Term Year – Course Number, Section

Philosophical Implications of Artificial Intelligence

Class Number: ####, Delivery Method: In Person

Course Information

Course Times + Location
Date and Time
Building and Room Number

Exam Times + Location
Date and Time
Building and Room Number

Contact Information
Instructor: Travis LaCroix
✉ tlacroix@uci.edu

Office Hours
Date and Time
Building and Office Number

Contents

Calendar Description

This course is an examination of some major issues in the philosophy of science.

Prerequisites

Some background in formal philosophy, especially propositional logic, will be valuable but is not strictly required.

Learning Outcomes

Upon successful completion of this course, students should be able to:

- Understand the basic logical framework of contemporary machine learning and AI,
- Be able to define important relevant terms and concepts in machine learning and AI
- Demonstrate knowledge of philosophical issues involved in ethics of AI,
- Demonstrate familiarity with relevant examples of AI systems,
- Identify problems where artificial intelligence techniques are applicable,
- Show ability to work in a small team, and participate in the design of systems that act intelligently and learn from experience.
- Understand and critically assess different conceptions of intelligence and learning.

Detailed Course Description

Artificial intelligence (AI) is developing at an extremely rapid pace. We should expect to see significant changes in our society as AI systems become embedded in many aspects of our lives. This course is meant as an introduction to some contemporary and pressing philosophical and ethical implications of the state-of-the-art in machine learning and artificial intelligence.

However, in philosophy of science, it is important to engage directly with the scientific work in the field one is considering. So, as a background to the philosophical questions which we will discuss in the final part of this course, one of the main objectives in this course is to gain some command of recent work in machine learning and AI in order to cultivate a sophisticated understanding of the real problems facing emerging technologies.

To provide a foundation for questions of artificial intelligence, the first part of the course is centred around the question "what is intelligence?" We will begin with historical conceptions of machine intelligence, and contrast intelligence and learning of this sort with human with animal intelligence as we move forward. In the second part of the course, we will delve into machine learning techniques in order to gain some understanding of how AI programmes are designed, what they are capable of doing, and (importantly) what they are not. Finally, in the third part of the course, we will discuss a number of philosophical issues that arise from machine learning and AI. In particular, we will

discuss (1) practical problems in technical AI safety—i.e., how to ensure machines are safe. This will lead to (2) theoretical problems concerning alignment — i.e., how to ensure that the goals of machines align with a common human good. Finally, we will discuss a number of problems that engage directly with philosophical work, including algorithmic bias and notions of epistemic injustice, legal and societal impacts of artificial intelligence, and (coming back around to our initial questions concerning intelligence), what it means to be human in an artificial world.

Some questions we will consider in each part of the course include the following.

Part I: (Natural) Intelligence and Learning

- What does it mean for an agent to be intelligent?
- What does it mean for an agent to learn?
- How does learning [intelligence] in humans compare to learning [intelligence] in animals or machines? Are these differences a matter of degree, or kind?

Part II: Machine Learning and AI

- What are neural networks?
- What is Backpropagation?
- How do machine learning algorithms work?
- What is the distinction between supervised learning, unsupervised learning, and reinforcement learning?

Part III: Ethical and Philosophical Implications of AI Systems

- How do we align the aims of autonomous AI systems with our own?
- Does the future of AI pose an existential threat to humanity?
- How do we prevent learning algorithms from acquiring morally objectionable biases?
- How should AI systems be embedded in our social relations?
- What sort of ethical rules should AI like a self-driving car use?
- Can AI systems suffer moral harms? And if so, of what kinds?
- Can AI systems be moral agents? If so, how should we hold them accountable?

DISCLAIMER:

Machine learning is a maths-heavy subject. Many of the readings (especially in Part II) are intended for CS students studying AI. As such, many of the readings for this course may seem incomprehensible at first glance. However, the emphasis of the lectures and assignments will focus on understanding the *fundamental concepts* that underlie machine-learning techniques. It is not necessary to understand all of the formalism associated with these readings.

In addition to the required readings, several helpful resources (such as videos and blog posts) will be outlined in the syllabus.

Grading

The final grade will derive from three main "projects" (one for each part of the course), as well as participation.

The final grade will derive primarily from a research paper, due at the end of the semester. Given this course may be cross-listed with a graduate section, the requirements for the undergraduates and the graduate students will vary somewhat.

