University of Bern
HS 2024
Claus Beisbart, Pierre Beckmann

CAS AML M5 Philosophy and Ethics of Extended Cognition and Artificial Intelligence

Conception and schedule (last update: 11.10.2024)

1. Objectives

The aim of this seminar is to discuss Extended Cognition and Artificial Intelligence (AI) from a broader perspective. AI comes with a lot of "food for thought": It raises deep philosophical issues and will at the same time have a significant impact on our society such that ethical challenges arise. Philosophical papers will make us aware of the key issues and contain proposals for solving problems connected to AI.

Topics:

(a) Philosophical conceptions of AI (weak vs. strong AI)

(b) Extended mind hypothesis

(c) Philosophical concepts of data and data analysis

(d) AI and scientific inference

(e) Ethical challenges due to AI in the light of ethical theories

(f) Machine ethics

Students who have successfully participated at the course

(a) have an overview of the history and the philosophy of artificial intelligence,

(b) know philosophical and scientific presuppositions of artificial intelligence,

(c) can relate techniques of AI to well-known scientific methods,

(d) know the main philosophical discussions on artificial intelligence,

(e) know the main moral challenges related to artificial intelligence and can discuss solutions from the perspective of ethics,

(f) master best practices for ethics dealing with artificial intelligence.

2. Way of proceeding

In the seminar sessions, we discuss philosophical papers about AI. The focus is on a few famous classics and more recent contributions. Please prepare each sessions by carefully studying the paper that has been announced and sent around beforehand. Some sessions will be started by a brief general introduction by the teacher. Most sessions are centered about a presentation of the paper by two participants.

3. Possible topics and with suggestions for readings

NB. One reading will be selected per session; the following suggestions include additional reading that are not required.

Theoretical Issues

(1) Can machines think? The Turing test

Text: Turing, A. (1950), Computing Machinery and Intelligence, Mind, LIX: 433–460; additional literature: Besold, Tarek R. (2013), Turing revisited: A cognitively-inspired decomposition, in: Müller, Vincent C. (ed., 2013), Philosophy and theory of artificial intelligence, Berlin and Heidelberg: Springer, 121–132

(2) Can machines understand language? The Chinese room argument

Text: Searle, John R. (1980), Minds, brains, and programs, Behavioral and Brain Sciences 3 (3), 417–424; additional texts: Responses to Searle, John R. (1980), Behavioral and Brain Sciences 3 (3), 424–457; Churchland, Paul M. & Churchland, Patricia S. (1990), Could a machine think? Scientific American 262 (1), 32–37

(3) Are there things that computers cannot do?

Text: Dreyfus, Hubert L. (1999), What computers "still" can't do: A critique of artificial reason, Cambridge (Mass.): MIT Press (extracts, particularly from Chs. 7–9, pp. 235–280)

(4) How does ML affect scientific method and data analysis?

Pietsch, Wolfgang (2015), Aspects of Theory-Ladenness in Data-Intensive Science, Philosophy of Science 82 (5):905-916; alternative: Sullivan, Emily, Understanding from Machine Learning Models, The British Journal for the Philosophy of Science 73:1 (2022), 109-133; additional text: Leonelli, Sabina (2015). What Counts as Scientific Data? A Relational Framework. Philosophy of Science 82 (5):810–821

(5) Should we expect superintelligence and a singularity?

Text: Chalmers, David J. (2010), The singularity: A philosophical analysis, Journal of Consciousness Studies 17 (9-10), 7–65; alternative: Bostrom, Nick (2014), Superintelligence: Paths, dangers, strategies, Oxford: Oxford University Press (extracts)

(6) Why are many AI algorithms black boxes and what can we do about it?

Fleisher, Will (2022). Understanding, Idealization, and Explainable AI, Episteme 19 (4):534-560; alternatives: Günther, M. & Kasirzadeh, A. (2022), Algorithmic and human decision making: For a double standard of transparency, AI & Soc 37, 375–381; Zednik, Carlos (2019), Solving the Black Box Problem: A Normative Framework for Explainable Artificial Intelligence, Philosophy and Technology 34 (2):265-288; Zerilli, John (2022), Explaining Machine Learning Decisions, Philosophy of Science 89 (1), 1–19.

(7) How far does the mind extend?

