THE ETHICS OF DATA AND ARTIFICIAL INTELLIGENCE (ME102)

Course Duration: 54 hours lecture and class time (Over three weeks)

Summer School Programme Area: Research Methods, Data Science, and Mathematics

LSE Teaching Department: Department of Philosophy, Logic and Scientific Method

Lead Faculty: Dr Kate Vredenburgh, Dr Ali Boyle, Professor Alex Voorhoeve and Dr Paoloa Romero (Dept. of Philosophy, Logic and Scientific Method, LSE)

Pre-requisites: There are no prerequisites for this course. No prior study in philosophy or computer science is assumed, and students from all disciplines - from law to engineering to business - are welcome.

Course Outline:

AI is now embedded in our day-to-day lives, influencing who we date, the new stories we read on our social media feed, how we invest in financial assets, our community's exposure to the police, the goods we consume online, and the tasks we do at work.

Ethics has often been the last step in the design and deployment of AI technologies. But, new and pending regulation, activism by civil society, and self-governance efforts by companies have sought to integrate values like fairness, safety, and privacy throughout the product development process or decision support system design.

This course introduces you to the core ethics concepts needed to build better technology and reason about its impact on the economy, civil society, and government. In the first half of the course, we consider ethical questions raised by different steps in the data science pipeline, such as:

- What is data, and how can we design better (ethical?) data governance regimes?
- Can technology discriminate? If so, what are promising strategies for promoting fairness and mitigating algorithmic bias?
- Can we understand black-box AI systems and explain their decisions? Why is it morally important that we do so?

In the second half of the class, we consider ethical questions raised by the use of AI systems to manage our work, political, and social lives, such as:

- How does automation impact economic inequality?
- Do employees have a right to privacy at work?
- How does AI concentrate power, and when is this concentration of power objectionable?
- How can we embed human values into AI systems?

This course is ideal if you're seeking a practical understanding of the ethical challenges and potential solutions posed by real-world AI systems.

This course is especially beneficial to those targeting a career in data science or computational social science, product management, managerial positions, AI policy, information technology law, or an academic career in a field related to the ethics of AI.

Course Outcomes:

By the end of the course participants will be taught to:

- Understand core ethics concepts and how those concepts apply to AI systems
- Analyse the ethical issues raised by a particular technology by applying core ethical reasoning techniques to real-world case
- Apply cutting-edge ethics research within the development process to build more ethical AI systems
- Communicate your own ethical viewpoint clearly and persuasively by reconstructing others' arguments, objecting to them, and providing your own solution.

Course Structure and Assessment:

- Lectures: 36 hours
- Classes: 18 hours

This course is designed as a combination of lectures, discussions, case studies, and readings. The course emphasises applied, rather than theoretical, issues, using fundamental concepts in ethics to analyse ethical issues raised by particular technologies. Case studies will be used and student participation in lectures and discussion sections is essential to the achievement of the course objectives.

This course is assessed by one essay and one examination: one mid-session essay (50%) and one final exam (50%). You will also write an extended outline, to prepare for the essay, and one class presentation. Neither the outline nor the presentation contributes to your grade but are designed to give you feedback on core skills – argument reconstruction, posing objections – and understanding of course content to improve your performance on the summative assessments.

Texts:
There are no set-texts for this course, indicative readings will be made available electronically.

Lecture schedule and assigned readings:

Note: The assigned readings are merely indicative; they may change before the start of the course.

Unit 1: What's so special about artificial intelligence?

Lecture 1: Justice and the control of technology [KV]

In some scholarly traditions, technology is viewed as *neutral,* or technological development as separate from a society's political or economic institutions. In this lecture, we will consider approaches to technology as neutral and approaches to technology as *value-laden*. We will also examine questions about power, justice, and the control of technology, and the relationship between technology and labor.