I will discuss the graded components in more detail during the first meeting, but here is the basic idea:

Assignments for Part I of the course will consist in short written responses to questions. I will distribute the questions at the end of class, and the response will be due the subsequent week. These are *short* responses and should only be around 250-300 words (1 page, double-spaced), and *certainly* no more than 500 words. You should expect to discuss the responses with your colleagues at the start of the class when it is due. **Note**: *the written responses will be graded on a PASS/FAIL basis. What is deemed to be a 'reasonable' effort to engage with the question will be sufficient for complete credit on that week's assignment.*

In Part II of the class, instead of weekly writing assignments, there will be weekly programming assignments. **Note**: there will be extensive instructions for each assignment so that they should be able to be completed with little-to-no coding background. These assignments will lead up to a "coding project", where you design a machine-learning programme to perform a 'simple' task. The larger part of this coding project will be submitted in groups. **Note**: the code for each group will be tested for efficiency, and a bonus 2% will be added to the final grade of the students in that group.

The assignment for Part III will be a short research paper. The final draft of the paper will be due near the end of the examination period, so that you have 1-2 weeks to revise an earlier draft. The earlier draft will be peer-reviewed by two of your colleagues (and likewise, you will peer-review two of your colleagues' papers). The two papers, which will be part of your required reading for the final meeting, will be sent to you one week before the prior meeting (therefore, a draft of your final paper will be due before the penultimate meeting). **Note**: *because of the peer-review component of the paper, the deadline for the first draft is a HARD deadline. Even if the paper is incomplete, you should submit what you have.* Note also that the final essay need not be on a subject that we discuss explicitly in class. For example, we are (sadly) not discussing the application of trolley-problem type thought experiments to self-driving cars. However, if you want to write about this, connecting the subject to the themes that we discuss, that is perfectly acceptable.

Required Readings

All of the readings are freely available online, or will be made available as .pdf files on the course webpage.

> **Note**: if you are auditing or on the waitlist and do not have access to the course web-page, please email me.

Participation	10%
Part I: Short Writing Assignments	15%
Part II: Programme Design (group project)	25%
Part III: Final Paper (according to the following scheme:)	50%

Proposal (500 words)	5%
First Draft (2-5k words)	15%
Peer Feedback (In class)	5%
Abstract of the Paper (< 250 words)	5%
Revised Final Draft (< 6k words)	20%

We will generally not go in detail through every single one of the readings, but it will be important that you have read them, since the lecture will not merely summarise that week's reading.

Policies

For general university-wide policies, see here: [Insert Link to University Policies].

Policies specific to this course are detailed below.

Use of Technology in Class

Laptops, tablets, etc **are** allowed in class; however, please be mindful that what is on your screen may be distracting to those around you.

In case you miss something from the lecture, the slides for the day's lecture, when applicable, will be posted after the lecture on the course web-page.

Ground Rules for Discussions

These ground rules form a set of expected behaviours for conduct in discussions and lectures. They are meant to foster an intellectual atmosphere where we work together to achieve knowledge. They are also meant to ensure that discussions are spirited without devolving into argumentation and to ensure that everyone has an opportunity to be heard.

- Respect yourself and others (share your viewpoint and allow others to share theirs).
- Show respect for others by learning and using their preferred names and pronouns.
- Give each other the benefit of the doubt. (Be charitable.)
- Be cautious of universal claims.
- Listen actively and attentively.
- Keep an open mind. (Expect to learn something new, or to have your views challenged by ideas, questions, and points of view different than your own.)

- Ask for clarification if you are confused.
- Do not interrupt one another—even when you are excited to respond.
- Challenge one another, but do so respectfully.
- Allow others (and yourself) to revise or clarify ideas and positions in light of new information.
- Critique ideas, not people.
- Do not offer opinions without supporting evidence.
- Try not to make assumptions; ask questions instead.
- Avoid put-downs.
- Take responsibility for the quality of the discussion.
- Build on one another's comments; work toward shared understanding.
- Always have your book or readings in front of you.
- Do not monopolise discussion.
- If you are offended by anything said during discussion, acknowledge it immediately.
- If you notice patterns that are troubling or might be impeding full engagement by others, please speak to me (or the T.A., when applicable) in office or via email. Such discussions should be understood as being strictly confidential. If it is not possible to speak to me, feel free to reach out to the department chair, and academic advisor, or a trusted mentor.

