PHILOSOPHY AND ARTIFICIAL INTELLIGENCE

SYLLABUS

COURSE DESCRIPTION. Artificial intelligence (AI) promises—or threatens—to transform every area of our lives and societies. It has already begun to upend our understanding of human nature, radically alter our social institutions, and revolutionize scientific practice. In some circles, there is increasing concern that AI is developing intelligence to rival our own. We will explore these issues through the philosophy of mind, cognitive science, and ethics.

There will be two main themes. First, it seems that AI systems will eventually possess minds like (or, more daunting, minds quite unlike) ours. Is this right? Could AI systems think, understand, and be creative? Would they do these things the same way we do? And if they did possess minds—especially if they became conscious—what obligations would we have to them? What obligations would *they* have to us? Would they agree with us about what those obligations were?

The second theme is a more pressing worry about how to use AI ethically. AI already influences what we buy, how we vote, and who we become friends with. Often, it's beneficial. It can provide solutions to difficult problems, sometimes overcoming bias (e.g., matching algorithms in dating apps have increased dating across socioeconomic classes). But it can also be harmful, even introducing or reinforcing bias. E.g., AI in court systems has been found to grant or deny parole based on the applicant's race. So what are our responsibilities when using AI? Are there problems we shouldn't allow AI to touch? And how can we tell whether AI systems are biased? Can we even know why they make the decisions they do?

LEARNING OBJECTIVES. By the end of this course you will understand some important advances in artificial intelligence, along with the philosophical and ethical questions they raise. You will have gained experience taking an interdisciplinary approach to difficult problems. And you will have developed your writing, communication, and research skills through projects directly relevant to your own goals and interests.

STRUCTURE OF THIS DOCUMENT. This whole document is required reading for the course. The next page contains basic information about people, grades, deadlines, and readings—the kind of thing to keep by your desk or on your desktop so you have important information handy.

After that, there are overviews of the assignments, grading scheme, and course policies. You should read through these at the start of the semester, and then refer back to them as they become relevant again. The final section contains optional readings for each week of the course.

PHILOSOPHY AND ARTIFICIAL INTELLIGENCE

BASIC INFORMATION

Time: Wed, 12:10-2:00
Location: Philosophy 716

Instructor: Andrew Richmond, ar3688@columbia.edu
Office hours: Online, Thurs 12:00-2:00, claim a spot **here**

ASSESSMENT *See next section for details*

10% Two Participation Logs, 5% each
50% Six Writing Exercises, from 4-12% each
40% Final Project (includes 5% from a *Project Proposal* and 5% from a *Project Reflection*)

SCHEDULE • = *reading*, ⊕ = *due date*; *all readings (plus optional ones) posted on Courseworks*

Jan 19 **Introduction & Background**
- This syllabus, top to bottom
- The *Introduction* and any two *commentaries* from this blog post, plus GPT-3's response here
- Mitchell, Chapters 1-3

Jan 26 **What AI is:** Symbols and Networks
- Marcus & Davis, Chapters 3-5
- Minsky, Steps Toward Artificial Intelligence
- Sharkey & Sharkey, "Connectionism"

Feb 2 **What AI is:** Symbols *vs* Networks
- Buckner, "Deep Learning"
- Schneider, "The Language of Thought"
- Marcus & Davis, Chapters 6-7

⊕ **Writing Exercise 1**

Feb 9 **What AI could be:** A Mind
- Mitchell, Chapters 14 & 15
- Turing, "Computing Machinery & Intelligence"
- Bayne et. al., "What is Cognition?"

⊕ **Writing Exercise 2**

Feb 16 **What AI could be:** Linguistic
- Mitchell, Chapters 11-13
- Boden, Chapter 6
- Floridi & Chiriatti, "GPT-3- Its Nature, Scope, Limits, and Consequences"

⊕ **Writing Exercise 3**

Feb 23 ***No Class — watch or read something(s) from the optional reading list and write a short reaction to it***

Mar 2 **What AI could be:** Creative
- Boden, "Creativity and Artificial Intelligence"
- Shevlin, "Rethinking Creative Intelligence"
- Halina, "Insightful artificial intelligence"

⊕ **Writing Exercise 4**

Mar 9 **What AI could be:** Conscious
- Jackson, "What Mary Didn't Know"
- Schneider, Chapters 1–4

⊕ **Writing Exercise 5**

Mar 16 ***No Class: Spring Break***

Mar 23 **Ethics of AI:** Moral Consideration
- Shevlin, "How Could We Know When a Robot Was a Moral Patient?"
- Andreotta, "The Hard problem of AI Rights"
- Gerdes, "The Issue of Moral Consideration in Robot Ethics"
- Hildt, "Artificial Intelligence: Does Consciousness Matter?"

