Philosophy 223:
Minds and Machines

PHIL 223, Credit hours: 3.0, Spring 2022, T&Th: 2-3:20pm, Room: 160 English Building or via Zoom (synchronous meetings)

Instructor: Guillermo Del Pinal (he/him); *Email:* delpinal@illinois.edu; *Office:* Gregory Hall 204 (or via Zoom); *Office Hours:* Thusdays 4-6pm

Teaching Assistants:

- Dan Durso (he/him); *Email:* ddurso3@illinois.edu;
- Henzel Kim (he/him); *Email:* hkim175@illinois.edu;

Prerequisites: None

Course Outline:

Can machines think?	2 weeks
Classical architectures and the mind	1 week
Connectionism, neural nets, and deep learning	2 weeks
Smart robots, embodied mind, extended mind	1 week
Induction, human learning, machine learning	2 weeks
Machine learning, big data, and society	2 weeks
Consciousness and strong AI	2 weeks
Personal identity, embodiment, immortality	2 weeks
Responsibility, control and The Singularity	2 week

Course Description: This course will examine (i) the nature of human minds and brains in light of what we know about machine 'minds' and 'hardware', and vice versa, and (ii) how the rise of intelligent machines is affecting and reshaping our own society. The course will explore questions like: Could a machine have a mind? What can human minds and brains teach us about how an intelligent machine might work, and vice versa? Can machines learn to master various different domains in ways that simulate and eventually even surpass the astonishing capacity and flexibility of human learning? Could a machine think in the ways humans do? How could we tell? How do machines and our interactions with them influence, affect and enhance how humans think, learn, and reason? What are the promises and perils of our increasing dependence on artificial intelligence, big data, and social networks? How should we, as a society, confront situations in which the underlying processes behind machine 'decisions' are not transparent to us? When machines are trained on human generated data such as news corpora, what kinds of human-like social biases—including race and gender—might they re-create/incorporate into their 'decisions'? Can we reduce the effect of race, gender, and other social biases in machine learning decisions without degrading their overall performance? From a normative perspective, how can traditional philosophical theories of fairness and justice help us think about machine biases, and understand the relevant trade-offs? What is the nature of our personal identity? Is it a serious possibility that human minds can be transferred

to mediums other than biological bodies? What might information processing, integration and flexibility have to do with consciousness? How can we tell if machines become conscious? How should machines that exhibit some non-trivial features of human minds be treated? How should machines treat us?

Learning outcomes: This course will achieve the following learning outcomes from the *Intellectual Reasoning and Knowledge* theme in UIUC's Campus-wide Student Learning Outcomes:

- Students will gain an appreciation of analytical and philosophical thinking.
- Students will acquire historical and philosophical perspectives on computing and computers, and related topics such as machine learning.
- Students will grasp from a philosophical perspective foundational ideas in computer science—e.g., the nature of computers, algorithms, machine learning, etc.
- Students will develop a reflective orientation toward conceptual issues at the intersection of computer science and philosophy.

This course will also achieve the following learning outcomes from the *Social Awareness and Cultural Understanding* theme:

- Students will reflectively engage with issues concerning the impact of computing on society.
- Students will develop a critical awareness of ethical issues surrounding computing, and of the moral responsibility of computing professionals and the wider society.
- Students will learn about philosophical foundations of moral principles underlying these issues: e.g., under what conditions should we protect intellectual property and under what conditions can big data be used to predict the behavior individuals.

Grade Policy and Schedule:

- Canvas responses worth 25% of final grade. Every week, students can submit critical reactions/comments on the weekly readings. Each response should be between 200-400 words and submitted before the corresponding class for the reading. Each response will be grade on a 'pass/fail' basis. Each students has to obtain 8 'pass' marks to obtain the full points for this module of the course. To obtain a pass mark, the responses should demonstrate a serious critical engagement with the reading.
- 3 take home assignments, mainly with short essay questions, each assignment is worth 25% of the final grade.
- *Grading scale:* Numerical grades are converted to final letter grades according to the following scale and rounding:

Letter:	F	D-	D	D+	C-	C	C+	B-	B	B+	A-	A	A+
%:	>60	60	63	66	70	73	76	80	83	86	90	93	96

Academic Integrity: Every student should familiarize themselves with the sections of the Student Code that define infractions of academic integrity and list the possible, associated penalties at:

- hhttps://studentcode.illinois.edu/article1/part4/1-401/

See also see

- https://www.las.illinois.edu/students/integrity/

and read the academic integrity policy of the College of Liberal Arts & Sciences.

