Topics in the Philosophy of AI

Convenor: Mike Stuart

Email: mike.stuart@york.ac.uk

Office location: Department of Philosophy (Sally Baldwin Buildings, Block A), Room 108

Office hours: Tuesdays, 11:00am – 12:00pm

Seminar details

Time: 14:00 - 16:00

Location: BS/008 Seminar Room, Campus West - Berrick Saul Building

See details about the module here.

Course content

This course will explore social, political, moral, metaphysical, and epistemological issues surrounding artificial intelligence. We will explore questions like: What would it take for machines to have subjective experiences? Could machines deserve moral treatment? Can machines create art? How have new technologies affected the roles of traditionally marginalized groups? Can technology be racist? How does technology affect our social interactions with each other? What can we learn about the human mind by inventing intelligent machines?

Communication

Course updates will be communicated by Announcements made on the VLE site.

To contact the instructor concerning course content, please email mike.stuart@york.ac.uk. Replies should come within 24-48 hours.

Students may also come to the instructor's office hours, Tuesdays from 11:00 to 12:00 in the Department of Philosophy (Sally Baldwin Buildings, Block A), Room 108.

Online appointments may also be scheduled by email.

Accessibility

We try to make course content as accessible as possible, but we are always keen to improve accessibility further, so please let the instructor know if there are additional ways we can support your learning process.

If you don't already have a student support plan in place, and you feel that you have specific learning/accessibility needs, please consider exploring whether this may be a good option for you. Your supervisor can help you navigate this.

Learn more about accessibility in the VLE, Reading Lists and in your department: Department of Philosophy Accessibility Statement

Departmental accessibility contact

If you have any accessibility queries or problems accessing module materials in a suitable format, please contact the accessibility contact for Philosophy:

- Email: philosophy@york.ac.uk

Use Ally to convert VLE content and documents to alternative formats by clicking on the Ally 'A' icon at the top of the page or next to the document.

You can use different file formats that are more accessible for you or to help you study effectively in different situations. For example, you could convert a PDF into an audio file so you can listen to it if your eyes are tired or you're walking, or you could convert it to an e-reader file that's easier to read on a tablet or to read online.

More information about Ally

Disability support

Contact the Disability Support Team to access academic support and adjustments if you have a disability or long-term health condition that has an impact on your ability to study.

Reading list

Week 1: Introduction to the course

Recommended

- 'Philosophers on GPT-3' Daily Nous
- Philosophers on Next-Generation Large Language Models' Daily Nous

Week 2: Foundations and definition of AI

Essential

- Norvig, P. and Russell, S. 2009. Artificial Intelligence: A Modern Approach, Prentice. Pp. 1-33.

Recommended

- Boden, Margaret. 2016. AI: Its Nature and future. Oxford University Press. Chapter 1.
- On Turing machines [YouTube video]
- On 'virtual machines' [YouTube video]
- Simulations of neural nets [YouTube video]
- Dasgupta, Subrata. 2016. *Computer Science: A Very Short Introduction*. Oxford: OUP.

Week 3: Modern AI and how it works

Essential

- Buckner, C. 2018. "Empiricism without magic: transformational abstraction in deep convolutional neural networks." *Synthese* 195, 5339–5372 (2018). https://doi-org.libproxy.york.ac.uk/10.1007/s11229-018-01949-1.

Recommended

- On neural networks [YouTube video]
- Deep learning [YouTube video]
- Gradient descent [YouTube video]
- Backpropagation [YouTube video]
- Generated Adversarial Networks (GANs) [YouTube video]
- Genetic Algorithms [YouTube video]
- Sheng Lu, Irina Bigoulaeva, Rachneet Sachdeva, Harish Tayyar Madabushi, and Iryna Gurevych. 2024. "Are Emergent Abilities in Large Language Models just In-Context Learning?" In Proceedings of the 62nd Annual Meeting of the Association for Computational Linguistics (Volume 1: Long Papers), pages 5098–5139, Bangkok, Thailand. Association for Computational Linguistics. https://aclanthology.org/2024.acl-long.279.
- Stanford's State of AI report 2023 (at least the chapter summaries, pages 11-19)

Background

- ‘A Beginner’s Guide to Neural Networks and Deep Learning’ A.I. Wiki.
- Vaswani, Ashish, et al. 2017. "Attention Is All You Need." 31st Conference on Neural Information Processing Systems, Long Beach, CA, USA.