⊕ **Writing Exercise 6 (DUE FRIDAY the 25th)**
⊕ **Participation Log 1**

Mar 30 **Ethics of AI:** Bias
- Angwin et. al., "Machine Bias"
- Zou & Schniebinger, "AI Can Be Sexist & Racist"
- Howard & Borenstein, "The Ugly Truth About Ourselves and Our Robot Creations"
- Fazelpour & Danks, "Algorithmic bias: Senses, sources, solutions"

Apr 6 **Interlude:** Explainable AI
- Voosen, "The AI Detectives"
- Rahwan et al, "Machine Behavior"
- Creel, "Transparency in Complex Computational Systems"

⊕ **Project Proposal (DUE FRIDAY the 8th)**

Apr 13 **Interlude:** Explainable AI *Cont.*
- Lipton, "The Mythos of Model Interpretability"
- Zednik, "Solving the Black Box Problem"
- Egan, *Draft TBD*

Apr 20 **Ethics of AI:** Bias *Cont.*
- Buolamwini & Gebru, "Gender Shades"
- Bender et al, "On the Dangers of Stochastic Parrots"

Apr 27 **TBD depending on interest:** the singularity; ethics of automation; mind uploading; privacy & data; AI in science; ownership of/responsibility for AI; … if more ethics, possibly:
- Danks & London, "Algorithmic Bias in Autonomous Systems"
- Castro, "What's wrong with machine bias?"
- Bietti "From Ethics Washing to Ethics Bashing"

⊕ **Final Project**

One Week After Feedback Received for Final Project

⊕ **Project Reflection**
⊕ **Participation Log**

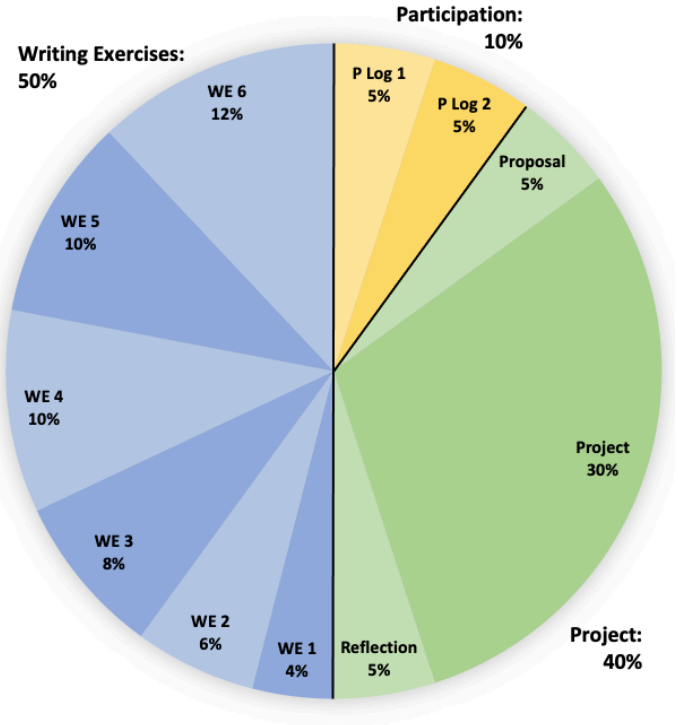
PHILOSOPHY AND ARTIFICIAL INTELLIGENCE

COURSE WORK AND GRADING

FINAL PROJECT, 40%, 5% of that from a *Project Proposal* and 5% from a *Project Reflection*.

You have a lot of latitude to decide what your project looks like. You're welcome to write a traditional philosophical paper, engaging in one of the debates we've discussed and arguing for a position on it. In that case the paper should be around 4500 words. But you might instead write a debate or dialogue with another member of the class, or a brief for an organization or company to advocate for or against the use of some AI technology. Or you could create something else entirely, a work of fiction, a computer program, etc. *The only requirement is that your project engages seriously and critically with the material and topics we've discussed in class.* The project proposal (next paragraph) and the writing exercises (below), especially the sixth, will help you develop your project.