Text: Clark, Andy & Chalmers, David J. (1998), The Extended Mind, Analysis 58 (1), 7–19

Ethics of Artificial Intelligence

(1) Are there things that computer should not do?

Texts: Moor, James H. (1979), Are there decisions computers should never make?, Nature and System 1, 217–29; alternative: Grote, Thomas & Berens, Philippe (2020), On the ethics of algorithmic decision-making in healthcare. Journal of Medical Ethics 46 (3):205-211.

(2) Should we implement morality into machines? The idea of machine ethics

Texts: Allen, Colin, Wallach, Wendell & Smit, Iva (2006), Why machine ethics? IEEE Intelligent Systems, 21/4, 12–17, reprinted in Anderson, Michael & Anderson, Susan L. (2011), Machine ethics, New York: Cambridge University Press, 51—61; and: van Wynsberghe, Aimee & Robbins, Scott (2019), Critiquing the reasons for making artificial moral agents, Science and Engineering Ethics 25, 719–735; alternatives: Etzioni, Amitai & Etzioni, Oren (2017), Incorporating ethics into artificial intelligence, The Journal of Ethics 21 (4), 403-418; Anderson, Michael, Anderson, Susan L. & Armen, Chris (2006), An approach to computing ethics, IEEE Intelligent Systems 21/4, 56—63

(3) Which values should be implemented into machines? The issue of value alignment

Text: Gabriel, Iason (2020), Artificial intelligence, values, and alignment. Minds and machines, 30(3), 411-437; alternative: Gabriel, Iason, and Vafa Ghazavi, The Challenge of Value Alignment, in: The Oxford Handbook of Digital Ethics. Oxford: Oxford University Press, 2022.

(4) Should we trust AI?

Texts: Hatherley, Joshua James (2020), Limits of trust in medical AI. Journal of Medical Ethics 46 (7):478-481; Starke, Georg ; van den Brule, Rik; Elger, Bernice Simone & Haselager, Pim (2021), Intentional machines: A defence of trust in medical artificial intelligence. Bioethics 36 (2):154-161.

(5) Biases and algorithmic fairness

Texts: Fazelpour, Sina & Danks, David (2021), Algorithmic bias: Senses, sources, solutions. Philosophy Compass 16 (8):e12760; additional literature: Johnson, Gabbrielle M., Algorithmic bias: on the implicit biases of social technology. Synthese 198 (2021), 9941–9961; Herzog, Lisa, Algorithmic bias and access to opportunities, in: The Oxford Handbook of Digital Ethics. Oxford: Oxford Academic, 2021, 413–432; Hedden, Brian (2021), On statistical criteria of algorithmic fairness. Philosophy and Public Affairs 49 (2):209-231; Binns, Reuben, What Can Political Philosophy Teach Us about Algorithmic Fairness? IEEE Security & Privacy 16/03 (2018), 73–80; Friedman, Batya & Nissenbaum, Helen (1996), Bias in computer systems, ACM Transactions on Information Systems 14, 330—347; Verma, S., & Rubin, J. (2018, May). Fairness definitions explained. In 2018 IEEE/ACM International Workshop on Software Fairness (FairWare) (pp. 1-7). IEEE; Wong, Pak-Hang (2020). Democratizing Algorithmic Fairness. Philosophy and Technology 33 (2):225-244.

(6) How do robots affect human work?

Texts: Smids, Jilles, Nyholm, Sven & Berkers, Hannah (2020), Robots in the workplace: a threat to – or opportunity for – meaningful work? Philosophy and Technology 33 (3), 503–522; Danaher, John (2017). 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.

(7) Should we become friends with robots?

Text: Danaher, John (2019), The Philosophical Case for Robot Friendship. Journal of Posthuman Studies 3 (1): 5–24.

(8) How should we deal with AI assistants and recommender systems?

Texts: Valentine, Lee, D'Alfonso, S. & Lederman, R. (2023), Recommender systems for mental health apps: advantages and ethical challenges, AI & Soc 38, 1627–1638; Milano, Silvia ; Taddeo, Mariarosaria & Floridi, Luciano (2020), Recommender systems and their ethical challenges. AI and Society (4):957-967.

(9) Who is responsible if AI makes a mistake?