Missing or Late Assignments

I will adopt the following policy regarding late assignments: For the (very short) weekly paper, no late assignments will be accepted. The purpose of this exercise is to encourage you to come to class prepared, and having read the assigned reading.

With respect to the final paper, each component, except for the peer feedback, will be penalised by **one full letter grade** per 24 hours past the deadline. (Note that this includes weekends and holidays.) Letter grades are given by A+, A, A-, B+, B, B-, C+, C, C-, D, F.

So, for example, if you wrote an A+ paper, and you hand it in 10 days late, you get an F! If you hand in a B- paper two days late, it gets a C, etc.

That being said, I am willing to grant some leeway (within reason), for **valid** excuses with documentation.

If you have any concerns at any point throughout the course, I encourage you to email me or come to my office hours to discuss. In general, if a special condition or circumstance in your life may affect your performance, please let me know about it as soon as possible. It will be treated with the strictest confidence. *Do not wait until the condition or circumstance is impending or has already happened before telling me about its impact on you.*

If something unanticipated occurs, bring it to my attention and we will work out a way of dealing with it.

Academic Integrity

Any form of academic misconduct that is shown on an assignment, exam, essay, etc. is sufficient for a failing grade on that assignment.

Demonstrable repetition of academic misconduct is sufficient for a failing grade in the course.

Depending on the severity of the misconduct, a letter recording the violation may be sent to the Dean.

[Refer to the university's policies concerning academic integrity.]

Detailed Schedule

Note:
This syllabus may be updated throughout the course with more detailed / accurate descriptions of the later weeks; I will send a notification via email and make an announcement in class whenever the syllabus is changed.

Part I: (Natural) Intelligence and Learning

In the first part of the course, we will primarily be looking at what it means to be intelligent or to be able to learn. We will compare discussions of whether machines are capable of intelligence with definitions and discussions of human and animal intelligence.

W0. **Introduction to Course**

I will go over the syllabus, and provide a general introduction to the theme and problems upon which we will focus in this course.

We will begin by looking at a classic perspective from early in the field on whether machines *can* be intelligent, and how we might measure this. One key thing that you should be thinking about in these first weeks is what *you* think machine might be reasonably capable of. Do you think that this behaviour can be classified as intelligence? Some of the early descriptions of machine intelligence are dated, and may seem somewhat silly; however, by the end of the course, we will re-address our own assumptions about what machines are capable of, and see whether these assumptions hold up in light of state-of-the-art technologies in AI.

Required Reading

Alan M. Turing. (1992/1948) "Intelligent Machinery" in D. C. Ince (ed.) *Mechanical Intelligence: Collected Works of A. M. Turing*

Irving John Good. (1966) "Speculations Concerning the First Ultraintelligent Machine" *Advances in Computers* 6: 31–88.

Optional Reading

Alan M. Turing. (1950) "Computing Machinery and Intelligence" *Mind* 49: 433-460.

W1. **Intelligent Agents**

This week, we will discuss what it means for agents to be intelligent, comparing machine programs with human intelligence.

Required Reading

John R. Searle. (1980) "Minds, Brains, and Programs" *Behavioral and Brain Sciences* 3(3): 417-457.

Herbert A. Simon. (1990) "Invariants of Human Behavior". *Annual Review of Psychology* 41:1-19.

W2. **Intelligent Animals**

These week, we discuss the role of learning in intelligence; in particular, we will look at empirical and philosophical work on animal intelligence.

Required Reading

Ido Erev and Alvin E. Roth. (1998) "Predicting How People Play Games: Reinforcement Learning in Experimental Games with Unique Mixed Strategy Equilibria". *The American Economic Review* 88(4): 848-881.

Selmer Bringsjord, Clarke Caporale and Ron Noe. (2000) "Animals, Zombanimals, and the Total Turing Test: The Essence of Artificial Intelligence" *Journal of Logic, Language, and Information* Special Issue on Alan Turing and Artificial Intelligence 9(4): 397–418.

Optional Reading

Edward L. Thorndike. (1905) "The Law of Association" Chapter XIII in *The Elements of Psychology*. New York: A. G. Seiler.