Indicative readings:

- Gabriel, "Towards a Theory of Justice for Artificial Intelligence", *Daedalus*
- Friedman, Kahn, and Borning, "Value Sensitive Design and Information Systems"
- Levinson, *The Box* [selections]
- Cohen, *If You're an Egalitarian, How Come You're So Rich?* [selections]

Lecture 2: What is intelligence? [AB]

Demis Hassabis, founder of DeepMind, says his mission is 'to solve intelligence […] to fundamentally understand intelligence and recreate it artificially'. But what exactly is intelligence supposed to be? In this lecture, we'll consider three concepts of intelligence. First, there's the commonsense idea of intelligence – the one we use in everyday life. Second, there's intelligence as it appears in psychology, which theorists have proposed can be measured using constructs like IQ and *g*. Finally, there's intelligence as it appears in artificial intelligence. We'll explore how these concepts relate to one another, what to make of their (in some cases, troubling) history, and whether the project of solving intelligence should be approached with scepticism.

Indicative readings:

- Devin Sanchez Curry (2021). *g* as bridge model. *Philosophy of Science* 88 (5), 1067-1078
- Davide Serpico (2018). What kind of kind is intelligence? *Philosophical Psychology* 31 (2), 232-252
- Davide Serpico & Marcello Frixione (2018). Can the *g* factor play a role in artificial general intelligence research? *Proceedings of the Society for the Study of Artificial Intelligence and Simulation of Behaviour*. 301-305
- Mark Alfano, Latasha Holden & Andrew Conway (2016). Intelligence, race and psychological testing. *The Oxford Handbook of Philosophy and Race* (Oxford: OUP)
- Henry Shevlin, Karina Vold, Matthew Crosby & Marta Halina (2019). The limits of machine intelligence. *EMBO Reports*20, e49177
- Stephen Cave (2017). Intelligence: a history. *Aeon*. https://aeon.co/essays/on-the-dark-history-of-intelligence-as-domination

Lecture 3: Evaluating intelligence in AI systems [AB]

In the early 20th Century, public attention was sparked by a horse named Clever Hans. When Clever Hans was presented with mathematical puzzles, he would beat out the correct answer using his hooves. Eventually, the German board of education commissioned a scientist, Oskar Pfungst, to investigate Clever Hans' abilities. Pfungst discovered that Hans would only get the answers correct when his questioner knew the answer: Hans was responding to subtle behavioural changes from his questioner, which the questioner was unaware of. This case has become famous as a demonstration that apparently intelligent behaviour doesn't always require the sort of intelligence we expect.

If AI researchers are attempting to build intelligent machines, how will they know when they've succeeded? Even the best AI systems do Clever Hans-like things. The best large language models say extremely jarring things, and the best image classifiers are tripped up by imperceptible image distortions. Does this mean they aren't intelligent after all? In this lecture we'll discuss the challenge of AI evaluation and explore one possible solution: evaluating AI systems using frameworks and methods developed over the last century in the science of animal behaviour.

Indicative readings:

- Cameron Buckner (2020). Black boxes or unflattering mirrors? Comparative bias in the science of machine behaviour. *British Journal for the Philosophy of Science*
- François Chollet (2019). On the measure of intelligence. *ArXiv*: 1911.01547
- Matthew Crosby et al. (2020). The animal-AI testbed and competition. *Proceedings of Machine Learning Research* 123, 164-176
- Marta Halina (2021). Insightful artificial intelligence. *Mind & Language* 36 (2), 315-329
- Michael Trestman (2015). Clever Hans, Alex the Parrot and Kanzi: What can exceptional animal learning teach us about human cognitive evolution?

Unit 2: Ethics when building AI systems

Lecture 4: Participatory AI

AI systems model a social world with normativity and with thick concepts, such as human flourishing, the connection between social identity and outcomes, and inequality. This normativity raises a number of problems for technologists. One such problem is that many of these concepts, such as justice, equality, or welfare, are essentially contested: there are a number of permissible versions of the concept that could drive modeling. Another is a problem of legitimacy: AI systems systematically and seriously impact individuals' rights and welfare. We will examine and critique one solution, in the form of participatory AI.

