PHILOSOPHY OF ARTIFICIAL INTELLIGENCE

Fall 2022 Syllabus

Instructor: Ignacio Ojea Quintana.
Meeting Time: Tuesdays 12pm to 2pm
Location: Geschw.-Scholl-Pl. 1 (F) - F 007 (Geschossplan)
Office Hours: Tuesdays 10am-12pm
Contact: ignacio.ojea@lrz.uni-muenchen.de
Uni Link: Link

Course Description

In this course we will examine central philosophical issues around Artificial Intelligence. The course will have two parts.

The first part will focus on artificial intelligence and the mind. Some of the guiding questions are:

- How can we establish that a computer is intelligent?
- More broadly, how would a theory of the mental in computational terms would look like?
- What problems do computational theories of the mind have?
- What is the difference between symbolic AI and contemporary approaches?
- What can we say about consciousness and artificial intelligence?

In order to answer these questions we will study the Turing Test, Turing Machines, Computationalism and its challenges, Connectionism, and some issues around consciousness, super-intelligence and the Computer Simulation Hypothesis.

The second part of the course will focus on moral and social issues around AI. Some of the guiding questions are:

- What kind of risks does AI pose for humanity?
- What is the moral status of AI?
- How does AI affect fairness relations in society? Can AI be fair?
- AI and Machine Learning applications can be opaque, in the sense that when thing go wrong it is hard to say what failed. How can we improve on AI transparency?
- What role does AI have in popular culture?

We will examine all of these questions by looking at journal articles and chapter selections

Prerequisites

The course will have both a technical component and a philosophy component. Mathematics without philosophy is empty, philosophy without mathematics is blind. Students are expected to understand the basics of logic as well as having the basic ability to write a philosophical essay. Having taken a introductory logic course *and* an introductory philosophy course should suffice.

Required Texts

All texts will be made available digitally.
But you should definitely own a printed version of Descartes *Meditations on First Philosophy.*

Grading

The course will have two evaluation instances near the end of the semester.
First, a quiz/exam about the content of the class. A list of questions will be provided in sufficient advance, and the exam will consist in a subset of those questions. It is worth 30% of the total grade.
Second, an essay worth 70% of the grade. You can find a guide on how to write the essay here.

TENTATIVE COURSE SCHEDULE

Please note that all readings and due dates are subject to change.
Please do the readings before attending to class.
Note: Dates will be corrected according to the academic calendar. This course is about twelve weeks long, factoring in public holidays.

Part 1: Artificial Minds

Week 1: *The Turing Test* **Tu 18/10**

Required:

- A. Turing, "Computing Machinery and Intelligence", [25].

Optional:

- J.H. Moor, "An Analysis of the Turing Test", *Philosophical Studies*, [17].
- M. Halina, "Insightful artificial intelligence", [13].

Week 2: *Turing Machines: Universal Turing Machines, Halting Problem, Church-Turing Tesis.* **Tu 25/10**

Required:

- G.S. Boolos, J.P. Burgess, R.C. Jeffrey - *Computability and Logic*, **Chapter 3.**, [4].
- Check this out: https://turingmachine.io/

Optional:

- G.S. Boolos, J.P. Burgess, R.C. Jeffrey - *Computability and Logic*, **Chapter 4.**, [4].
- G.S. Boolos, J.P. Burgess, R.C. Jeffrey - *Computability and Logic*, **Chapter 8.**, [4].

Week 3: *Computationalism* **Tu 1/11**

Required:

- H. Putnam, "The Nature of Mental States", [21].

Optional:

- M. Rescorla, *The Computational Theory of Mind*, Stanford Encyclopedia of Philosophy. (link)
- J. Fodor, "Propositional Attitudes", [11].

Week 4: *Challenges to Computationalism* **Tu 8/11**

Required:

- J. Searle, "Can Computers Think?", [23].

Optional:

- M. Boden, "Escaping from the Chinese Room", [3].
- D. Dennett, "Can Machines Think?", [10].

Week 5: *Connectionism and Artificial Neural Networks* **Tu 15/11**

Required:

- P. Smith Churchland, excerpts from Chapter 7 of *Brain-Wise*, "How do Brains Represent?", [8].

Optional:

- Notes on Neural Networks (to be uploaded)
- A. Clark, "Connectionism, Competence, and Explanation", [9].
- D. E. Rumelhart, "The architecture of mind: a connectionist approach", [22].

Week 6: *Consciousness and AI* **Tu 22/11**

Required:

- R. Descartes, *Meditations* 1st & first half of 2nd. Any translation is good, I suggest you get this book if you do not have it.
- D. Chalmers, "Facing up the problem of consciousness", opinion article in *Scientific American*, 1995.

Optional:

- T. Nagel, "What is it like to be a bat", [19].
- F. Jackson, "Epiphenomenal Qualia", [15].

Week 7: *Superintelligence and The Computer Simulation Hypothesis* **Tu 29/11**

Required:

- D. Chalmers, "The Singularity: A Philosophical Analysis", [7].
- N. Bostrom, "Are We Living in a Computer Simulation?", [5].

Optional:

- J. Prinz. "Singularity and Inevitable Doom", [20].

Part 2: Moral and Social Issues Around AI

Week 8: *Existential Risk* **Tu 6/12**

Required:

- V. Müller & M. Cannon, "Existential risk from AI and orthogonality: Can we have it both ways?", [18].