Edward L. Thorndike. (1905) "The Law of Dissociation or Analysis" Chapter XIV in *The Elements of Psychology*. New York: A. G. Seiler.

W3. **Contemporary Perspectives**

Many of the readings that we have discussed at this point have been historical notions of intelligence, spanning the 20th Century. This week, we conclude the first part of the course by discussing a more contemporary notion of what machine intelligence consists in, comparing it to more contemporary notions of animal and human intelligence.

Required Reading

Stuart Russell. (2019) "Intelligence in Humans and Machines" Chapter 2 in *Human Compatible: Artificial Intelligence and the Problem of Control*.

Nicolas Gauvrit, Hector Zenil, and Jesper Tegnér. (2017) "The Information-Theoretic and Algorithmic Approach to Human, Animal, and Artificial Cognition" in Gordana Dodig-Crnkovic and Raffaela Giovagnoli (eds.) *Representation and Reality in Humans, Other Living Organisms and Intelligent Machines* Studies in Applied Philosophy, Epistemology and Rational Ethics (SAPERE) Vol. 28. Cham, Switzerland: Springer Nature. 117–140.

Part II: (Artificial) Intelligence and (Machine) Learning

In the second part of the course, we will get an overview of the fundamentals of machine learning and contemporary artificial intelligence. We will begin with the history of machine learning and AI before discussing supervised learning, unsupervised learning, neural networks, reinforcement learning, and deep learning. Because of the maths-heavy nature of many of these readings, I will outline additional resources that may be helpful. As noted at the start of the syllabus, it is not necessary to understand all of the formalism, but it is important to understand the concepts underlying machine learning techniques, and how they essentially work.

W4. **History and Foundations of Artificial Intelligence**

In order to understand where we are currently in artificial intelligence research, it is important to understand (at least a small part of) the history of advancements in AI. The first required reading this week focuses upon history in terms of the sort of philosophical questions in which we are interested, whereas the second is a more technical description of the history of AI from a computer-science perspective.

Required Reading

Mariusz Flasiński (2016) "History of Artificial Intelligence" Chapter 1 in *Introduction to Artificial Intelligence*.

Stewart Russell and Peter Norvig. (2020) "Introduction" Chapter 1 in *Artificial Intelligence: A Modern Approach*.

Optional Reading

Paul Thagard. (1990) "Philosophy and Machine Learning". *Canadian Journal of Philosophy* 20(2): 261–276.

Vishal Maini. (2017) "Why machine learning matters". *Medium*. https://medium.com/machine-learning-for-humans/

W5. **Neural Networks and Machine Learning**

In the first genuinely technical week, we will discuss the distinctions between different types of machine learning programs—particularly, supervised learning and unsupervised learning—and what applications they have. The idea is to get a sense of how contemporary machine learning programs work and what they do.

Required Reading

Michael A. Nielsen. (2015) "Using Neural Networks to Recognize Handwritten Digits". Chapter 1 in *Neural Networks and Deep Learning*, Determination Press.

Michael A. Nielsen. (2015) "How the backpropagation algorithm works". Chapter 1 in *Neural Networks and Deep Learning*, Determination Press.

Additional Resources:

Grant Sanderson (3Blue1Brown) has an excellent series of four videos on YouTube describing what neural networks are and how they essentially work. This provides a nice visual explanation for neural networks and the fundamentals of many machine-learning algorithms.

Series 3, Episode 1. "But What is a Neural Network" *Deep Learning, Ch. 1.* https://youtu.be/aircAruvnKk.

Series 3, Episode 2. "Gradient Descent, How Neural Networks Learn" *Deep Learning, Ch. 2.* https://youtu.be/IHZwWFHWa-w.

Series 3, Episode 3. "What is Backpropagation Really Doing?" *Deep Learning, Ch. 3.* https://youtu.be/Ilg3gGewQ5U.

Series 3, Episode 4. "Backpropagation Calculus" *Deep Learning, Ch. 4.* https://youtu.be/tIeHLnjs5U8.

Additionally, Noah Yonackl has an excellent non-technical introduction to machine learning that is posted on *Medium* (via SafeGraph). This should be consulted at *some* point during Part II of this course.

Noah Yonack (2017) "A Non-Tehcnical Introduction to Machine Learning" *SafeGraph (Medium).* https://blog.safegraph.com/.

