PHIL 10167(UG)/PHIL 11186 (PG) Ethics of Artificial Intelligence (2020-2021)

Course organiser: Prof. Shannon Vallor

Course secretary [UG]: Ann-Marie Cowe [philinfo@ed.ac.uk]
Course secretary [PG]: Becky Verdon [Rebecca.Verdon@ed.ac.uk]

Schedule: All teaching of this semester's version of the course will be online. Recordings of lectures and the associated Powerpoint slides will be posted weekly (asynchronous) according to the schedule on the Learn course page. You will also be assigned to a weekly synchronous online tutorial.

Office Hour: TH 17h-18h on Teams (access details to be shared at first tutorial)

COURSE SUMMARY

Artificial intelligence (AI) tools and systems are developing at a rapid pace. We expect to see significant changes in our society as AI systems become embedded in various aspects of our lives. This course will cover philosophical issues raised by current and future AI systems, with a special focus on normative concerns.

Questions we consider include:

- What larger sociotechnical systems, historical forces, cultural values, and power relations have shaped the design, development and use of AI systems, and how might these be shaped by AI in the future?
- What sort of ethical rules, principles, rights or norms should govern AI systems and decisions?
- How do we prevent learning algorithms from acquiring morally objectionable biases?
- Should autonomous AI systems ever be used to kill in warfare, or to make other decisions with irrevocable and morally grave consequences?
- How will AI systems affect human dignity, skills, virtues, purpose, and work?
- What kinds of social roles (e.g. teacher, friend, supervisor, caregiver, lover) are ethically permissible for AI systems to occupy?
- Should AI systems be allowed to deceive or manipulate people, even if for beneficial rather than malicious purposes? Should they be allowed to imitate human emotions?
- Can an AI system suffer moral harm, or be a morally responsible agent?
- Does the future of AI pose an existential threat to humanity? Can we keep the values of AI systems safely aligned with our own?
- How should the benefits and risks of AI systems be distributed in societies and globally?

COURSE AIMS

The aim of this course is to introduce students to a range of ethical issues that arise regarding current and future artificial intelligence (AI), and to develop the critical reflexivity needed to collaboratively investigate and evaluate the moral permissibility or desirability of various technical developments and practical applications of AI. The main questions we will

consider are listed in the course summary. No previous familiarity with the literature on AI will be assumed. While the course readings and lectures will engage the basic functionalities and designs of current AI systems and techniques, the course methodology is qualitative and normative analysis, rather than quantitative/formal methods. Previous coursework in computer science, mathematics and related fields is not required.

The classes will be primarily discussion based, so students are expected to have done the reading in advance of each tutorial. During tutorials, students will work in small teams in breakout rooms to answer a question (approximately 1 per team) based on the reading for the week. They may be instructed to argue for a particular case (pro or contra). They may be asked to assess the merits of a given view. They may be asked to look for counterexamples to a generalisation or fallacies with a specific argument. In second part of the class, we will come together to discuss what each group has achieved to see how it helps us to answer our questions.

Topics covered in class:

- Ethics of AI in government, health, education, transportation, media, finance, and warfare
- Ethics of AI prediction, classification, manipulation, and surveillance of humans
- Ethics of social robots
- AI and the future of work
- Justice and human rights in AI decision-systems
- AI safety, reliability and existential risks
- AI agency, responsibility and moral status

LEARNING OUTCOMES

On completion of this course, the student will be able to:

- Demonstrate knowledge of philosophical issues involved in ethics of AI
- Demonstrate familiarity with relevant examples of AI systems
- Demonstrate ability to bring philosophical considerations to bear in practical contexts
- Demonstrate ability to work in a small team
- Demonstrate skills in research, analysis and argumentation

COMMUNICATION POLICY

Email communication to the instructor is not recommended except in cases where information must be discussed privately. Otherwise, for routine communications and questions for the instructor, please use the Discussion Board forums on LEARN dedicated to *Administrative Questions* (for questions about matters such as the format of assessments, reading schedule, etc.), the forum dedicated to *Philosophical Questions* for general questions related to the course theme but not addressed in the readings or lectures, and the forum named *Course Discussions* to start threads and ask questions concerning lectures, readings, and tutorial activities.