Presenting other people's work as your own on any assignment, quiz, essay, or exam is a serious offense and will be treated as such. While it is permitted to collaborate with others on assignments and essays (unless noted otherwise), the collaboration must not result in any part of the submitted work that cannot be ascribed to you and you alone.

Accommodations for students with disabilities: In compliance with the University of Illinois's policy and equal access laws, I am available to discuss appropriate academic accommodations that may be required for student with disabilities. To obtain disability-related academic adjustments and/or auxiliary aids, students with disabilities must contact the course instructor and the Disability Resources and Educational Services (DRES) as soon as possible. To contact DRES, you may visit 1207 S. Oak St., Champaign, call 333-4603, e-mail disability@illinois.edu or go to the DRES website.

Inclusivity statement: Our members represent a rich variety of backgrounds and perspectives. The Philosophy Department is committed to providing an atmosphere for learning that respects such diversity. While working together to build this community we ask all members to: (i) share their unique experiences, values and beliefs, (ii) be open to the views of others, (iii) honor the uniqueness of their colleagues, (iv) appreciate the opportunity that we have to learn from each other in this community, (v) value each other's opinions and communicate in a respectful manner, (vi) keep confidential discussions that the community has of a personal (or professional) nature, (vii) use this opportunity together to discuss ways in which we can create an inclusive environment in this course and across the UIUC community.

Extra Help: Please do not hesitate to come to the instructor's (virtual) office hours to discuss any aspect of the course. Also, if you are interested in obtaining information to improve writing skills and organization, the following campus resources are available to all students: Writer's Workshop, Undergrad Library, 217-333-8796. For a very useful guide to writing a philosophy or any theoretical paper, see Jim Pryor's Guidelines on Writing a Philosophy Paper.

University Class Attendance Policy: Students are expected to attend classes regularly. A student who incurs an excessive number of absences may be withdrawn from a class at the discretion of the instructor. For details, see the University of Illinois's Student Code Part 5.

Zoom meetings, netiquette and accessibility: This course will meet in person, but depending on campus instructions, sometimes we might have to go live (synchronously) on T&Th, 2-3:20pm. The Zoom link will be available on Canvas course site! Some of you may be new to Zoom, and here are some guidelines for Zoom Netiquette:

- Do have the video and microphone turned on in every synchronous class. This way we can keep the class as similar as possible to an in-person class and get the most out of this class. Of course, this isn't set in stone—I encourage you to contact me prior to the class session to discuss expectations and accommodations needed.

- Do use your 'hand' emoji if you want to indicate that you'd like to contribute to the class, ask questions, etc. Do that as often as you like!
- Please make sure that you are muted when you're not speaking, especially if your background is noisy.
- Don't multitask and use electronic devices for purposes not related to the class content. It's tempting, but don't text with friends, check your emails, etc. You may use your electronic devices only for Zoom, note-taking, and the in-class tasks I give you.
- Try (if possible) to be in a quiet environment without much going on in the background; avoid chewing into the camera; avoid moving around. Feel free to use Zoom backgrounds if you'd like! (But, of course, they shouldn't contain anything inappropriate.)
- The synchronous sessions will be recorded but only provided to students who missed the session for some justified reason. Do not share any recorded content outside the classroom.
- Please make sure your electronic device and Zoom version allows you to use the main Zoom functions such as break-out rooms, the chatbox, and polls.

Another important issue concerns Zoom Accessibility:

- The 'digital divide' is accentuated by online teaching approaches. For many reasons, a considerable fraction of students do not have the technology hardware, expertise and/or access that instructors may take for granted.
- Campus has programs to help bridge the digital divide. LAS has the 'ATLAS Share' program, to assists students obtain adequate computers and/or internet connectivity. It is described here: https://atlas.illinois.edu/student-information/atlas-share

Readings: All readings will be available at the course canvas website.

Tentative reading schedule: (**Last Update: January 26, 2022**)

PART 1:

Week 1 (T: 1/18, Th: 1/20): Can machines think?

- **Main Readings:**
  - Turing, A. (1950). Computing machinery and intelligence. *Mind*, LIX(236): 433–460.
- **Further Readings:**
  - Descartes, Discourse on the method, ch. V. and Letter to the Marques of Newcastle. Reprinted in Schieber (ed.) *The Turing Test: Verbal Behavior as a Hallmark of Intelligence* ch. 1-2.

***Notes:** We might also discuss parts of Block, N. (1989). Psychologism and Behaviorism in *Philosophical Review*, XC(1): 5-43, and a recent response by Schieber, S. M. (2014). There can be no Turing-test-passing memorizing machines in *Philosophers Imprint*, 14(16): 1-13.