The project proposal, worth 5%, is due three weeks before the project itself. It should be roughly 300 words, and it should do two things. (1) Describe the project you're going to take on: the type of project (academic paper, computer program, etc.); the content (the specific course issues and material you'll discuss); and the expected result (an argument for a certain view, a program that solves some problem, etc.). (2) Explain why the project is a good fit given your learning goals, and what you hope to get out of it. Maybe you just like writing philosophy papers, so you're writing a traditional paper. Or maybe you're going into a career in government, and you're writing a brief to a government organization because you want to build skills you'll need in that career. Within a week I'll review your proposal and give you feedback. I will also propose a grading rubric or a set of expectations based on the details you've given me. We'll discuss both the project and the rubric as necessary, and we'll modify them together in the following week to settle on a concrete plan for your final project. (Or you may rewrite the proposal from scratch if you want.)



The project itself will be worth 30% of your final grade. It will be graded according to the expectations we discussed, and I'll send the grade and feedback a week after the project is submitted.

The project reflection, worth 5%, is due one week after you receive your grade and feedback. In roughly 300 words, you'll be asked to reflect on the comments I gave you, and more generally on the goals you had for the project, how well it achieved them, and what lessons you'll take forward.

WRITING EXERCISES. Six, from 4-12% each.

These will be between 100 and 750 words long and should be handed in before class on the day they're due. They're here so that you can practice some of the skills you'll need in your larger project. *The exercises start very low stakes, with simple goals. As the semester progresses, you will combine the skills you practiced in earlier exercises, building toward the kind of work your final project will involve.*

The first skills you'll need in your final project are *clear thinking* and *argument analysis*. So the first writing exercise (100 words, worth 4%) will ask you to examine a paper we're reading in class and clearly report its conclusions. The second exercise (250 words, 6%) will ask you to do the same for another reading, and also report the basic arguments for those conclusions. And the third writing exercise will ask you to dig deeper into the structure of an argument by discussing points where it is susceptible to objections (350 words, 8%). On each of these, you'll also reflect (in another ~100 words) on how you went about doing the assignment, what came easily

and what was more difficult, and how the paper in question made it easier or more difficult to identify its arguments and conclusions.

Another essential skill for your final project is *interdisciplinary thinking*. Your projects will be informed by an understanding of both philosophical debates and developments in AI, among other things. So in your fourth and fifth writing exercises (each 500 words, 10%) you'll be asked to examine one of our readings, do everything that you did in the first three exercises, and then consider the interdisciplinary nature of the arguments—the way different sources of evidence come to bear on the same questions, and the challenges for both author and reader in handling those different sources of evidence. On each of these, you'll again reflect (in another ~100 words) on how you went about doing the assignment, what came easily and what was more difficult, etc.

The sixth writing exercise will ask you undertake a short project of the kind you're planning to do for your final project (750 words or equivalent, 12%). The topic will be up to you (though I can provide suggestions). This is a chance to experiment so that you know what you're getting into with your final project, and you have a chance to identify skills you'll need to use and pitfalls you'll need to avoid. Like the earlier exercises and the final project itself, this exercise will have to engage critically with arguments about AI—examining their conclusions, points of contention, sources of evidence, etc. And again, you'll reflect (in another ~100 words) on how you went about doing the assignment, what came easily and what was more difficult, etc., and what lessons you can take into your final project. You may rewrite the sixth writing exercise, if you wish.

PARTICIPATION LOGS. Two, 5% each.

You should come to class having done the readings, and ready to discuss them. You'll use these participation logs to track and reflect on your participation in the course. The logs will look like this:

Participation. Log at least 15 distinct examples of your participation in class. Each one need only be one or two sentences.

	Date	*Example*
1	Sep 16	Added an item to our breakout group's list of questions about AI
2	Sep 18	Asked a clarifying question about the "Frame Problem"
…	…	…

Reflection. Briefly (6-10 sentences) reflect on the way you contributed to the class in the examples above.

… … …

You'll receive full marks for a participation log if it is *accurate* and *complete*.

The examples needn't be more than a sentence. They should be spread out over the course of the semester and should include some examples of participation in both small group and whole class activities. *The examples do not need to be times when you shared a thought of your own.* Other good examples include: asking someone else to elaborate on or clarify their comment; making a point of order (e.g. that someone didn't get to finish their point); *holding in* a comment when others haven't had a chance to speak; emailing Andrew about missing files on Courseworks; mentioning to Andrew that something made you uncomfortable speaking up in class; and so on. There are many ways to contribute to a class. What I want to see in these logs are examples of behavior that you think contribute to the class's discussions or broader goals, *however you think it is useful to contribute to them*.