You may also wish to visit the Philosophy student learning support page at https://www.ed.ac.uk/ppls/philosophy/current/undergraduate/handbooks for further

information on department assessment and marking, academic support, and standard reference practices.

CORE TEXTS (e-versions available via Library, see Library Resource List)

Please see the weekly course reading assignments in the Class Readings and Topics table below for specific chapters assigned from these edited collections. Chapters from these core texts will be marked by the following tags at the end of the title: *(in Lin, Abney & Jenkins)*, *(in Dubber, Pasquale & Das)*, or *(in Liao).*

- Lin, P., Abney, K. and Jenkins, R. (2017) *Robot Ethics 2.0*, Oxford Univ Press
- Dubber, M.D., Pasquale, F. and Das, S. eds. (2020), *The Oxford Handbook of Ethics of AI*, Oxford Univ Press.
- Liao, M., ed. (2020) *Ethics of Artificial Intelligence*, Oxford Univ Press.

Additional electronic readings (journal articles, reports, etc.) will also be required or recommended, and are indicated in the weekly table below.

Note: So that you may refer to specific pages and passages easily, please have the required readings for that week (and any recommended readings you have done) open in a browser tab or as a PDF on your computer *before* starting each weekly tutorial session.

ASSESSMENT

Midterm 1500 Words (40%); deadline Thursday 25th February, by 12pm UK time topic and criteria TBA on LEARN page

Final 2500 Words (55%); deadline Thursday 15th April, by 12pm UK time
topic and criteria TBA on LEARN page
Participation (5%) – based on quality of contributions to tutorials/discussion boards as judged by:
- relevance of contributions to course themes
- depth of engagement with ideas from course readings, lectures, and discussions
- constructive engagement with other student discussants

CLASS READINGS AND TOPICS

All course readings are available in electronic, downloadable form through DiscoverED; consult the library resource list on LEARN to access the relevant links.

Required Readings (in bold) are <u>needed</u> in order to benefit from the recorded lecture and to be prepared for the weekly tutorial. Lectures, discussions and tutorial activities will presuppose a careful prior reading of these texts.

Recommended Readings (in italics) will broaden and/or deepen your understanding of that week's theme, as well as allowing you to contribute further to tutorial sessions in those topics that are of special interest to you. They are also relevant sources that you can draw

upon in your midterm and final essays, if you choose. These readings will be referenced during lectures and tutorials but not in a manner that presumes that the entire class has read them.

Research Tips: You may also benefit from consulting some of the interesting-looking readings that are cited in the bibliographies/notes of the required and recommended readings. If you read a chapter or paper that is particularly relevant to the topic of your course essay, or simply appeals strongly to your own interests, it helps to know that in addition to reviewing the works cited within that source, you can also often find a link on Google Scholar or the original journal website that will let you view a list of works that *have since cited that paper or chapter.* This is an excellent way to discover what reviews, objections to and critiques of the work you just read have appeared in the subsequent literature, and/or how the work is being defended, applied or modified by others. These tips are part of the basic toolkit that researchers in philosophy develop, and PG students in particular should practice them whenever possible.

WEEK 1. AI SYSTEMS IN MORAL, POLITICAL AND HISTORICAL CONTEXT

Liao, S. Matthew (2020), "A Short Introduction to the Ethics of Artificial Intelligence," *Ethics of Artificial Intelligence*, Oxford University Press (Introduction p. 1-42 in Liao). * don't worry, the notes/bibliography account for 14 of these pages!

Benjamin, Ruha (2019), Race After Technology: Polity Press (Chapters 1-2, p. 49-96).

WEEK 2. AI TODAY: DOMAINS OF APPLICATION AND KEY ETHICAL ISSUES

Powers, Thomas M. and Ganascia, Jean-Gabriel (2020), "The Ethics of the Ethics of AI," *The Oxford Handbook of Ethics of AI*, Oxford University Press (Chapter I.2 in Dubber, Pasquale & Das).

Donath, Judith (2020), "Ethical Issues in Our Relationship with Artificial Entities," *The Oxford Handbook of Ethics of AI*, Oxford University Press (Chapter I.3 in Dubber, Pasquale & Das).