See also Vishal Maini's explanations in the series "Machine Learning for Humans".

Vishal Maini. (2017) "Machine Learning for Humans, Part 2.1: Supervised Learning". *Medium.* https://medium.com/machine-learning-for-humans/

Vishal Maini. (2017) "Machine Learning for Humans, Part 2.2: Supervised Learning II". *Medium.* https://medium.com/machine-learning-for-humans/

Vishal Maini. (2017) "Machine Learning for Humans, Part 2.2: Supervised

Learning III". *Medium*.
https://medium.com/machine-learning-for-humans/

Vishal Maini. (2017) "Machine Learning for Humans, Part 2.2: Supervised Learning III". *Medium*.
https://medium.com/machine-learning-for-humans/

W6. **Reinforcement Learning**

This week, we will extend our knowledge of the repertoire of machine learning programs from (un)supervised learning to reinforcement learning—much work in this field is highly exploratory. We will discuss bandit problems as a basis of understanding the practical trade-off between 'exploration' and 'exploitation', as well as delayed rewards, environments, and the general architecture of Markov decision processes.

Required Reading

Richard S. Sutton and Andrew G. Barto. (2017) "Introduction" Chapter 1 in *Reinforcement Learning: An Introduction*. 2nd Ed. 1–17.

Richard S. Sutton and Andrew G. Barto. (2017) "Multi-Armed Bandits" Chapter 2 in *Reinforcement Learning: An Introduction*. 2nd Ed. 18–36.

Optional Reading

Richard S. Sutton and Andrew G. Barto. (2017) "Finite Markov Decision Processes" Chapter 3 in *Reinforcement Learning: An Introduction*. 2nd Ed. 1–17.

Christopher John Cornish Hellaby Watkins. (1989) Chapters 1 and 2 of "Learning from Delayed Rewards" (PhD Thesis).

Additional Resources

Vishal Maini. (2017) "Machine Learning for Humans, Part 5: Reinforcement Learning". *Medium*.
https://medium.com/machine-learning-for-humans/

W7. **Deep Learning**

In the final week of our technical background in contemporary artificial intelligence, we come to the current state-of-the-art in machine learning: so-called *deep* learning.

Required Reading

Cameron Buckner. (2019) "Deep Learning: A Philosophical Introduction" *Philosophy Compass* 14(10): e12625.

Additional Resources

Vishal Maini. (2017) "Machine Learning for Humans, Part 4: Neural Networks & Deep Learning". *Medium.* https://medium.com/machine-learning-for-humans/

Part III: Ethical, Philosophical, and Societal Implications of AI

Having a conceptual understanding of what intelligence and learning consist in, and having a technical understanding of what artificial intelligence and machine learning actually are (and what they can actually do), we are now in a position to begin discussing the philosophical and ethical implications of artificial intelligence. In the final weeks, we will look at a variety of technical, conceptual, ethical, societal, and legal issues surrounding the implementation of artificial intelligence programs.

W8. **Technical Aspects of AI Safety**

One significant problem with deep-learning in general is the ubiquity of so-called 'black-box' algorithms. In many cases, as we have seen, deep-learning techniques work significantly better than classic techniques in AI. However, in many cases it is not understood *what* the algorithms are actually doing, or *how* they are learning. Measures of success in ML are often given by surpassing benchmarks, which implies that it is the *ends* that matter rather than the *means*. Hence, from an industry standpoint it matters not how a program achieves better performance, but rather *that* it does so. However, this raises the question *How can we ensure that an AI is safe?* We begin by looking at some technical issues, such as *reward hacking*.

Required Reading

Stuart Russell. (2019) "If We Succeed" Chapter 1 in *Human Compatible: Artificial Intelligence and the Problem of Control.*

Stephen M. Omohundro. (2008) "The Basic AI Drives". *Proceedings of the 2008 conference on Artificial General Intelligence.* Amsterdam: IOS Press Amsterdam. 483-492.

Dario Amodei and Chris Olah and Jacob Steinhardt and Paul Christiano and John Schulman and Dan Mané. (2016) Concrete Problems in AI Safety. *arXiv preprint.* https://arxiv.org/pdf/1606.06565.pdf.