Week 4: How close is AGI?

Essential

- Felin, Teppo and Holweg, Matthias. 2024. "Theory Is All You Need: AI, Human Cognition, and Causal Reasoning." Available at SSRN: https://ssrn.com/abstract=4737265 or http://dx.doi.org/10.2139/ssrn.4737265

Recommended

- Bostrom, N. 2012. ‘The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents.’ *Minds and Machines* 22: 71–85.
- Dehaene, S., Lau, H. and Kouider, S. 2017. ‘What is consciousness, and could machines have it?’ *Science* 358: 486-492.
- Oktar, Kerem; Sucholutsky, Ilia; Lombrozo, Tania; Griffiths, Thomas L. 2024. "Dimensions of disagreement: Divergence and misalignment in cognitive science and artificial intelligence." *Decision*. 10.1037/dec0000244.
- Vaidya, A.J. 2024. "Can machines have emotions?" *AI & Soc*. https://doi-org.libproxy.york.ac.uk/10.1007/s00146-024-02022-x.
- Dennett, D. 1984. ‘Cognitive wheels: The frame problem of AI’ In Christopher Hookway (ed.), Minds, Machines and Evolution. Cambridge University Press.
- Bermudez, Luis. 2021. ‘Overview of Embodied Artificial Intelligence.’ *Medium*.
- Harnad, S. 1990. ‘The Symbol Grounding Problem.’ *Physica D* 42: 335-346.
- Taddeo. Mariarosaria and Floridi, Luciano. 2005. ‘Solving the Symbol Grounding Problem: A Critical Review of Fifteen Years of Research.’ Journal of Experimental & Theoretical Artificial Intelligence 17(4).
- The Frame Problem in AI [YouTube video]
- ‘The biggest problem in AI? Machines have no common sense’ [YouTube video]

Background

- Shanahan, Murray, ‘The Frame Problem.’ The Stanford Online Encyclopedia of Philosophy.

Week 5: What does "artificial" mean?

Essential

- Bechtel, William, Bich, Leonardo. 2024. "Eating and Cognition in Two Animals without Neurons: Sponges and Trichoplax." *Biological Theory* 10.1007/s13752-024-00464-6. https://www-scopus-com.libproxy.york.ac.uk/record/display.uri?eid=2-s2.0-85196100327.

Recommended

- Leonardo Bich, Alvaro Moreno. 2016. "The role of regulation in the origin and synthetic modelling of minimal cognition." *Biosystems*, Volume 148, 2016, Pages 12-21. https://www.sciencedirect.com/science/article/pii/S0303264715001148. https://doi.org/10.1016/j.biosystems.2015.08.002.
- Maley, Corey J. 2023. "Analogue Computation and Representation." *The British Journal for the Philosophy of Science* Volume 74, Number 3. https://www-journals-uchicago-edu.libproxy.york.ac.uk/doi/full/10.1086/715031.
- Harraway, Donna. 1991. "A Cyborg Manifesto: Science, Technology, and Socialist Feminism in the Late Twentieth Century," in *Simians, Cyborgs and Women: The Reinvention of Nature.* New York; Routledge, pp.149-181.
- Hu, Charlotte. 2023. "Inside the lab that's growing mushroom computers" *Popular Science*. https://www.popsci.com/technology/unconventional-computing-lab-mushroom/.
- 'The Mushroom Motherboard: The Crazy Fungal Computers that Might Change Everything' [YouTube video]
- 'Can You Upload Your Mind & Live Forever?' [YouTube video]
- 'Scientists Put the Brain of a Worm into a Robot…and It MOVED' [YouTube video]
- 'These Self-Aware Robots Are Redefining Consciousness' [YouTube video]
- "Platt, Charles. 2023. "The Unbelievable Zombie Comeback of Analog Computing." *Wired.* https://www.wired.com/story/unbelievable-zombie-comeback-analog-computing/.

Background

- Van Gulick, Robert. 2014. 'Consciousness.' *Stanford Online Encyclopedia of Philosophy*.

Week 6: AI ethics

Essential

- Bryson, J. 2010. 'Robots Should Be Slaves.' In Close Engagements with Artificial Companions, Y. Wilks (ed.), pp 63-74.