Indicative readings:

- Alexandrova and Fabian, "Democratizing Measurement: Or Why Thick Concepts Call for Coproduction"
- "Envisioning Communities: A Participatory Approach Towards AI for Social Good" Coauthors: Elizabeth Bondi, Lily Xu, Jackson A. Killian. Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society. July 2021

Lecture 5: Data and privacy

Building AI systems requires massive amounts of data. Some scientists claim that big data is fundamentall revolutionising science, and, more generally, the ability to learn from past observations. In this lecture, we will examine whether big data is fundamentally changing the methodology of the social sciences. We will also examine ethical questions about who owns the data, and what sort of governance regimes can protect the rights and interests of data subjects.

Indicative readings:

- Northcott, “Big Data and Prediction: Four Case Studies”
- Simons and Alvarado, “Can we trust Big Data? Applying philosophy of science to software”
- boyd and Crawford, “Six Provocations for Big Data”
- “Privacy,” Stanford Encyclopedia of Philosophy
- Viljoen, “A Relational Theory of Data Governance”

Lecture 6: Fair prediction

For some theorists of science, scientific models and theorists ought to aim at and be assessed by epistemic values alone, such as truth and accuracy. In this week, we will examine questions of what accuracy is, as well as whether non-epistemic values ought to guide algorithm design. One core value that some have argued ought to guide algorithm design is anti-discrimination. We will examine what discrimination in AI could be, whether AI has a greater potential to be more (or less!) discriminatory than human decision-makers, and how to build algorithms to do not discriminate.

Indicative readings:

- Johnson, “Are Algorithms Value Free?”
- Barocas, Hardt, and Narayanan, *Fairness and Machine Learning: Limitations and Opportunities* [selections]
- Dotan, “Theory Choice, Non-Epistemic Values, and Machine Learning”
- Bright, “Dubois’ democratic defense of a value free ideal”
- Douglas, “Rejecting the Ideal of Value-Free Science”

Lecture 7: Explainable AI

Many learned models are highly complex, exploit unintuitive correlations for decision-making, and contain features that are not interpretable by human beings. This has led scientists and policy-makers to worry that such models are not in principle understandable by human beings, and that this lack of explainability raises serious epistemic and ethical concerns. In this lecture, we will examine what explainability might mean, why it is valuable, and whether current AI techniques succeed in advancing those values.

Even if AI models are understandable, however, there are further questions about the kinds of explanations that classificatory models make available. Classificatory models have become dominant in the social sciences and computer science. Furthermore, classifying and counting people into groups is central to policy-making, especially targeted policy to reduce inequality and discrimination. Classification into groups requires choosing and precisifying some properties of group members, and ignoring others, and raises concerns about essentializing. In terms of ethics and policy, classification can elide facts that are important for good decision-making using social science, such as facts about individual responsibility. We will examine whether other paradigms of explanation, such as analytic narratives, can offer an important compliment to classification-based explanations.

Indicative readings:

- Creel, "Transparency in Complex Computational Systems"
- Vredenburgh, "The Right to Explanation"
- Pozen, "Transparency's Ideological Drift", *Yale Law Journal*
- Kiviat, "The Moral Affordances of Construing People as Cases"
- Kiviat, "The Art of Deciding with Data: Evidence from How Employers Translate Credit Reports into Hiring Decisions"

Unit 3: The ethics of deployment: using AI systems for predictions and decisions

Lecture 8: AI, Privacy, and Consent to Personal Data Processing on Social Media.