**Montreal AI Ethics Institute, The State of AI Ethics Report, June 2020*
https://montrealethics.ai/wp-content/uploads/2020/06/State-of-AI-Ethics-June-2020-report.pdf

**Montreal AI Ethics Institute, The State of AI Ethics Report, October 2020*
https://montrealethics.ai/wp-content/uploads/2020/10/State-of-AI-Ethics-Oct-2020.pdf

***The MAIEI 'State of AI Ethics' reports are not recommended as cover-to-cover reading; they are custom-made for browsing/skimming and selective reading. Start by looking through the table of contents for each report. They provide a broad quarterly survey snapshot of recent academic, media, policy and industry perspectives on AI Ethics, so it's a good way to keep up to date on the latest developments and controversies in the field. You can subscribe for free at https://montrealethics.ai/state-of-ai-ethics/ to receive regular quarterly reports.*

WEEK 3. AI-ENABLED PREDICTION, MANIPULATION, CLASSIFICATION AND DISCRIMINATION

O'Neil, Cathy & Gunn, Hanna (2020), "Near Term Artificial Intelligence and the Ethical Matrix," *Ethics of Artificial Intelligence*, Oxford University Press (Chapter 8, p. 237-270 in Liao).

Hoffmann, Anna Lauren (2019), "Where Fairness Fails: Data, Algorithms and the Limits of Antidiscrimination Discourse," *Information, Communication and Society* 22 (7), 900-915.

Susser, Daniel (2019). "Invisible Influence: Artificial Intelligence and the Ethics of Adaptive Choice Architectures." 2019 AAAI/ACM AI Ethics and Society Conference Proceedings, 403-408. (PDF will be uploaded to LEARN)

Pasquale, Frank (2017). The Black Box Society: The Secret Algorithms that Control Money and Information. Chapters 1-2 (1-58).

WEEK 4. AI, AUTOMATION AND HUMAN LABOUR

James, Aaron (2020). "Planning for Mass Unemployment," *Ethics of Artificial Intelligence*, Oxford University Press (Chapter 6, 183-211 in Liao).

Zoller, David (2017). "Skilled Perception, Authenticity and the Case Against Automation," *Robot Ethics 2.0*, Oxford University Press (Chapter 6 in Lin, Abney and Jenkins).

Vallor, Shannon (2017). "AI and the Automation of Wisdom." Philosophy and Computing: Essays in Epistemology, Philosophy of Mind, Logic, and Ethics, ed. Thomas M. Powers. Cham: Springer Chapter 8, p. 161-178.

Danaher, John (2016). "Will Life Be Worth Living in a World Without Work? Technological Unemployment and the Meaning of Life." Science and Engineering Ethics 23(1), 41-64.

WEEK 5. AI, POWER, AND GLOBAL INEQUALITY

Mohamed, Shakir, Png, Marie-Therese & Isaac, William (2020). "Decolonial AI: Decolonial Theory and Sociotechnical Foresight in Artificial Intelligence,"*Philosophy and Technology* (33), 659-684.

Gebru, Timnit (2020). "Race and Gender." *The Oxford Handbook of Ethics of AI*, Oxford University Press (Chapter III.6 in Dubber, Pasquale & Das).

Birhane, Abeba (2020). "The Algorithmic Colonization of Africa." scripted 17(2), 389-409. https://script-ed.org/wp-content/uploads/2020/08/birhane.pdf

WEEK 6. JUSTICE, FAIRNESS AND HUMAN RIGHTS IN AI ETHICS AND GOVERNANCE

Le Bui, Matthew & Noble, Safiya (2020), "We're Missing a Moral Framework of Justice in Artificial Intelligence: On the Limits, Failings and Ethics of Fairness," *The Oxford*

***Handbook of Ethics of AI*, Oxford University Press (Chapter III.1 in Dubber, Pasquale & Das).**

Wong, Pak Hang (2020), "Cultural Differences as Excuses? Human Rights and Cultural Values in Global Ethics and Governance of AI," *Philosophy and Technology* 33, 705-715.

Gabriel, Iason (2020). "Artificial Intelligence, Values, and Alignment," Minds and Machines 30, 411-437.

European Commission High Level Expert Group on Artificial Intelligence (2019): Ethics Guidelines for Trustworthy AI (PDF in public domain)
https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419

WEEK 7. ETHICS OF SOCIAL ROBOTS

Meacham, Darian and Studley, Matthew (2017). "Could a Robot Care? It's All in the Movement,"*Robot Ethics 2.0*, Oxford University Press (Chapter 7 in Lin, Abney and Jenkins).

