honours) in mathematics and computer science from the University of Saskatchewan and a Master's in computer science from the University of Toronto.

Michael Strong is the founder of Socratic Experience, a secondary school for parents seeking a warm, motivating, high-performance virtual educational experience for their children. Prior to Socratic Experience, he founded several successful and innovative schools over the past 28 years based on Montessori, Socratic, and entrepreneurial principles. He is the author of *The Habit of Thought: From Socratic Seminars to Socratic Practice* and lead author of *Be the Solution: How Entrepreneurs and Conscious Capitalists Can Solve All the World's Problems*.

Zoe Weinberg is founder and managing partner of ex/ante, an early-stage venture fund to build agentic tech that counters digital authoritarianism. Previously, Zoe served on the National Security Commission on Artificial Intelligence and at Google AI. Prior to that, she focused on national security, working in Mosul Iraq during the counter-ISIL operation in 2017. She was an investor at the International Finance Corporation of the World Bank and at Goldman Sachs. Her work on national security has been published in the *New York Times* and *Foreign Affairs* and she co-hosts the podcast *Next in Foreign Policy*. She is an inaugural John Stuart Mill Fellow at the Mercatus Center. Zoe received her Bachelor's from Harvard, MBA from Stanford, and JD from Yale.

Alex Komoroske is founder of a startup in stealth that will reimagine the web for the AI era. Previously, he was Head of Corporate Strategy at Stripe. He is a strategic leader who merges the practice, theory, and mindset necessary to tackle complex problems. After studying the emergent power dynamics of open-source communities, he went to Google as a Product Manager. Over 13 years at Google, he worked on Search, DoubleClick, led Chrome's Open Web Platform team for 8 years, led Augmented Reality in Google Maps, and developed new toolkits to align company-wide strategy from the bottom up. He received his Bachelor's in social studies and computer science from Harvard.