Social networking services (SNS) enable individuals to expand their social relations to the online environment, for example by connecting with potential employers on LinkedIn, engaging in political debates with strangers on Twitter, or staying up to date with relatives through Facebook. SNS offer a service through which users can build a digital identity, develop relationships, and access opportunities and media. They are also a gateway to a range of goods, services, and applications. SNS typically offer access to their platforms in exchange for the opportunity to collect, process, use, and commercialize users' personal data (or "process personal data," for short). In this lecture and seminar, we investigate under what conditions this exchange is permissible. According to the Privacy Self-Management paradigm, the decision of how much personal information to cede control over to an SNS in return for the benefits provided should be up to the data subjects. The privacy policy of an SNS outlines which forms of personally identifiable information can be collected, how this data is stored, and how it may be used and shared. Users are held responsible for agreeing to this policy when accessing services. On this view, the *consent* of the user plays a large role in justifying the personal data processing practices of SNS. Within wide boundaries, it is taken to be sufficient to make the collection and use of personal data both legally and morally legitimate. We will examine various reasons to question the Privacy Self-Management Paradigm. We will also investigate two alternative views: (1) that valid consent to the collection of personal data requires autonomous consent, which entails that users must be adequately informed and knowledgeable and free from coercion and manipulation; (2) that what renders the trade of personal data for access to services permissible is instead that users face options that they can navigate to their advantage despite their lack of knowledge and their decision-making biases.

Indicative readings:

- Daniel J. Solove, Privacy Self-Management and the Consent Dilemma, 126 Harv. L. Rev. 1880 (2013)
- Wolmarans, L. and Voorhoeve, A. 2022. What Makes Personal Data Processing by Social Networking Services Permissible? Canadian Journal of Philosophy 52: 93–108, doi:10.1017/can.2022.4
- Alessandro Acquisti, Laura Brandimarte and George Loewenstein. Privacy and human behavior in the age of information Science 347 (6221), 509-514. DOI: 10.1126/science.aaa1465
- Franklin Miller and Alan Wertheimer The Ethics of Consent: Theory and Practice Oxford Scholarship Online 2010. DOI: 10.1093/acprof:oso/9780195335149.001.0001

Lecture 9: Surveillance and workplace privacy

Surveillance technologies have given managers and third parties unprecedented levels of oversight over their employees. Workers’ keystrokes are monitored; their movements in cars (Uber; mail delivery) or on the factory floor are tracked in real time; their activities outside of work are tracked on workplace devices. Proponents of such technologies argue that they increase economic productivity; detractors insist they violate workers’ privacy. In this lecture, we will delve deeper into why privacy is valuable, whether workers have a claim to privacy at work, and whether worker privacy is compatible with the modern capitalist workplace.

Indicative readings:

- Shoshana Zuboff (2015) "Big other: surveillance capitalism and the prospects of an information civilization", Journal of Information Technology 30: 75-89
- Anders J. Persson and Sven Ove Hansson, “Privacy at Work: Ethical Criteria”, *Journal of Business Ethics*, Vol. 42, No. 1 (Jan., 2003), pp. 59-70.
- Harrigan, Michael, “Privacy Versus Justice: Amazon's First Amendment Battle in the Cloud”, *Western State University Law Review*, 2017, Vol.45 (1), p.91
- Jeffrey Reiman, "Privacy, Intimacy and Personhood," *Philosophy & Public Affairs* 6 (1) (1976), pp. 26.

Lecture 10: AI and value alignment

AI is used to inform decisions that impact citizens rights and welfare around the global. Some have argued that AI can only be used in such settings if it is aligned with human values, i.e., if AI either outputs predictions or decisions for reasons that humans can understand and endorse, or if its predictions and decisions promote the relevant values. We will examine why value alignment is claimed to be important, and the different understandings of value alignment, with an emphasis on the case of self-care apps. Furthermore, in a globalised world, this claim for the importance of value alignment raises a number of questions, such as whose values ought to be promoted by AI and whether the values underpinning AI models and AI apps are too limited to a Western understanding of values.