- Schwitzgebel and Mara, 2015. ‘A Defense of the Rights of Artificial Intelligence.’ *Midwest Studies in Philosophy* XXXIX

Recommended

- Five principles for AI ethics: https://hdsr.mitpress.mit.edu/pub/l0jsh9d1/release/8
- Allen, C., Smit, I. & Wallach, W. 2005. "Artificial Morality: Top-down, Bottom-up, and Hybrid Approaches." *Ethics Inf Technol* **7**, 149–155. https://doi.org/10.1007/s10676-006-0004-4
- Bostrom and Yudkowsky, 2014. ‘The Ethics of Artificial Intelligence’ In *Cambridge Handbook of Artificial Intelligence*, edited by Keith Frankish and William Ramsey. New York: Cambridge University Press.
- Etzioni, A. and Etzioni, O. 2017. ‘Incorporating Ethics into Artificial Intelligence.’ *The Journal of Ethics*.
- Thoma, J. 2021. ‘How should artificial agents make risky choices on our behalf?‘
- van Wynsberghe, A., Robbins, S. 2019. ‘Critiquing the Reasons for Making Artificial Moral Agents.’ *Sci Eng Ethics* 25, 719–735.
- Hao, Karen, and Stray, Jonathan. 2019. ‘Can You Make AI Fairer Than a Judge?’ *MIT Technology Review*.
- Müller, Vincent C. 2023. "Ethics of Artificial Intelligence and Robotics", *The Stanford Encyclopedia of Philosophy*, Edward N. Zalta & Uri Nodelman (eds.), URL = <https://plato.stanford.edu/archives/fall2023/entries/ethics-ai/>.

Week 7: AI, art, and creativity

Essential

- Langland-Hassan, Peter. 2024. 'Imagination, Creativity, and Artificial Intelligence' In Amy Kind & Julia Langkau (eds.), Oxford Handbook of Philosophy of Imagination and Creativity. Oxford University Press. https://philarchive.org/rec/LANICA-5.

Recommended

- Coeckelbergh, M. 2017. ‘Can Machines Create Art?‘ *Philos. Technol*. 30, 285–303.
- Brainard, Lindsay. 2024. 'What is creativity?' *The Philosophical Quarterly*, pqae075, https://doi.org/10.1093/pq/pqae075.
- Mikalonytė, Elzė Sigutė & Kneer, Markus. 2022. ‘Can Artificial Intelligence Make Art?‘ *ACM Transactions on Human-Robot Interaction* Volume 11(4) Article No. 43, pp. 1–19.

- Neef, N.E., Zabel, S., Papoli, M. et al. Drawing the full picture on diverging findings: adjusting the view on the perception of art created by artificial intelligence. AI & Soc (2024). https://doi-org.libproxy.york.ac.uk/10.1007/s00146-024-02020-z
- Halina, Marta. 2021. "Insightful Artificial Intelligence." *Mind & Language* 36: 315-329. DOI:10.1111/mila.12321.
- 'How Does A.I. Art Stack Up Against Human Art?' [YouTube video]
- McFadden, Christopher. 2019. '7 of the Most Important AI Artists That Are Defining the Genre.' *Interesting Engineering*.
- Kelly, Sean Dorrance. 2019. 'A philosopher argues that an AI can't be an artist.' *MIT Technology Review*.
- Delacroix, Sylvie. 2021. 'Computing Machinery, Surprise and Originality.' *Philosophy & technology*.
- Anna Ridler: 'Myriad Tulips', 'Bloemenveiling', 'Mosaic Virus' and 'Laws of Ordered Form'. http://annaridler.com/works
- Moruzzi, Caterina. 2022. "Perceptions of Creativity in Artistic and Scientific Processes." xCoAx: 10th Conference on Computation, Communication, Aesthetics & X. DOI 10.24840/xCoAx_2022_5.

Background

- Adajian, Thomas. 2018. 'The Definition of Art.' *Stanford Online Encyclopedia of Philosophy*.

Week 8: AI, war, and responsibility

Essential

- Sparrow, R. 2007. 'Killer Robots.' *Journal of Applied Philosophy*, Vol. 24, No. 1.
- Horowitz, M. 2016. 'The Ethics & Morality of Robotic Warfare: Assessing the Debate over Autonomous Weapons.' *Daedalus* 145 (4): 25–36.

