Artificial Intelligence and Normative Theory
PHIL3073

Convener: Nick Schuster
nick.j.schuster@anu.edu.au

Description: This course addresses moral, social, and political questions about artificial intelligence (AI). After an overview of AI aimed at non-specialists, we'll begin our normative inquiry with the question, "can AI systems be moral agents?" We'll then turn to the closely related question of how these complex machines should be programmed and implemented for high-stakes applications. This in turn raises questions about fairness in algorithmic systems. After grappling with the technical and social dimensions of algorithmic fairness, we'll transition to the more general challenge of value alignment. We'll focus especially on how the general public might have a say in which values get embedded in AI systems and how the use of AI systems in governance might be legitimized. Finally, we'll consider how AI systems stand to affect human agency and autonomy as they become increasingly pervasive and impactful.

Content Advisory: This course will discuss topics that students might find distressing, including racism, sexism, and violence. The instructor is primarily responsible for ensuring a safe and inclusive learning environment, but students' cooperation is essential for civil and productive discussion of such issues. Thanks in advance!

Learning Outcomes: Upon successful completion, students will be able to:

1. Demonstrate familiarity with normative issues related to artificial intelligence.
2. Argue for a philosophical position related to the material covered in the course.
3. Display skill in writing research papers in philosophy.
4. Discuss ideas verbally and engage in interactive dialogue.

Assessment Summary:

Assessment Task	Value	Due Date	Return Date	Learning Outcomes
Essay 1	40 %	22 March	31 March	1, 2, 3
Essay 2	40 %	24 May	16 June	1, 2, 3
Participation	20 %			1, 2, 4

Essays: Each essay will be approximately 2000 words long and will reconstruct, challenge, and assess an argument from one of the course readings. Detailed assignments will be circulated well in advance of due dates.

Participation: You are expected to participate in tutorial discussions (10% of final grade). You may miss 2 tutorials without penalty. Additionally, you will submit 10 short reading responses by the end of the semester (1% of final grade each). These responses will be approximately 100 words long and will focus on what you found most confusing, dubious, or controversial in the assigned readings for the week. Responses are due 24 hours before the weekly lecture. You may submit only one response per week. Since this course includes 12 lectures, this means that you may choose two weeks to not submit a reading response. However, you can earn extra credit for submitting additional responses.

Reading Schedule:

Week	Material

1. Introduction and Overview	• PBS Crash Course: Artificial Intelligence Episodes:
	  ○ 2. Supervised Learning
	  ○ 6. Unsupervised Learning
	  ○ 9. Reinforcement Learning
	• Virtual lecture: "Ethics & Politics of ML"

2. Introduction and Overview (cont'd)	• PBS Crash Course: Artificial Intelligence Episodes:
	  ○ 3. Neural Networks
	  ○ 4. Backpropagation and Optimization
	  ○ 7. Natural Language Processing
	• Agüera y Arcas, "Do Large Language Models Understand Us?"

3. Artificial Moral Agents	• Floridi & Sanders, "On the Morality of Artificial Agents" • Véliz, "Moral zombies: why algorithms are not moral agents"
4. Artificial Moral Agents (cont'd)	• Talbot, Jenkins, & Purves, "When Robots Should Do the Wrong Thing" • Cameron et al, "The Social Triad model of Human-Robot Interaction"
5. Algorithmic Fairness	• PBS Crash Course: Artificial Intelligence Episode 18: Algorithmic Bias and Fairness • Lum & Isaac, "To Predict and Serve" • Binns, "Fairness in Machine Learning: Lessons from Political Philosophy"
6. Algorithmic Fairness (cont'd)	• Fazelpour & Lipton, "Algorithmic Fairness from a non-ideal perspective" • Selbst et al, "Fairness and abstraction in sociotechnical systems"
7. Value Alignment	• Gabriel & Ghazavi, "The Challenge of Value Alignment: From Fairer Algorithms to AI Safety" • Russell, "Artificial Intelligence: A Binary Approach"
8. Value Alignment (cont'd)	• Sinnott-Armstrong & Skorburg, "How AI Can Aid Bioethics" • Awad et al., "Crowdsourcing Moral Machines"
9. AI and Governance	• Pascquale, "A Rule of Persons, Not Machines" • Himmelreich, "Against 'Democratizing AI'"
10. Algorithmic Influence	• PBS Crash Course: Artificial Intelligence Episode 15: Recommender Systems • Bhargave & Velasquez, "Ethics of the Attention Economy: The Problem of Social Media Addiction" • Benn & Lazar, "What's Wrong with Automated Influence?"
11. Privacy and Manipulation	• Barocas & Nissenbaum, "Big Data's End Run around Anonymity and Consent" • Susser, Roessler, & Nissenbaum, "Technology, autonomy, and manipulation"
12. The Future of Humanity	• Vallor, "Moral Deskilling and Upskilling in a New Machine Age" • Danaher, "The rise of the robots and the crisis of moral patiency"