Phil2110: Introduction to Philosophy of Artificial Intelligence

Instructor: Prof. David Thorstad
Office hours: Th 4-5PM, Furman Hall 113
Last syllabus update: January 23, 2024

1. About this course

This course is an introduction to the philosophy of artificial intelligence. We will cover the ethics and nature of artificial intelligence. I hope that by the end of this course you will learn to (1) think philosophically about recent developments in artificial intelligence, (2) understand a range of philosophical problems regarding the nature and ethics of artificial intelligence, and (3) situate topics in the philosophy of artificial intelligence within broader philosophical, academic and societal perspectives.

This is a scaled-down pilot for a course that I hope to offer every year, so your feedback will be instrumental in shaping the future of this course. Please be open and free in your feedback when possible. (Tell me anything: http://tinyurl.com/Phil2110-Feedback).

2. Course materials

All readings are available on the course website. There is no need to purchase any books for this course.

3. Course structure

3.1. Assignments

I **Participation (10%):** Based on attendance and active participation during lectures and other course components.

II **Case study (40%):** Working in groups, students will prepare an ethical case study examining a significant issue in the ethics of artificial intelligence.

III **Case study presentation (20%):** Each group will present their case study as a class exercise.

IV **Final exam (30%):** A final exam will test knowledge of course content and ability to draw on course content to answer philosophical questions raised by artificial intelligence. April 29, 9AM.

4. Schedule

4.1. Introduction

January 9: Course introduction.

4.2. Ethics of artificial intelligence

January 11: Transparency and explanation, introduction.

i **Reading:** Will Knight, "The dark secret at the heart of AI", *MIT Technology Review* (2017).

January 16: Transparency and explanation, case study.

i **Reading:** None.

January 18: Snow day. No class.

January 23: Transparency and explanation, contemporary literature.

i **Reading:** Kate Vredenburgh, "The right to explanation," *Journal of Political Philosophy* (2022).

January 25: No class.

January 30: Bias, introduction.

i **Reading:** Julia Angwn et al., "Machine bias," ProPublica (2016).

February 1: Bias, case study.

i **Reading:** None.

February 6: Bias, contemporary literature.

i **Reading:** Tom Kelly, "The norm-theoretic account of bias," from *Bias: a philosophical study* (2022).

February 8: Privacy, introduction.

i **Reading 1:** Carissa Véliz, "The future of privacy", Philosophy 24/7 podcast

ii **Reading 2:** Glenn Greenwald, "Why privacy matters", TED Conference (2014).

February 13: Privacy, case study.

i **Reading:** None.

February 15: Privacy, contemporary literature.

i **Reading:** Beate Rosessler, "Privacy as a human right," *Proceedings of the Aristotelian Society* (2017).

February 20: Statistical generalizations, introduction.

i **Reading 1:** John Hope Franklin, "Racism in the 1990s," https://www.youtube.com/watch?v=30go2yHooQo.

ii **Reading 2:** Quimbee, "Smith vs. Rapid Transit, Inc., Case brief summary," https://www.youtube.com/watch?v=aWq3uMUSh1I.

iii **Reading 3:** Ian Gordon, "Law and stats: What's a statistician doing in court?," Random Sample Podcast.

February 22: Statistical generalizations, case study.

i **Reading:** None.

February 27: Statistical generalizations: contemporary literature.

i **Reading:** Renée Jorgensen Bolinger, "The rational impermissibility of accepting (some) racial generalizations," *Synthese* (2020)

4.3. Nature of artificial intelligence

February 29: Can machines think? Introduction.

i **Reading:** Sebastien Bubeck, "Sparks of AGI: Early experiments with GPT-4", https://www.youtube.com/watch?v=qbIk7-JPB2c.

March 5: Can machines think? The Chinese Room argument.

i **Reading:** John Searle, "Can computers think?," *Minds, Brains, and Science* (1984).

March 7: Can machines think? Responses to the Chinese Room argument.

i **Reading:** Paul Churchland and Patricia Churchland, "Could a machine think?" *Scientific American* (1990).

March 19: Can machines think? The case of GPT-3.

i **Reading:** Luciano Floridi and Massimo Chiriatti, "GPT-3: Its nature, scope, limits and consequences," *Minds and Machines* (2020)

March 21: Can machines be conscious? Introduction.

i **Reading:** David Chalmers, "Are language models sentient?," https://www.youtube.com/watch?v=-BcuCmf00_Y.

March 26: Can machines be conscious? Structuring the debate.

i **Reading:** Susan Schneider, Ch2 of *Artificial you* (2019).

March 28: Can machines be conscious? Timelines and ethical implications.

i **Reading:** Jeff Sebo and Rob Long, "Moral consideration for AI systems by 2030," *AI and Ethics* (2023), SECTION 3 ONLY.

April 2: Case study presentations

i **Reading:** None.

April 4: No class.

April 9: Case study presentations

i **Reading:** None.

April 11: Superintelligence: Introduction.

i **Reading 1:** Nick Bostrom, "What happens when our computers get smarter than we are?," https://www.youtube.com/watch?v=MnT1xgZgkpk.

ii **Reading 2:** Ray Kurzweil, "The coming singularity," https://www.youtube.com/watch?v=1uIzS1uCOcE.