Recommended

- Musgrave, Z. and Roberts, B. 2015. 'Humans, Not Robots, Are the Real Reason Artificial Intelligence Is Scary' *The Atlantic*.
- 'A.I. Is Making it Easier to Kill (You). Here's How' [YouTube video]
- 'The future of modern warfare: How technology is transforming conflict | DW Analysis' [YouTube video]

Week 9: AI, science, and epistemology

Essential

- Andrews, Mel. 2023. "The Devil in the Data: Machine Learning & the Theory-Free Ideal." Preprint. https://philsci-archive.pitt.edu/22690/1/ML_Atheoreticity.pdf

Recommended

- Tamir, M., Shech, E. 2023. "Machine understanding and deep learning representation." *Synthese* 201, 51 (2023).
- Sullivan, Emily. 2022. "Understanding from Machine Learning models." The British Journal for the Philosophy of Science 73(1).
- Yildirim, Ilker, and Paul, Laurie. 2024. "From task structures to world models: what do LLMs know?" *Trends in Cognitive Sciences*, Volume 28, Issue 5, 404 - 415. https://www.cell.com/trends/cognitive-sciences/abstract/S1364-6613(24)00035-4.
- King, R. et al. 2009. "The Automation of Science." *Science* 324(5923): 85-89.
- Krenn, M. et al. 2022. "On scientific understanding with artificial intelligence." *arXiv*.
- Vamathevan, Jessica, et al. 2019. "Applications of machine learning in drug discovery and development." Nature Reviews: Drug Discovery 18: 463-477.
- Stuart, Michael T. 2022. "The future won't be pretty: The nature and value of ugly, AI-designed experiments." In Milena Ivanova and Alice Murphy (eds). *The Aesthetics of Scientific Experiments*. Routledge.https://doi.org/10.1093/bjps/axz035
- Nyrup, Rune. 2022. "Explanatory Pragmatism: A Context-Sensitive Framework for Explainable Medical AI." *Ethics and Information Technology* 24(13).

Week 10: Philosophical writing

Recommended

- Guidelines on Writing a Philosophy Paper (James Pryor)
- Writing a Philosophy Paper (Peter Horban)
- Guide to the Study of Philosophy (Garth Kemerling)
- Tips on Writing a Philosophy Paper (Douglas Portmore)
- 7 Steps to a Better Philosophy Paper (Bryan Roberts)
- A Guide to Philosophical Writing (Elijah Chudnoff)

Week 11: Review, essay preparation

- No assigned readings
- In-person help session during regular class hours

Assessments

Task	Length	% of module mark
Summative assessment: One **ESSAY**	MAX **4000** words. Due date TBA.	100%
Formative assessment: **What's philosophical about that?**	Share one interesting AI news item, case study, academic paper, or simply a precise question about a recent development in AI, for group discussion (happens twice per student over the course of the semester)	0%
Formative assessment: **Class presentation**	10 minutes, once	0%

The **summative assessment**, weighted at 100% of the module mark, is a 4,000-word essay.

- Students will be required to identify their own research question for this essay, in consultation with the instructor.
- Students are strongly recommended to agree on their essay question/topic/thesis with the instructor before the end of Week 11 (mike.stuart@york.ac.uk).
- We will discuss the requirements for this essay in more detail during the second half of semester.

The **formative assessments** are there to develop crucial philosophical skills. There are two.

- First, there is **What's philosophical about that?** For this, each student will be asked to bring in some kind of news item which is not explicitly of a philosophical nature (e.g., a development in AI, law, politics, ethics, or something technology-related), and we will consider together which philosophical issues it raises. The class will be divided into two groups for discussion.
- For accessibility reasons, please post your news item/question on the dedicated discussion board on the course VLE before the class begins (under assessments).

- Second, there is a **class presentation**. Each student is asked to access the sign-up sheet and choose an essential reading. If there are no essential readings left, you can choose a recommended reading. The presentation should be 10 minutes (or less!), and it should introduce the main thesis of the reading and the argument for that thesis, as well as some critical remarks and questions for discussion.

Exceptional circumstances affecting assessment

Sometimes things happen that can seriously impair your performance in an assessment or prevent you undertaking the assessment at the scheduled time. If these events are unforeseeable and exceptional (i.e., serious and unusual) you may be able to defer an assessment or take it again.

Find out more about claiming extenuating circumstances.

Where to get help with assessments

Questions about the assessment task or where/how to submit: contact module staff

Technical help with the VLE: email vle-support@york.ac.uk

Useful resources

Harvard referencing style guide

MLA referencing style guide

Academic Writing Practical Guide (especially the Assessment & feedback page)

Book a 1:1 appointment to discuss your work with the Writing Centre