Week 2 (T: 1/25, Th: 1/27): Classical architectures and the mind

- **Main Readings:**
  - Block N. (1995). The mind as the software of the brain. Ch. 11 in *Thinking: An Invitation to Cognitive Science. 2nd ed. Vol. 3.*
- **Further Readings:**
  - Marr, D. (1982). Understanding complex information processing systems. From *Vision*, ch 1.

***Notes:**

Week 3 (T: 2/1, Th: 2/4): Can machines think? A skeptical interlude

- **Main Readings:**
  - Searle. J. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, 3: 417-457.
  - Boden, M. (1990). Escaping the Chinese Room. Ch. 4 in *The Philosophy of Artificial Intelligence.*
- **Further Readings:**
  - Chalmers, D. (1992). Sub-symbolic computation and the Chinese room. In *The Symbolic and Connectionist Paradigms: Closing the Gap.*

***Notes:**

Week 4 (T: 2/8, Th: 2/10): Connectionism and Deep Learning I

- **Main Readings:**
  - Rumelhart, (1989/1998). The architecture of mind: A connectionist approach. from (ed. Haugeland) *Mind Design II.*
- **Further Readings:**
  - Hinton, G. (1992). How Neural Networks Learn from Experience. *Scientific American.*

***Notes:** For a useful and entertaining overview of Connectionism, see Pinker (1997: pp. 98-131), section 'Replaced by a machine & Connectoplasm'. That material also includes a nice discussion of some objections to deep learning presented Marcus (2017), which we discuss week 5.

Week 5 (T: 2/15, Th: 1/17): Connectionism and Deep Learning II

- **Main Readings:**
  - Marcus, G. (2017). Deep Learning: A critical appraisal.
  - Marcus, G. (2017). Innateness, AlphaZero, and Artificial Intelligence.
- **Further Readings:**
  - DeepMind algorithm beats people at classic video games. *Nature.*
  - Buckner, (2019). Deep Learning: A philosophical introduction. *Philosophy Compass.*

***Notes:** An influential criticism of Connectionism is Fodor and Pylyshyn's 'Connectionism and cognitive architecture: A critical analysis'. For a response on behalf of Connectionism, see Smolensky's 'On the proper treatment of Connectionism'. That debate from the 80s and 90s is covered/updated in the readings by Buckner and Marcus. For some recent promising/reviews of DNNs (Deep neural nets) that learn/represent hierarchical structure, focusing on the domain on language, see Linzen and Baroni (2021) and Manning et al. (2020).

Week 6 (T: 2/22, Th: 2/24): Intelligent robots, embodied mind, extended mind

- **Main Readings:**
  - Brooks, R. A. (1991) Intelligence without representation. *Artificial Intelligence*, 47(1–3): 139-159. Reprinted in (ed. Haugeland) *Mind Design.*
  - Clark, A. and Chalmers, D. (1998). The extended mind. *Analysis* 58(1): 7-19.
- **Further Readings:**
  - Dreyfus. (1992). The role of body in intelligent behavior. In (1992 edition) *What Computers Still Can't Do*, ch 7.

***Notes:** For a fun presentation of some of the ideas in the extended/embodied mind readings for this week, check out Chalmer's famous Ted talk ('Is your phone part of your mind?'). You might also want to check out Weigmann (2012), 'Does intelligence require a body?'

Week 7 (T: 3/1, Th: 3/3): Induction, human/machine learning I

- **Main Readings:**

- Harman, G. and Kulkarni, I. (2007). The problem of induction, ch. 1 from *Reliable Reasoning*
- Harman, G. and Kulkarni, I. (2007). Induction and the VC dimension, ch. 2 from *Reliable Reasoning*

- **Further Readings:**
  - Goodman, N. (1953). The new riddle of induction, from *Fact, Fiction and Forecast*

***Notes:** For background on induction—and specially for those of you who didn't take Phil 222—see Leah Henderson, 'The Problem of Induction', in Edward N. Zalta, ed., *The Stanford Encyclopedia of Philosophy* .

Week 8 (T: 3/8, Th: 3/10): Induction, human/machine learning II

- **Main Readings:**
  - Harman, G. and Kulkarni, I. (2007). Induction and "Simplicity', ch. 3 from *Reliable Reasoning*.
  - Harman, G. and Kulkarni, I. (2007). Neural networks, support vector machines and transduction, ch. 4 from *Reliable Reasoning*.
- **Further Readings:**
  - Aaronson, S. (2013). Why philosophers should care about computational complexity, Section 7. from *Computability: Turing, Gödel, Church, and Beyond*.

