Topics in the Philosophy of Artificial Intelligence

Taught at the University of York, 2025-2026

Michael T. Stuart

(Feedback always welcome!)

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 of computing, and the history of AI

Essential

- Norvig, Peter and Russell, Stuart. 2022. "Introduction." *Artificial Intelligence: A Modern Approach*. Prentice. Pp. 19-53.

Recommended

- On Turing machines [YouTube video]
- On 'virtual machines' [YouTube video]
- Simulations of neural nets [YouTube video]
- Eater, Ben. "Build an 8-bit computer from scratch." Website and YouTube playlist: https://eater.net/8bit.
- Colburn, Timothy R. 2000. "AI and Logic." and "Computer Science and Mathematics" from *Philosophy and Computer Science*. M.E. Sharpe.

Background

- Dasgupta, Subrata. 2016. *Computer Science: A Very Short Introduction*. Oxford: OUP.
- Boden, Margaret. 2016. "Chapter 1." *AI: Its Nature and future*. Oxford University Press.
- Gardiner, Martin. 1962. "Mathematical Games," *Scientific American*. Available here.

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.

Recommended

- On neural networks [YouTube video]
- Gradient descent [YouTube video]
- Backpropagation [YouTube video]
- Generated Adversarial Networks (GANs) [YouTube video]
- Genetic Algorithms [YouTube video]
- Attention in transformers [YouTube video]
- "Convolutional Neural Networks: A Comprehensive Guide." *Medium*. https://medium.com/thedeephub/convolutional-neural-networks-a-comprehensive-guide-5cc0b5eae175.
- 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 2025 (at least the chapter summaries, pages 11-19)

Background

- Porter, Zoë, et al. 2025. "INSYTE: A Classification Framework for Traditional to Agentic AI Systems." *ACM Transactions on Autonomous and Adaptive Systems* 20(3): 1-39.
- '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: The nature of intelligence: Artificial and natural

Essential

- Lake BM, Ullman TD, Tenenbaum JB, Gershman SJ. 2017. "Building machines that learn and think like people." *Behavioral and Brain Sciences* 40:e253. doi:10.1017/S0140525X16001837.
- Leonardo Bich, Alvaro Moreno. 2016. "The role of regulation in the origin and synthetic modelling of minimal cognition." *Biosystems*, Volume 148: 12-21.

Recommended

- Botvinick M, Barrett DGT, Battaglia P, et al. 2017. "Building machines that learn and think for themselves." *Behavioral and Brain Sciences*. 40:e255. doi:10.1017/S0140525X17000048.
- Dennett, D. 1984. 'Cognitive wheels: The frame problem of AI' In Christopher Hookway (ed.), Minds, Machines and Evolution. Cambridge University Press.
- Bostrom, N. 2012. 'The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents.' *Minds and Machines* 22: 71–85.
- Yildirim, Ilker et al. 2024. "From task structures to world models: what do LLMs know?" *Trends in Cognitive Sciences*, Volume 28, Issue 5, 404-415.
- 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.
- 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).
- Felin, Teppo and Holweg, Matthias. 2024. "Theory Is All You Need: AI, Human Cognition, and Causal Reasoning." Available at SSRN: or https://ssrn.com/abstract=4737265 or http://dx.doi.org/10.2139/ssrn.4737265.
- Searle, J. R. 1980. 'Minds, Brains, and Programs.' *Behavioral and Brain Sciences* 3 (3): 417-57.

- Geraci, Robert M. 2010. "Spiritual Robots: Religion and Our Scientific View of the World." *Theology and Science* 8(3): 229-246.
- 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.
- 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*.
- '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/.
- 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.
- Van Gulick, Robert. 2014. 'Consciousness.' *Stanford Online Encyclopedia of Philosophy*.

Week 5: AI Ethics

Essential

- 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.

Recommended

- 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
- Danaher, John. 2017. "The Symbolic-Consequences Argument in the Sex Robot Debate." In *Robot Sex: Social and Ethical Implications*, edited by John Danaher and Neil McArthur. MIT Press.
- Five principles for AI ethics: https://hdsr.mitpress.mit.edu/pub/l0jsh9d1/release/8
- AI Incident Database. https://incidentdatabase.ai/.
- Ober, Josiah and Tasioulas, John. 2024. "AI Ethics with Aristotle" - White Paper, Lyceum Project. Available here.
- 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*.
- van Wynsberghe, A. 2021. "Sustainable AI: AI for sustainability and the sustainability of AI." *AI and Ethics* 1, 213–218.

Background

- Müller, Vincent C. 2023. "Ethics of Artificial Intelligence and Robotics", *The Stanford Encyclopedia of Philosophy*.
- 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.

Week 6: AI and power: Governance, labour, and the self

Essential

- Acemoglu, Daron, and Johnson, Simon. 2023. "The Wrong Kind of AI? Artificial Intelligence and the Future of Labor Demand." *Oxford Review of Economic Policy*, 39(4), 567–585.
- Danaher, John. 2016. "The Threat of Algocracy: Reality, Resistance and Accommodation." *Philosophy & Technology* 29(3): 245–68.

