Exeter College Oxford Summer Programme

The Ethics of Artificial Intelligence

Course Description

For the first time in history, non-biological entities appear capable of generating, processing, and sharing knowledge in a fashion that closely resembles human agency. But what exactly is happening here? What is the moral and epistemic status of AI systems? Are they genuine agents, or are they mere information processors? Can we trust their outputs in forming beliefs and making decisions? The urgency of answering these questions cannot be overstated: Artificial Intelligence systems have quickly become central parts of our lives. This course is an introduction to AI ethics, focusing on fundamental philosophical questions about the nature and implications of artificial intelligence. It will be structured around twelve topics, such as: What is the nature of artificial intelligence? Can AI systems be moral agents? Can AI systems have minds, consciousness, or knowledge? How can we collaborate with AI systems while maintaining trust and avoiding discrimination? Can AI systems speak meaningfully and refer to the world? Do AI systems pose an existential threat to humanity? In addition, we will discuss some of the most pressing concerns in contemporary AI ethics, including trustworthiness, explainability, discrimination, alignment with human values, and the potential risks posed by advanced AI systems.

Course Objectives

The aim of the course is to introduce students to central topics in the ethics of AI. At the end of the course students should:

- understand central issues in the ethics of AI, including central concepts, theses, arguments and positions
- understand relevant relations between these issues
- be able to clearly and concisely expound these issues in their own words
- be able to critically discuss questions and assess arguments relating to these issues
- be able to offer their own well-founded views on these issues

Teaching Methods

- 12 x 1.25hr Lectures (15 hrs)
- 6 x 1.25hr Seminars (7.5 hrs)
- 4 x 1.25hr Tutorials (5 hrs)

Lectures introduce the material, and seminars provide an opportunity for more in-depth discussion. We will cover one topic per lecture (see below). In addition, there will be six seminars. Each seminar will focus on two readings that are unified by a common theme. Finally, students will attend four tutorials, which are one-on-two meetings between a tutor and two students.

Assessment

This course will be assessed by a three-hour end-of-course exam (45%) in addition to an essay (2,500 - 3,000 words) asking students to formulate and defend a philosophical thesis related to AI ethics (45%). Participation in seminar and tutorial discussions will account for the remaining 10% of the final grade.

Lecture Topics

1. AI Minds: Can AI systems have minds? Required Reading: 1.
2. AI Consciousness: Can AI systems have consciousness? Required Reading: 2.
3. AI Meaning: Can AI systems speak meaningfully and refer to the world? Required Reading: 3.
4. AI Assertion: Can AI systems make assertions? Required Reading: 4.
5. AI Knowers: Can AI systems be knowers? Required Reading: 5.
6. AI Opacity: How can we acquire knowledge from AI systems? Required Reading: 6.
7. Trustworthy AI: Can AI systems be trusted? Required Reading: 7.
8. Human-AI Collaborations: How can we collaborate with AI systems? Required Reading: 8.
9. AI and Discrimination: Do AI systems systematically discriminate? Required Reading: 9.
10. AI Alignment: How can AI systems align with human values? Required Reading: 10.
11. AI Moral Agents: Can AI systems be moral agents? Required Reading: 11.
12. AI Threat: Do AI systems pose an existential threat to humanity? Required Reading: 12.

Seminar Topics

1. AI and Mind. Readings: 1,
2. AI and Language. Readings: 3, 4
3. AI and Knowledge. Readings: 5, 6
4. Human AI Relations. Readings: 7, 8
5. AI Norms and Values. Readings: 9, 10
6. Artificial and Human Moral Agents. Readings: 11, 12

Required Readings

All readings will be made available online before the first lecture. There is one reading per lecture.

1. Goldstein, S. & Levinstein, B.A. Does ChatGPT Have a Mind? arXiv preprint *arXiv*:2407.11015 (2024).
2. Butlin, P., Long, R., Elmoznino, E., et al. (2023). Consciousness in Artificial Intelligence: Insights from the Science of Consciousness. arXiv preprint *arXiv*:2308.08708.
3. Bender, E. M., & Koller, A. (2020). Climbing towards NLU: On Meaning, Form, and Understanding in the Age of Data. In *Proceedings of the 58th Annual Meeting of the Association for Computational Linguistics* (pp. 5185–5198).
4. Butlin, P., & Viebahn, E. (2025). AI Assertion. *Ergo*, 12, 969–988.
5. Kelp, C. and Simion, M. (2026). *Knowledge and Artificial Intelligence*. Cambridge University Press.
6. Messeri, L., & Crockett, M. J. (2024). Artificial intelligence and illusions of understanding in scientific research. *Nature*, 627(8002), 49–58.
7. Durán, J. M., & Formanek, N. (2018). Grounds for Trust: Essential Epistemic Opacity and Computational Reliabilism. *Minds and Machines*, 28(4), 645-666.
8. Simion, M., & Kelp, C. (2023). Trustworthy artificial intelligence. *Asian Journal of Philosophy*, 2(1), 1–12.
9. Johnson, Gabbrielle M. (2025). The hard proxy problem: proxies aren't intentional; they're intentional. *Philosophical Studies* 182 (5):1383-1411.
10. Kasirzadeh, Atoosa and Gabriel, I. (2023). In Conversation with Artificial Intelligence: Aligning language Models with Human Values. *Philosophy and Technology* 36 (2):1-24.
11. van Wynsberghe and S. Robbins, 'Critiquing the Reasons for Making Artificial Moral Agents,' *Science and Engineering Ethics*, vol. 25, no. 3, pp. 719–735, 2019.
12. Cappelen, H., Goldstein, S., and Hawthorne, J. (2025). AI Survival Stories: A Taxonomic Analysis of AI Existential Risk. *Philosophy of AI*, 1.

Optional Readings

1. Cappelen, H., Dever, J. (2026). *Going Whole Hog: A Philosophical Defence of AI Cognition*. Oxford: Oxford University Press.
2. Cappelen, H. and Dever, J. (2021). *Making AI Intelligible: Philosophical Foundations*.
3. Cappelen, C. and Sterken, R. (eds., 2026) *Communicating with AI: Philosophical Perspectives*. Oxford: Oxford University Press.
4. Carter, J.A. (2024). *Digital Knowledge: A Philosophical Investigation*. New York: Routledge.

5. Hicks, M., Humphries, J., and Slater, J. (2024). ChatGPT Is Bullshit. *Ethics and Information Technology* 26/38.
6. Simion, M. (2024). Knowledge and Disinformation. *Episteme* 21: 1208-1219.
7. Vallor, S. (2016). *Technology and the Virtues; A Philosophical Guide to a Future Worth Wanting*. Oxford: Oxford University Press.
8. Vallor, S. (2024). *The AI Mirror: How to Reclaim Our Humanity in an Age of Machine Thinking*. Oxford: Oxford University Press.

NOTE: This is a fast-evolving field. More readings will be made available in advance of the Summer School.