***Notes:**

PART 2: (Week 9: Spring break 3/14-3/20)

Week 10 (T: 3/22, Th: 3/24): Machine learning, Big Data, and society I

- **Main Readings:**
  - O'Neill, C. (2017). *Weapons of Math Destruction*, ch. 1 and ch. 3-4
  - O'Neill, C. (2017). *Weapons of Math Destruction*, ch. 5-7
- **Further Readings:**
  - O'Neill, C. (2017). *Weapons of Math Destruction*, ch 9, 10, and 'Conclusion'

***Notes:** In this part of the course, we explore the social impact of outsourcing various decisions to machines. This include issues such as medical decisions, credit ratings and interests, and various kinds of rankings—including relevance of personal webpages relative to queries, movie recommendations, etc. A central topic this week and the next concerns the ways in which human generated data can includes certain social biases—including racial and gender biases—which are then incorporated into machine learning-based decisions. We will discuss issues of fairness and whether we can reduce such social biases.

Week 11 (T: 3/29, Th: 3/31): Machine learning, Big Data, and society II

- **Main Readings:**

- Caliskan et al. (2017). Semantics derived automatically from large corpora contain human-like biases. *Science*
- Garg et al. (2018). Word embeddings quantify 100 years of gender and ethnic stereotypes. *Proceedings of the National Academy of Sciences*. vol. 115(16)

- **Further Readings:**
  - Barocas and Selbst (2016). Big data's disparate impact. *California Law Review*, vol. 104: 671-729.
  - Zou, J. and Schiebinger, L. (2018). AI can be sexist and racist—it's time to make it fair. *Nature*

***Notes:**

Week 12 (T: 4/5, Th: 4/7): Consciousness and strong AI I

- **Main Readings:**
  - Smart, J. J. C. (1954). Sensations and brain processes. *The Philosophical Review*, 68(2): 141-156.
  - Nagel, T. (1974). What is it like to be a bat? *The Philosophical Review*, 83(4): 435-450.
- **Further Readings:**
  - Dennett, D. C. (1988). Quining qualia. In *Consciousness in modern science*.
  - Chalmers, David J. (2003). Consciousness and its place in nature. In *Blackwell Guide to the Philosophy of Mind*. pp. 102-142.

***Notes:** Additional/alternative readings: General background: Van Gulick (2018), 'Consciousness', in *The Stanford Encyclopedia of Philosophy*.

Week 13 (T: 4/12, Th: 4/14): Consciousness and strong AI II

- **Main Readings:**
  - Tonomi et al. (2019). Integrated information theory: from consciousness to its physical substrate. *Nature Reviews Neuroscience*, 17(7), 450
  - Dehaene et al. (2017). What is consciousness, and could machines have it? *Science*, 358(6362): 486-492.
- **Further Readings:**
  - Danaher, J. (2020). Welcoming Robots into the Moral Circle: A Defense of Ethical Behaviorism. *Science and Engineering Ethics* 26, pp. 2023–2049.
  - Shevlin, H. (forthcoming). How could we know when a robot was a moral patient? in *Cambridge Quarterly of Healthcare Ethics*.

***Notes:**

Week 14 (T: 4/19, Th: 4/21): Identity, embodiment, immortality (in the cloud) I

- **Main Readings:**
  - Parfit, D. (1984). What we believe ourselves to be, ch. 10 from *Reasons and Persons.*
  - Watch 'San Junipero' from *Black Mirror* Series 3, episode 2 and 'USS Callistel' from *Black Mirror* Series 4, episode 1.
- **Further Readings:**
  - Locke, J. An Essay Concerning Human Understanding, Book II, Chapter XXVII ("On Identity and Diversity")

***Notes:** General background: Olson (2019), 'Personal Identity', in *The Stanford Encyclopedia of Philosophy.*

Week 15 (T: 4/26, Th: 4/28): Identity, embodiment, immortality (in the cloud) II

- **Main Readings:**
  - Parfit D. (1984). How we are not what we believe, ch. 11 from *Reasons and Persons.*
  - Corabi, J. and Schneider, S. (2012). The Metaphysics of Uploading, *Journal of Consciousness Studies.*
- **Further Readings:**
  - Chalmers (2014). Mind uploading: A Philosophical analysis. In *Intelligence Unbound: The Future of Uploaded and Machine Minds* (an excerpt from the longer 'The Singularity: A philosophical analysis' paper)

***Notes:**

Week 16 (T: 5/3): Responsibility, control, and The Singularity

- **Main Readings:**
  - Bostrom and Yudkowsky, (2011). The Ethics of Artificial Intelligence. In *The Cambridge Handbook of Artificial Intelligence.*
- **Further Readings:**
  - Chalmers, D. (2010). The Singularity: A Philosophical Analysis. *Journal of Consciousness Studies* 17:7-65.

***Notes:**