The reflection should be 6-10 sentences. You should think about what you hope to contribute to the collective learning endeavor that is this class, and whether there are things you could do differently to reach that goal. If this is the second reflection of the semester, you should discuss how you've addressed the areas for change you identified in the first one.

PHILOSOPHY AND ARTIFICIAL INTELLIGENCE

COURSE POLICIES

OFFICE HOURS. *Please take advantage of these.* I hope to see you all in office hours at some point this semester. You can come with questions about course material, about philosophy, about grad school, or anything else. Or you can just drop by to chat—no need to have questions prepared in advance. If you can't make the time listed, get in touch. I'm very happy to set up another time to meet.

LATE ASSIGNMENTS. By default, late Writing Exercises will receive zero marks, and late projects will be docked 1/3 of a grade per day (so a B+ two days late becomes a B–). *I will be very generous with extensions for both, as well as the Participation Logs*. But if you need an extension, you should contact me as soon as possible about it.

REWRITES. You can rewrite the Project Proposal and Writing Exercise 6. To do this you just have to inform me that you're doing a rewrite within a week of getting your feedback on the original submission and describe the issues with your previous submission that your rewrite will improve, and how you plan to improve them.

ELECTRONICS. The scientific literature on electronics in the classroom is mostly unequivocal: *they are bad for learning*. If you use a computer in class—even if you just plan to take notes on it—you remember less about the material, particularly about conceptual as opposed to factual issues, i.e. the ones that are most important in a philosophy course. Using your phone is worse. Even having your phone *on your person* has a slight distracting effect. Using electronic devices, especially laptops, can also be distracting for the people around you. *I won't have a strict policy about this, so if you need, or if you just really want to use your computer, you can.* But you should be aware of all the above and try to mitigate the negative effects. (Obviously this is all complicated by the online environment, so if any of this advice conflicts with your ability to attend class or do the work, ignore the advice.)

ACADEMIC SUPPORT. You should take advantages of the resources you have here at Columbia. As usual, the Writing Center can be consulted for advice about academic writing, and the Libraries can be a great source of advice on the practicalities of research. You can also come to me about either of those, or about other sorts of advice—e.g. on building work habits or finding opportunities for interdisciplinary collaboration—and I'll try to direct you somewhere helpful.

ACCESSIBILITY. You can find the Faculty Statement on Disability Accommodation here. If you have a DS-certified Accommodation Letter, please get in touch with me as soon as possible about any accommodation needs I should be aware of. If you think you might have a disability that requires accommodation, you should contact Disability Services at 212-854-2388 or disability@columbia.edu. You should also feel free to come to me directly with any issues. More generally, I hope you'll let me know if there's anything I can do to make the class more accessible or inclusive, or if there's any way I can make it easier for you to participate and thrive.

ACADEMIC INTEGRITY. You can find the Faculty Statement on Academic Integrity here. The work in this course will all be individual, unless you choose a final project in which you'll work as a group. For individual work you can consult with each other, but the work you turn in must be your own, and any sources you draw from must be explicitly credited. If you do a group project you will have to let me know ahead of time how everyone in the group will contribute.

PHILOSOPHY AND ARTIFICIAL INTELLIGENCE

OPTIONAL READINGS

Non-fiction is in **black**
Fiction is in **green**
Movies, TV series, and documentaries are in **orange**

General Background Reading

- Harris, "Tips on Reading a Scientific Paper"
- Purugganan & Hewitt, "How to Read a Scientific Article"
- Wallisch, "How to read a scientific paper"
- Rippon, "A Brief Guide to Writing the Philosophy Paper"
- Article: Writing A Philosophy Paper
- Pryor: Guidelines on Writing a Philosophy Paper

Jan 19

- Margaret Boden, *AI: It's Nature and Future*
- Buchanan, "A (Very) Brief History of Artificial Intelligence"
- Landgrebe & Smith, "An argument for the impossibility of machine intelligence"
- Garfinkel et al, "On the Impossibility of Supersized Machines"
- SEP, "Artificial Intelligence"
- Mitchell, "Why AI is Harder Than We Think"
- *Westworld*, Seasons 1 and 2 *(and stop there, for your own sake)*