Text: Königs, Peter (2022). Artificial intelligence and responsibility gaps: what is the problem? Ethics and Information Technology 24 (3):1-11; Nyholm, Sven (2018), Attributing agency to automated systems: Reflections on human–robot collaborations and responsibility-loci, Science and Engineering Ethics 24 (4), 1201–1219; alternatives: Danaher, John (2016), Robots, Law and the Retribution Gap, Ethics and Information Technology 18 (4):299–309; Müller, Luise. "Domesticating Artificial IntelligenceMoral Philosophy and Politics, vol. 9, no. 2, 2022, pp. 219-237.

(10) Medicine as a use case

Text: Sparrow, Robert & Hatherley, Joshua James (2019), The promise and perils of AI in medicine, International Journal of Chinese and Comparative Philosophy of Medicine 17 (2):79-109; alternative: London, Alex John (2019), Artificial Intelligence and Black-Box Medical Decisions: Accuracy versus Explainability. Hastings Center Report 49 (1):15-21.

4. Schedule (proposal)

1. **18.10.2024** Introduction. Theory 1: Can machines think? The Turing test
2. **25.10.2024** Theory 2: Singularity
3. **01.11.2024** Theory 3: Explainability / scientific method
4. **08.11.2024** Theory 4: Extended Mind
5. **15.11.2024** Ethics 1: Machine decisions
6. **22.11.2024** Ethics 2: Value alignment
7. **29.11.2024** Ethics 3: Responsibility
8. **06.12.2024** Ethics 4: Algorithmic bias
9. **13.12.2024** Ethics 5: Robot friendship

5. Course requirements

You earn 2 credit points if, and only if,

(a) you participate on a regular basis (unless you cannot participate due to health reasons etc.);

(b) you give a short presentation (in a group of two people);

(c) you write a short essay (ca. 1'500 words).

6. Guidelines for presentations

(a) Presentations are done by groups of 2 people.

(b) Task: Please specify the main claims/ideas of the paper and provide 1–2 discussion points. You can't cover everything; so concentrate on what you take to be important. Discussion points are either non-trivial questions of understanding or possible objections, suggestions for further development.

(c) There are two alternative ways of doing the presentation:

- Either you present your slides at the beginning of the course in max. 10 minutes (strict; the presentation will be finished after 12 minutes the latest); afterwards, the teacher takes over and chairs the discussion.
- Alternatively, you take responsibility for the whole session. You can then split your presentation in small parts and start a discussion after each part. If you do, the presentation as such can take a bit longer.

(d) We recommend to formulate the most important 3–5 theses of a paper in your own words. Sometimes it's helpful to draw a little map.

(e) Media: Please come along with a couple of slides or a handout of 1 page (it's strongly recommended that you send your material to the teachers beforehand).

(f) Wording: Please don't copy passages from the paper without quoting with quotation marks; using your own words to feature the content of the paper is preferable.

(g) Option: The presentation is interrupted for the sake of questions of clarification posed by other people.

7. Work load

2 ECTS points correspond to a work load of about 60 h. Here is how this work load might be split:

(a) attendance: $9 \times 1{,}5$ h = 13.5 h;

(b) preparation of sessions (reading of max. 25 pages): 9×3 h = 27 h;

(c) presentation: 6 h;

(d) essay (ca. 1500 words; due Jan 31st, 2025; guidelines will follow): 10–15 h;

Sum: ca. 60 h. The course is not graded.

8. Materials

Electronic versions of the readings will be available on the ILIAS platform for this course: https://ilias.unibe.ch/ilias.php?baseClass=ilrepositorygui&ref_id=2930805.

9. Room and link

B001, ExWi, Sidlerstrasse 5, Link:
https://unibe-ch.zoom.us/j/63704458763?pwd=Q3poVINXFAieaJVVgppUyt55E3sVYg.1

10. Contact

Prof. Dr. Dr. C. Beisbart, Institute of Philosophy, University of Bern, Länggassstr. 49a, CH-3012 Bern, room B223. Telephone: 031 / 684 35 90. Email: Claus.Beisbart@unibe.ch.
Office hours: Tue, 15:00–15:30 or upon agreement (in any case email communication beforehand is recommended).

Pierre Beckmann, Institute of Philospohy, University of Bern, Länggassstr. 49a, CH-3012 Bern. Email: pierre.beckmann@students.unibe.ch