W9. **The Alignment Problem**

In the second part of our foray into technical AI safety, we examine the "alignment problem" (sometimes called the "control problem". The main question concerns how we might ensure that an AI agent's *goals* are aligned with our own, or with the goals of a common human good. We will see an analogy with economics in terms of so-called "incomplete contracts". We will further discuss the distinction between AI agents and human agents—namely, why misalignment of interests is so concerning with respect to AI, though it is evidently ubiquitous in interactions with humans.

Required Reading

Nick Bostromn (2017/2014) "The Control Problem" Chapter 9 in *Superintelligence: Paths, Dangers, Strategies* Oxford: Oxford University Press.

Dylan Hadfield-Menell and Gillian Hadfield. (2018) "Incomplete Contracting and AI Alignment". *arXiv pre-print* https://arxiv.org/abs/1804.04268.

W10. **Fairness and Bias**

This week, we will examine problems of fairness and bias with respect to machine-learning algorithms (and whether or not it is possible to avoid bias or discrimination completely).

Required Reading

Sorelle A. Friedler, Carlos Scheidegger, and Suresh Venkatasubramanian. (2016) "On the (im)possibility of fairness. *arXiv pre-print.* https://arxiv.org/abs/1609.07236 .

Reuben Binns. (2018) "Fairness in Machine Learning: Lessons from Political Philosophy". *Journal of Machine Learning Research* 81: 1–11. https://arxiv.org/pdf/1712.03586.pdf

Gabrielle M. Johnson. "Algorithmic Bias: on the implicit biases of social technology" (Unpublished manuscript, pending approval from the author)

Optional Reading

Chris Russell, Matt J. Kusner, Joshua R. Loftus, and Ricardo Silva. (2017) "When Worlds Collide: Integrating Different Counterfactual Assumptions in Fairness" *31st Conference on Neural Information Processing Systems (NeurIPS 2017).*

Andrew D. Selbst, Danah Boyd, Sorelle A. Friedler, Suresh Venkatasubramanian, and Janet Vertesi. (2019) "Fairness and Abstraction in Sociotechnical Systems" *ACM Conference on Fairness, Accountability, and Transparency* 1(1).

W11. **Legal and Societal Impact**

This week, we discuss broader social impacts of the implementation of artificial intelligence. We will examine issues such as human rights with regard to universal basic income (as AI displaces large swaths of the workforce), and potential concerns surrounding misuses of AI technology (such as surveillance, warfare, fake news media, etc.) We will also briefly discuss legal frameworks for dealing with these issues and touch upon the legal and moral standing of artificial agents.

Required Reading

Gillian Hadfield. (2019) "New Introduction" *Rules for a Flat World: Why Humans Invented Law and How to Reinvent It for a Complex Global Economy.* 2nd ed. Oxford: Oxford University Press.

Stuart Russell. (2019) "The Misuses of AI" Chapter 4 in *Human Compatible: Artificial Intelligence and the Problem of Control.*

Matthias Rolf, Nigel Crook, and J. J. Steil. (2018) "From social interaction to ethical AI: A developmental roadmap." *IEEE conference: Development and learning and epigenetic robotics.*

Optional Readings

Mathias Risse. (2019) "Human Rights and Artificial Intelligence: An Urgently Needed Agenda". *Human Rights Quarterly* 41(1): 1–16.

Neil M. Richards and William D. Smart (2013) "How Should the Law Think About Robots?" *SSRN Electronic Journal.*

W12. **Concluding (Further Philosophical Questions)**

In the final meeting, we will conclude on a (somewhat) high note, and discuss some of the *positive* implications of artificial intelligence in society. We will come back to the initial questions we discussed at the start of the semester, about what it means to be human, what human intelligence consists in, and how (or whether) machine (or animal) intelligence differs significantly from this. In particular, we will discuss the *integration* of AI in terms of human cognitive capacities.

Required Reading

Mariarosaria Taddeo and Luciano Floridi. (2018) "How AI can be a force for good". *Science* 361(6404): 751-752.

José Hernández-Orallo and Karina Vold. (2019) "AI Extenders: The Ethical and Societal Implications of Humans Cognitively Extended by AI". Proceedings of the the 2019 AAAI/ACM Conference.

Two Papers from your peers.

Disclaimer

This document is meant to be binding; however, in the event of circumstances beyond my control, the course contents, evaluation scheme and other parts of this syllabus are subject to change.