April 16: Superintelligence: Implications.

i **Reading:** Nick Bostrom, "Existential risk prevention as global priority," *Global Policy* (2013), Sections 1-2 only.

April 18: Final exam review.

i **Reading:** None.

5. Course policies

5.1. Office hours

I will hold office hours each Thursday from 5-6PM in Furman Hall 113. I would encourage you to stop by!

5.2. Technology policy

(The following policy is loosely modified from a policy by Prof. Michael Bess).

Recent technological developments have transformed the way that students learn. My goal in this course is to enable you to make appropriate use of AI tools as a learning aid, while submitting work that is entirely your own.

For the purposes of this course, the use of AI tools such as GPT4, Bing, Claude or Bard falls under two categories:

I Text-generation tool (prohibited).

II Research, brainstorming and editorial aid (permitted).

Prohibited uses include:

I **Entire AI-generated assignment:** A student instructs an AI text-generation tool to write an entire essay or assignment, then hands in the assignment as if it had been written by the student.

II **Partial AI-generated assignment:** A student instructs an AI text-generation tool to write a portion of an essay or assignment, then hands in the assignment as if it had been entirely written by the student.

III **Modified AI-generated assignment:** A student extensively modifies an AI-generated text in ways that result in a hybrid of the student's own phrasing intermingled with AI-generated text, then hands in the assignment as if it had been entirely written by the student.

IV **Paraphrased AI-generated assignment:** A student completely paraphrases an AI-generated essay or assignment in ways that result in a new text that is entirely written by the student, but that is merely a thorough rewording of a text generated by the AI. This paraphrased text still follows the same overall structure and organization as the AI-generated text, and closely echoes the main ideas presented in the AI-generated text. The student then hands in the assignment as if it had been entirely written by the student.

Permitted uses, *so long as the use of AI tools is acknowledged in writing*, include:

I **AI tools used for background research:** A student consults an AI tool by asking it basic factual or thematic questions about a topic, then seeing what kinds of material the AI presents in response. This is similar to consulting Wikipedia at the outset of a project, in order to get a quick sense of the main factual and thematic contours of the subject matter.

II **AI tools used for brainstorming:** A student consults an AI tool with questions about basic concepts, ideas, principles, theories, or scholarly debates relating to a topic the student wishes to explore. This is similar to consulting scholarly articles online, in order to get a sense of the main conceptual or theoretical contours of the subject matter.

III **AI-generated essay or text is consulted before student writes their own essay:** A student prompts an AI text-generation tool to write an essay on a topic assigned for this course, but only uses the AI-generated text for 'consultative' purposes, in order to see what kinds of ideas or arguments the AI has come up with. After reading the AI-generated essay, and reflecting critically about it, the student then conducts their own research and reflection about the topic, using the kinds of scholarly tools available to humans before the advent of advanced AI (for example, books, journal articles, online sources, debates with classmates, conversations with professors). The AI-generated essay thus becomes merely one element within the broad array of other resources that the student consults, and critically reflects upon, in researching, debating and crafting their own final product. (Note: with this usage, students need to be careful not to merely paraphrase portions of the AI-generated text or closely echo its organizational structure. The final product needs to be the student's own work of critical reflection and synthesis.)

IV **AI-edited version of student's essay or assignment:** A student composes their own entirely original essay or assignment, then submits the assignment to an AI tool to see what kinds of stylistic edits and/or grammatical modifications the AI recommends.

5.3. Accessibility

This class respects and welcomes students of all backgrounds, identities, and abilities. If there are circumstances that make your learning environment and activities difficult, if you have medical information that you need to share with me, or if you need specific arrangements in case the building needs to be evacuated, please let me know. I am committed to creating an effective learning environment for all students, but I can only do so if you discuss your needs with me as early as possible. I promise to maintain the confidentiality of these discussions. If appropriate, also contact Student Access Services to get more information about specific accommodations.

5.4. Grade disputes

Students have the right to know why they have received the grades that they have been given, and to seek redress if necessary. If you are unsure why you have received a given grade, please follow exactly the procedure below:

I **Wait 24 hours:** Please wait a minimum of 24 hours after grades are assigned before contacting me to discuss grades. This wait period is often helpful for processing feedback.

II **Submit written request for clarification:** Write to me specifying the portions of the assignment and its grading that you would like to discuss. Please submit requests in writing to guide future conversations in a clear direction.

III **(Optional) Request re-grading:** If you are still unsatisfied with your grade, please send a written request to me to have your assignment re-graded, and include a specification of any points in the initial grading that you are unhappy with.

i **Re-grading:** If there are satisfactory grounds for re-grading, I will fully re-grade the paper, making a holistic assessment of its merits in light of our discussion.

5.5. Academic integrity

All classes at Vanderbilt are governed by the undergraduate honor policy. The library has a helpful guide to avoiding plagarism (https://researchguides.library.vanderbilt.edu/plagiarism).

I recognize that students sometimes find themselves in difficult situations with too many deadlines to meet at once. If this happens to you, I would warmly encourage you to speak to your teaching assistant, or to myself. Often it is possible to arrange an extension.

I take academic dishonesty very seriously. Please don't cheat in my class.