Isaac, Alistair M.C. and Bridewell, Will (2017), "White Lies on Silver Tongues: Why Robots Need to Deceive," *Robot Ethics 2.0*, Oxford University Press (Chapter 11 in Lin, Abney and Jenkins).

Scheutz, Matthias (2012), "The Inherent Dangers of Unidirectional Emotional Bonds between Humans and Social Robots," Robot Ethics: The Ethical and Social Implications of Robotics, eds. P. Lin, K. Abney and G.A. Bekey, Cambridge, MA: MIT Press, 205-221.

Vallor, Shannon (2011). "Carebots and Caregivers: Sustaining the Ethical Ideal of Care in the Twenty-First Century." Philosophy and Technology 24: 251-268.

WEEK 8. THE ETHICS OF ARTIFICIAL MORAL AGENCY

Giubilini, Alberto & Savulescu, Julian (2017), "The Artificial Moral Advisor: The 'Ideal Observer' Meets Artificial Intelligence," *Philosophy and Technology* 31, 169-188.

Van Wynsberghe, Aimee & Robbins, Scott (2018), "Critiquing the Reasons for Making Artificial Moral Agents, *Science and Engineering Ethics* 25(3), 719-735.

Poulsen, Adam, Anderson, Susan Leigh, Anderson, Michael, Ben Byford, Fabio Fossa, Erica L. Neely, Alejandro Rojas & Alan Winfield (2019), "Responses to a Critique of Artificial Moral Agents." Cornell University. arXiv preprint at https://arxiv.org/abs/1903.07021

Wallach, Wendell & Allen, Colin (2009), Moral Machines: Teaching Robots Right from Wrong. New York: Oxford University Press, 13-42.

WEEK 9. AI MORAL STATUS AND RIGHTS

Basl, John & Bowen, Joseph (2020). "AI as a Moral Right-Holder" ***Oxford Handbook of Ethics of AI*****, Oxford University Press (Chapter III.8 in Dubber, Pasquale & Das).**

Liao, S. Matthew (2020). "The Moral Status and Rights of Artificial Intelligence" ***Ethics of Artificial Intelligence*****, Oxford University Press (Ch 17, 480-503 in Liao).**

Coeckelbergh, Mark (2010). "Robot Rights? Towards a Social-Relational Justification of Moral Consideration. Ethics and Information Technology 12(3), 209-221.

Bryson, Joanna & Dignum, Virginia (2018). "Patiency is Not a Virtue: The Design of Intelligent Systems and Systems of Ethics." Ethics and Information Technology 20, 15-26.

Birhane, Abeba & van Dijk, Jelle (2020). "Robot Rights? Let's Talk About Human Welfare Instead." Proceedings of the AAAI/ACM Conference on AI Ethics and Society 2020, 207-213. PDF will be uploaded to LEARN.

WEEK 10. SAFETY, LETHALITY AND UNCERTAINTY IN AUTONOMOUS AI SYSTEMS

Asaro, Peter (2020), "Autonomous Weapons and the Ethics of Artificial Intelligence," (Chapter 7, pp. 212-236 in Liao).

Bhargava, Vikram & Kim, Tae Wan (2017), "Autonomous Vehicles and Moral Uncertainty," ***Robot Ethics 2.0*****, Oxford University Press (Chapter 1 in Lin, Abney & Jenkins).**

Sparrow, Robert (2016). "Robots and Respect: Assessing the Case against Autonomous Weapons Systems," Ethics & International Affairs 30 (1) 93 -116.

Jenkins, Ryan and Purves, Duncan (2016). "Robots and Respect: A Response to Robert Sparrow," Ethics & International Affairs 30 (3) 391-400.

WEEK 11. HORIZONS, TRAJECTORIES AND OPEN QUESTIONS IN AI ETHICS

Coeckelbergh, Mark (2019), "Artificial Intelligence, Responsibility Attribution and a Relational Justification of Explainability." ***Science and Engineering Ethics*** **26(4), 2051-2068.**

Berberich, Nicolas, Nishida, Toyoaki and Shoko Suzuki (2020), "Harmonizing Artificial Intelligence for Social Good," ***Philosophy and Technology*** **33, 613-638.**