Recommended

- Zuboff, Shoshana. 2015. "Big Other: Surveillance Capitalism and the Prospects of an Information Civilization." *Journal of Information Technology* 30(1): 75–89.
- Turkle, Sherry. 2011. "Growing up tethered." in *Alone Together: Why We Expect More from Technology and Less from Each Other*. Basic Books.
- Verbeek, Peter-Paul. 2008. "Materializing Morality: Design Ethics and Technological Mediation." *Science, Technology, and Human Values* 31 (3):361-380.
- Paola Ricaurte. 2022. "Artificial Intelligence and the Feminist Decolonial Imagination." *Bot Populi*.
- Fiske, A., P. Henningsen, and E. Buyx. 2019. "Your Robot Therapist Will See You Now: Ethical Implications of Embodied Artificial Intelligence in Psychiatry, Psychology, and Psychotherapy." *Journal of Medical Internet Research* 21 (5): e13216.
- Sunstein, Cass R. 2018. "The Daily Me." In *#Republic: Divided Democracy in the Age of Social Media*. Princeton University Press.
- Kudina, O., Ballsun-Stanton, B. & Alfano, M. 2025. "The use of large language models as scaffolds for proleptic reasoning." *Asian J. Philos.* **4**, 24.
- Luckin, R., Holmes, W., Griffiths, M., & Forcier, L. B. 2016. Intelligence Unleashed: An argument for AI in Education. Pearson.
- Srnicek, Nick. 2017. *Platform Capitalism*. Polity.
- Rouvroy, Antoinette and Berns, Thomas. 2013. "Algorithmic governmentality and prospects of emancipation: Disparateness as a precondition for individuation through relationships?" Translated by Liz Carey-Libbrecht. Original: "Gouvernementalité algorithmique et perspectives d'émancipation" in *Réseaux* 177(1): 163-196.

Background

- Hao, Karen. 2025. *Empire of AI: Inside the reckless race for total domination*. Allen Lane.

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.

Recommended

- Coeckelbergh, M. 2017. 'Can Machines Create Art?' *Philos. Technol*. 30, 285–303.
- Brainard, Lindsay. 2024. 'What is creativity?' *The Philosophical Quarterly*, 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. 2024. "Drawing the full picture on diverging findings: adjusting the view on the perception of art created by artificial intelligence." *AI & Society*
- 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]
- Colucci, Mariachiara, Vecchi, Alessandra & Bonetti, Francesca. 2025. "The Transformative Influence of Artificial Intelligence on the Fashion Industry." In Giovanni Emanuele Corazza, *The Cyber-Creativity Process: How Humans Co-Create with Artificial Intelligence*. Cham: Springer Nature Switzerland. pp. 163-201.
- 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’.
- 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.
- Yee, Adrian K. 2025. "Construct Validity in Automated Counterterrorism Analysis." Philosophy of Science. 2025;92(3):566-583. doi:10.1017/psa.2024.65

Recommended

- Horowitz, M. 2016. ‘The Ethics & Morality of Robotic Warfare: Assessing the Debate over Autonomous Weapons.’ *Daedalus* 145 (4): 25–36.
- Musgrave, Z. and Roberts, B. 2015. ‘Humans, Not Robots, Are the Real Reason Artificial Intelligence Is Scary‘ *The Atlantic*.
- Paris Marx and Laleh Khalili. 2025. "How the US Weaponizes Tech in the Middle East." *Tech Won't Save You*.
- Paris Marx and Spencer Ackerman. 2025. "Gaza Is a Laboratory for Future Warfare." *Tech Won't Save You*.
- ‘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]
- Brand, J.L.M. 2025. "Air Canada’s chatbot illustrates persistent agency and responsibility gap problems for AI." AI & Soc 40, 3361–3363.

Week 9: AI, science, epistemology, and trust

Essential

- Andrews, Mel. 2023. "The Devil in the Data: Machine Learning & the Theory-Free Ideal." Preprint.
- Nyrup, Rune. 2022. "Explanatory Pragmatism: A Context-Sensitive Framework for Explainable Medical AI." *Ethics and Information Technology* 24(13).

Recommended

- Doshi-Velez, F. and Kim, B. 2017. "Towards a rigorous science of interpretable machine learning." arXiv preprint arXiv:1702.08608.
- Páez, A. 2019. "The Pragmatic Turn in Explainable Artificial Intelligence (XAI)". *Minds & Machines* **29**, 441–459.
- 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.
- Soltan, Andrew A S et al. 2021. "Rapid triage for COVID-19 using routine clinical data for patients attending hospital: development and prospective validation of an artificial intelligence screening test." *The Lancet Digital Health*, Volume 3, Issue 2, e78 - e87.
- 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
- Stuart, Michael T. 2025. "A New Account of Pragmatic Understanding, Applied to the Case of AI-Assisted Science." *Philos Stud*.
- Chen, Huili, Grimm, Stephen, Russakovsky, Olga and Lombrozo, Tania. Preprint. "Machine Understanding."
- Wolfram, Stephen. 2024. "Can AI Solve Science?"