Jan 26

- Churchland & Sejnowski, "Neural Representation and Neural Computation"
- Fodor & Pylyshyn, "Connectionism and Cognitive Architecture"
- Newell et al, "Report on a General Problem-Solving Algorithm"
- Newell et al, "A General Problem-Solving Program for a Computer"
- Rosenblatt, "The Perceptron"
- Braitenberg, *Vehicles: Experiments in Synthetic Psychology*

Feb 2

- Palvus, "Common Sense Comes Closer to Computers"
- Article: How Deep Learning Works
- McCarthy & Hayes, "Some Philosophical Problems from the Standpoint of Artificial Intelligence"
- Maloney, "In Praise of Narrow Minds: The Frame Problem" (in Fetzer, *Aspects of Artificial Intelligence*)
- Fodor & Pylyshyn, "Connectionism and Cognitive Architecture"
- Goertzel, "Artificial General Intelligence"
- Goertzel, "The General Theory of General Intelligence"
- Boden, "Has AI helped psychology?"
- Buckner, "Black Boxes or Unflattering Mirrors?"
- DeepMind, "Player of Games"

Feb 9

- Searle, "Minds, Brains, and Programs" (in Levitin, *Foundations of Cognitive Psychology*)
- Boden, "Escaping from the Chinese room"
- Saxe et al, "If deep learning is the answer, what is the question?"
- Dennett, "Can Machines Think?" (in Levitin, *Foundations of Cognitive Psychology*)
- Winograd, "Thinking Machines"
- Block, "The Mind as Software in the Brain"

- *Her*, 2013

Feb 16

- Briggs, "Knowledge Representation in Sanskrit and Artificial Intelligence"
- Alarie & Cockfield, "Machine-Authored Texts and the Future of Scholarship"

Mar 2

- Scheines, "Automating Creativity" (in Fetzer, *Aspects of Artificial Intelligence*)
- Boden, "Creativity in a nutshell"
- Dijkstra, The Value of Creativity
- Article: AI Designs Quantum Physics Experiments beyond What Any Human Has Conceived
- Richard Powers, *Galatea 2.2*

Mar 9

- Maudlin, "Computation and Consciousness"
- Shevlin, "Non-human consciousness and the specificity problem"
- Dehaene et. al., "What is Consciousness, and Could Machines Have It?"
- Udell & Schwitzgebel, "Susan Schneider's Proposed Tests for AI Consciousness"
- Article: Machine in the ghost
- What is it like to be a computer? An interview with GPT-3 [link]

Mar 23

- Gerdes, "The Role of Phronesis in Robot Ethics"
- Bostrom & Yudkowsky, "The ethics of artificial intelligence"
- Neely, "Machines and the Moral Community"
- Article: AI cannot be the inventor of a patent, appeals court rules
- Article: From Mind-as-Computer to Robot-as-Human: Can metaphors change morality?
- Mary Shelley, *Frankenstein*
- Kazuo Ishiguro, *Klara and the Sun*
- Ted Chiang, *Dacey's Patent Automatic Nanny*
- *Ex Machina*, 2014

Mar 30

- Coghlan et al, "Good Proctor or 'Big Brother'? AI Ethics and Online Exam Supervision Technologies"
- Article: 'Orwellian' AI lie detector project challenged in EU court
- Article: The role of the arts and humanities in thinking about artificial intelligence

Apr 6

- Phillips et al, "Four Principles of Explainable Artificial Intelligence"
- Article: How much should we trust technology?
- Ted Chiang, *The Evolution of Human Science*

Apr 13

- Coyle & Weller, "'Explaining' machine learning reveals policy challenges"
- Poursabzi-Sangdeh et al, "Manipulating and Measuring Model Interpretability"
- Rudin, "Stop Explaining Black Box Machine Learning Models for High Stakes Decisions and Use Interpretable Models Instead"

Apr 20

- Raji & Buolamwini, "Actionable Auditing"
- Lima & Cha, "Responsible AI and Its Stakeholders"
- Fazelpour & Lipton, "Algorithmic Fairness from a Non-Ideal Perspective"
- Article: This is how AI bias really happens—and why it's so hard to fix
- Article: AI Recognises Race in Medical Images
- Article: Tech-industry AI is getting dangerously homogenized, say Stanford experts

Other

- Eden et al, *Singularity Hypotheses*