Phil3891: Ethics 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 a selective survey of the ethics of artificial intelligence. We'll begin each unit with an ethical case study, then read perspectives from the philosophical and scientific literature. I hope that by the end of this course you will learn to (1) think philosophically about ethical implications of artificial intelligence, (2) understand a range of philosophical problems regarding the ethics of artificial intelligence, (3) situate topics in the philosophy of artificial intelligence within broader philosophical, academic and societal perspective, (4) communicate ethical ideas orally and in writing in a clear and argument-driven way.

This course is highly discussion-based. Please feel free to participate actively and help to make this course your own. This is a pilot of a course that I hope to offer regularly, so please be open and honest with your feeedback. Your feedback will help shape future iterations of the course. (Tell me anything: http://tinyurl.com/Phil3891-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 (15%):** Based on attendance and active participation during all components of the course. Some ways to participate actively include (but are not limited to) attending lecture, keeping up with readings, contributing to discussions, contributing to group work, attending office hours, and giving course feedback. I understand that different students participate in different ways. In particular, participation is not (purely) a quantity game: I'm looking for constructive, informed, kind, helpful participation.

II **Discussion leading (25%):** Each student will sign up to lead class discussion on the assigned readings during one course meeting. Students may register individually, or in pairs. See assignment for details.

III **Essay 1: 30%** 3-5 pages, double spaced. Assigned 2/15, due 2/29.

IV **Essay 2: 30%** 3-5 pages, double spaced. Assigned 3/21, due 4/9.

V **Essay revisions:** Students may choose to submit a revision of either (but not both) of Essay 1 or Essay 2. If the revision receives a higher grade than the original, the grade for the revision will stand in place of the grade for the original. Revised papers

4. Schedule

The course is divided into six units, covering some core topics in the ethics of artificial intelligence. Each begins with a case study (conducted in class, no reading assigned). This is followed by 3-4 days of readings on the chosen topic.

4.1. Introduction

January 9: Course introduction.

4.2. Transparency and explainability

Big questions: Should decisions made by algorithms be transparent and explainable? If so, why? And what does it mean to say that a decision is transparent or explainable, anyways?

Jan 11: Case study (no reading).

Jan 16: Explanation and self-advocacy.

i **Reading:** Kate Vredenburgh, "The right to explanation."

Jan 18: Unexplainable snow. No class.

Jan 23: Explanation and outcomes.

i **Reading:** Elanor Taylor, "Explanation and the right to explanation."

Jan 25: No class.

Jan 30: Explanation and due consideration.

i **Reading:** David Gray Grant et al., "What we owe to decision-subjects: Beyond transparency and explanation in automated decision-making", (Sections 1-6 only).

4.3. Bias

Big questions: Are existing AI systems biased against certain groups of individuals? What does it mean to say that a system is biased? What's wrong with bias? How can bias be mitigated?

Feb 1: Case study (no reading).

Feb 6: Algorithmic bias and the proxy problem.

i **Reading:** Gabrielle Johnson, "Algorithmic bias: On the implicit biases of social technology."

Feb 8: Algorithmic bias and risk.

i **Reading:** Clinton Castro, "What's wrong with machine bias?"

Feb 13: What's in a bias?

i **Reading:** Tom Kelly, "The norm-theoretic account of bias."

4.4. Privacy and surveillance

Big questions: What is the right to privacy? Why is there a right to privacy? In what ways do AI systems infringe on the right to privacy? Special case: What are (im)permissible uses of AI systems in surveillance?

Feb 15: Case study (no reading).

Feb 20: Online privacy.

i **Reading:** Carissa Véliz, "The internet and privacy."

ii **Assigned:** Essay 1 assigned.

Feb 22: Interrogating the right to privacy.

i **Reading:** Judith Jarvis Thomson, "The right to privacy."

Feb 27: Privacy as a human right.

i **Reading:** Beate Roessler, "Privachy as a human right."

Feb 29: Surveillance.

i **Reading:** Kevin Macnish, "Just surveillance? Towards a normative theory of surveillance."

4.5. Labor and the future of work

Big questions: What role will AI play in the future of work? How can we find meaning in a world where the role of human workers has changed? What are permissible uses of AI in the workplace?

Mar 5: Case study (no reading).

i **Due:** Essay 1 due.

Mar 7: Economic effects of AI in historical perspective.

i **Reading:** Joel Mokyr et al., "The history of technological anxiety and the future of economic growth: Is this time different?"

Mar 19: Meaning without work.

- **Reading:** John Danaher, "Will life be worth living in a world without work? Technological unemployment and the meaning of life."

Mar 21: AI in the workplace.

- **Reading:** Kate Vredenburgh, "Freedom at work: Understanding, alienation, and the AI-driven workplace.'

4.6. Moral patiency

Big questions: Can AI systems be moral patients (have rights, claims, etc.)? If so, why? What should we do if we are unsure of the answer to these questions?

I **Mar 26:** Case study (no reading).

- **Assigned:** Essay 2 assigned.

II **Mar 28:** Can AI systems have rights?

- **Reading:** John Basl and Joseph Brown, "AI as a moral right-holder".

III **Apr 2:** Uncertainty about AI rights.

- **Reading:** Jeff Sebo and Rob Long, "Moral consideration for AI systems by 2030?".

IV **Apr 4:** No class.

4.7. Transhumanism

Big questions: Could AI systems exceed their makers? If so, what dangers might AI systems pose, and how should those dangers be mitigated? What do we owe to future generations who will be affected by the AI systems we are building today?

I **Apr 9:** Case study (no reading).

- **Due:** Essay 2 due.

II **Apr 11:** The singularity hypothesis.

- **Reading:** David Chalmers, "The singularity, a philosophical analysis," (Sections 1-2 only).

III **Apr 16:** What would a superintelligence want?

- **Reading:** Nick Bostrom, "The superintelligent will."

IV **April 18:** Examining the evidence.

- **Reading:** David Thorstad, "Against the singularity hypothesis," (Sections 1-3 only). **Note:** Please feel very free to disagree with me!

5. Course policies

5.1. Office hours

I will hold office hours each Thursday from 4-5PM 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.