Indicative readings:

- Gabriel, “Artificial Intelligence, Values, and Alignment.” [link:https://www.deepmind.com/publications/artificial-intelligence-values-and-alignment]
- M. J. Dennis (2020) "Cultivating Digital Well-Being & the Rise of Self-Care Apps” in *The Ethics of Digital Well-Being: A Multi-Disciplinary Approach.* C. Burr & L. Floridi (eds.). New York: Springer Publishing
- M. J. Dennis & E. Ziliotti (forthcoming)"Living Well Together Online: Digital Well-Being from a Confucian Perspective.", *Journal of Applied Philosophy*
- Bondi et. al. “Envisioning Communities: A Participatory Approach Towards AI for Social Good”
- Shengnan Han, Eugene Kelly, Shahrokh Nikou & Eric-Oluf Svee, “ Aligning artificial intelligence with human values: reflections from a phenomenological perspective”, *AI & SOCIETY* 37, 1383- 1395 (2022)

Lecture 11: AI and democracy: political discourse and social media

AI has fundamentally re-shaped political discourse in democracies and non-democracies. Social media and encrypted messaging allow members of a political community to coordinate protests; ad targeting allows politicians to deliver contradictory messages to different types of people, or for actors to sow disinformation; the sharing of news on social media puts people in so-called *filter bubbles* or *echo chambers*. In this lecture, we will first survey the evidence on whether social media enhances or in contrast, precludes a robust political arena of discourse and action.

Indicative readings:

- Gorwa et. al "Algorithmic content moderation: Technical and political challenges in the automation of platform governance."
- Manheim, Karl; Kaplan, Lyric "Artificial Intelligence: Risks to Privacy and Democracy", *Yale Journal of Law and Technology* 21 (2019)
- Richard A. Mills "Pop-up political advocacy communities on reddit.com: SandersForPresident and The Donald", *AI & SOCIETY* 33, 39-54 (2018).
- TB Kane, "Artificial intelligence in politics: establishing ethics", IEEE Technology and Society Magazine, 2019

Lecture 12: AI and democracy: regulating power

Legislators and citizens have become concerned that technology companies such as Apple, Google, Meta, Amazon, and Twitter have become too powerful, and that the inequalities of power between companies and consumers and companies and citizens undermines the proper functioning of markets and democracy. We will examine how data and AI enable companies to have outsized market and political power, as well as potential regulatory solutions.

Indicative readings:

- James Smith & Tanya de Villiers-Botha , "Hey, Google, leave those kids alone: Against hypernudging children in the age of big data", *AI & SOCIETY,* (2021)
- Buhmann, A., & Fieseler, C., "Deep Learning Meets Deep Democracy: Deliberative Governance and Responsible Innovation in Artificial Intelligence." *Business Ethics Quarterly,* 1-34, (2022)
- Dirk Helbing, Bruno S. Frey, Gerd Gigerenzer, Ernst Hafen, Michael Hagner, Yvonne Hofstetter, Jeroen van den Hoven, Roberto V. Zicari, Andrej Zwitter, "Will democracy survive big data and artificial intelligence?", *Scientific American* Magazine,February 25, 2017
- Fukuyama, F, "Making the Internet Safe for Democracy", *Journal of democracy*, 2021, Vol.32 (2), p.37-44

THE LONDON SCHOOL OF ECONOMICS AND POLITICAL SCIENCE

Credit Transfer: If you are hoping to earn credit by taking this course, please ensure that you confirm it is eligible for credit transfer well in advance of the start date. Please discuss this directly with your home institution or Study Abroad Advisor.

As a guide, our LSE Summer School courses are typically eligible for three or four credits within the US system and 7.5 ECTS in Europe. Different institutions and countries can, and will, vary. You will receive a digital transcript and a printed certificate following your successful completion of the course in order to make arrangements for transfer of credit.

If you have any queries, please direct them to summer.school@lse.ac.uk