Optional:

- N. Bostrom, "The Superintelligent Will: Motivation and Instrumental Rationality in Advanced Artificial Agents", [6].
- S. Armstrong , "General Purpose Intelligence: Arguing the Orthogonality Thesis", [2].

Week 9: *The Moral Status of AI* **Tu 13/12**

Required:

- D.G. Johnson (2006), "Computer systems: Moral Entities But Not Moral Agents", [16].

Optional:

- J.P. Sullins, "When is a Robot a Moral Agent?", [24].
- M.A. Warren. *Moral Status: Obligations to Persons and Other Living Things*, Excerpt pp. 4–17, [26].

Week 10: *AI Fairness* **Tu 20/12**

Required:

- B. Hedden, "On statistical criteria of algorithmic fairness", [14].

Week 11: *Transparency and Data Bias* **Tu 10/1**

Required:

- M. Günther & A. Kasirzadeh, "Algorithmic and human decision making: for a double standard of transparency", [12].

Optional:

- Zerilli et al. "Transparency in algorithmic and human decision-making: Is there a double standard?" [27].

Week 12: *Artificial Intelligence in Pop Culture* (Terminator, 2001 Space Odyssey, *Her*, *Ex Machina*. Also *Asimov*, *The Matrix*). [For now we use Asimov but we might change that] **Tu 17/1**

Required:

- Isaac Asimov, *The Bicentennial Man. In The Bicentennial Man and Other Stories*, Doubleday, pp. 138–172, 1976.

- S.L. Anderson, "The Unacceptability of Asimov's Three Laws of Robotics as a Basis for Machine Ethics", [1].

Examination Week:

Final Quiz Due
Final Paper Due

References

[1] Susan Leigh Anderson. *The Unacceptability of Asimov's Three Laws of Robotics as a Basis for Machine Ethics*, page 285–296. Cambridge University Press, 2011.

[2] Stuart Armstrong. General purpose intelligence: Arguing the orthogonality thesis. *Analysis and Metaphysics*, 12:68–84, 01 2013.

[3] Margaret A. Boden. Escaping from the chinese room. In John Heil, editor, *Computer Models of Mind*. Cambridge University Press, 1988.

[4] George S. Boolos, John P. Burgess, and Richard C. Jeffrey. *Computability and Logic*. Cambridge University Press, 5 edition, 2007.

[5] By Nick Bostrom. Are we living in a computer simulation? *Philosophical Quarterly*, 53(211):243–255, 2003.

[6] Nick Bostrom. The superintelligent will: Motivation and instrumental rationality in advanced artificial agents. *Minds and Machines*, 22(2):71–85, 2012.

[7] David J. Chalmers. The singularity: A philosophical analysis. *Journal of Consciousness Studies*, 17(9-10):9–10, 2010.

[8] Patricia Smith Churchland. *Brain-Wise: Studies in Neurophilosophy*. MIT Press, 2002.

[9] Andy Clark. Connectionism, competence, and explanation. *The British Journal for the Philosophy of Science*, 41(2):195–222, 1990.

[10] Daniel C. Dennett. *Can Machines Think?*, pages 295–316. Springer Berlin Heidelberg, Berlin, Heidelberg, 2004.

[11] Jerry A. Fodor. Propositional attitudes. *The Monist*, 61(October):501–23, 1978.

[12] Mario Günther and Atoosa Kasirzadeh. Algorithmic and human decision making: for a double standard of transparency. *AI & SOCIETY*, 37, 03 2022.

[13] Marta Halina. Insightful artificial intelligence. *Mind & Language*, 36, 01 2021.

[14] Brian Hedden. On statistical criteria of algorithmic fairness. *Philosophy and Public Affairs*, 49(2):209–231, 2021.

[15] Frank Jackson. Epiphenomenal qualia. *The Philosophical Quarterly (1950-)*, 32(127):127–136, 1982.

[16] Deborah G. Johnson. Computer systems: Moral entities but not moral agents. *Ethics and Information Technology*, 8(4):195–204, 2006.

[17] James H. Moor. An analysis of the turing test. *Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition*, 30(4):249–257, 1976.

[18] Vincent C. Müller and Michael Cannon. Existential risk from ai and orthogonality: Can we have it both ways? *Ratio*, 35(1):25–36, 2022.

[19] Thomas Nagel. What is it like to be a bat? *The Philosophical Review*, 83(4):435–450, 1974.

[20] Jesse Prinz. Singularity and inevitable doom. *Journal of Consciousness Studies*, 19(7-8):77–86, 2012.

[21] Hilary Putnam. *17. The Nature of Mental States*, pages 223–231. Harvard University Press, Cambridge, MA and London, England, 2013.

[22] David E. Rumelhart. The Architecture of Mind: A Connectionist Approach. In *Mind Design II: Philosophy, Psychology, and Artificial Intelligence*. The MIT Press, 03 1997.

[23] John R. Searle. Can computers think? In David J. Chalmers, editor, *Philosophy of Mind: Classical and Contemporary Readings*. Oup Usa, 2002.

[24] John P. Sullins. When is a robot a moral agent. *International Review of Information Ethics*, 6(12):23–30, 2006.

[25] A. M. Turing. Computing machinery and intelligence. *Mind*, 59(236):433–460, 1950.

[26] Mary Anne Warren. *Moral Status: Obligations to Persons and Other Living Things*. Oxford University Press, 03 2000.

[27] John Zerilli, Alistair Knott, James Maclaurin, and Colin Gavaghan. Transparency in algorithmic and human decision-making: Is there a double standard? *Philosophy & Technology*, 32, 12 2019.