Intelligence, Natural and Artificial:
Philosophical Implications of AI

(ABRIDGED) COURSE OUTLINE

Detailed Course Description.

'What is AI such that it may be called 'intelligent', and what is intelligence such that it may be artificial?'

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? How are they like/unlike the human brain?
- 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?

> ***NOTE***: 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.

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.

Grading Details

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.

In addition to these three assignment components, timely arrival, attendance, and engagement in class will count toward the participation mark. It is expected that your contributions will be respectful and constructive.

Bonus Marks. There will be two opportunities for obtaining bonus marks in this course. One bonus mark will be awarded for a perfect score on a quiz on the content of this syllabus (completed on the first day of class). A second bonus mark will be awarded to everyone registered just in case a quorum (at least 2/3) of students completes the year-end course evaluations. More details will be given in class.

Grading Overview. The grading for this course is broken down as follows:

Participation	10 %
PART I: Short Writing Assignments	15 %
PART 2: Program Design (group project)	25 %
PART 3: Final Paper (see below for breakdown)	50 %
Bonus Marks	2 %
TOTAL	**100 %**

The final paper has several components, and the grade for the final paper will be broken down as follows:

Proposal (500 words)	5 %
First Draft (2-3K words)	15 %
Peer Feedback (in Class)	5 %
Abstract of the Final Paper (<250 words)	5 %
Revised final Draft (<6K words)	20 %
TOTAL	**50 %**

Detailed Course Schedule

PART I: (Natural) Intelligence and Learning

Week 1 **Course Introduction**

Topic Overview

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.

Req. Reading

Syllabus

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.

Alan M. Turing. 1950. ‘Computing Machinery and Intelligence’. *Mind* 49: 433-460.

PART I: Classical Game Theory and the Lewis Signalling Game

Week 2 **Intelligent Agents**

Topic Overview

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

Req. 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.

Week 3 **Intelligent Animals**

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

Req. 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.

Opt. Reading Edward L. Thorndike. 1905. 'The Law of Association', Ch. XIII in *The Elements of Psychology*. New York: A. G. Seiler.

Edward L. Thorndike. 1905. 'The Law of Dissociation or Analysis', Ch. XIV in *The Elements of Psychology*. New York: A. G. Seiler.

Week 4 **Contemporary Perspectives**

Topic Overview 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.

Req. Reading Stuart Russell. 2019. 'Intelligence in Humans and Machines', Ch. 2 in *Human Compatible: Artificial Intelligence and the Problem of Control*. 13-61.

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 help- ful. 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.

Week 5 History and Foundations of Artificial Intelligence

Topic Overview

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.

Req. Reading

Mariusz Flasinski. 2016. ‘History of Artificial Intelligence’, Ch, 1 in *Introduction to Artificial Intelligence*.

Stewart Russell and Peter Norvig. 2020. ‘Introduction’, Ch. 1 in *Artificial Intelligence: A Modern Approach*.

Opt. 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/

Week 6 Neural Networks and Machine Learning

Topic Overview

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.

Req. Reading

Michael A. Nielsen. 2015. ‘Using Neural Networks to Recognize Handwritten Digits’, Ch. 1 in *Neural Networks and Deep Learning*, Determination Press. http://neuralnetworksanddeeplearning.com

Michael A. Nielsen. 2015. ‘How the backpropagation algorithm works’, Ch. 2 in *Neural Networks and Deep Learning*, Determination Press. http://neuralnetworksanddeeplearning.com

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-Technical Introduction to Machine Learning’ SafeGraph (Medium). https://blog.safegraph.com/.

See also Vishal Maini’s explanations in the series “Machine Learning for Humans”. https://medium.com/machine-learning-for-humans/.

> Vishal Maini. 2017. ‘Machine Learning for Humans, Part 2.1: Supervised Learning’.
>
> Vishal Maini. 2017. ‘Machine Learning for Humans, Part 2.2: Supervised Learning II’.
>
> Vishal Maini. 2017. ‘Machine Learning for Humans, Part 2.2: Supervised Learning III’.
>
> Vishal Maini. 2017. ‘Machine Learning for Humans, Part 2.2: Supervised Learning III’.

Week 7 **Reinforcement Learning**

Topic Overview 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.

Req. Reading Richard S. Sutton and Andrew G. Barto. 2017. ‘Introduction’, Ch. 1 in *Reinforcement Learning: An Introduction*. 2nd Ed. 1-17.

Richard S. Sutton and Andrew G. Barto. 2017. ‘Multi-Armed Bandits’, Ch. 2 in *Reinforcement Learning: An Introduction*. 2nd Ed. 18-36.

Opt. Reading Richard S. Sutton and Andrew G. Barto. 2017. ‘Finite Markov Decision Processes’ Ch. 3 in *Reinforcement Learning: An Introduction*. 2nd Ed. 1-17.

Christopher John Cornish Hellaby Watkins. 1989. Ch. 1 and Ch. 2 of ‘Learning from Delayed Rewards’, *PhD Thesis*. King’s College, University of Cambridge.

Week 8 **Deep Learning**

Topic Overview 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*.

Req. 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, Social, and Philosophical 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.

Week 9 Technical Approaches to AI Safety

Topic Overview 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.

Req. Reading Stuart Russell. 2019. 'If We Succeed', Ch. 1 in *Human Compatible: Artificial Intelligence and the Problem of Control*. New York: Viking. 1-12.

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, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. 2016. 'Concrete Problems in AI Safety'. *arXiv preprint* 1606.10565: 1-29. https://arxiv.org/pdf/1606.06565.pdf

.

Week 10 The Alignment Problem

Topic Overview In the second part of our foray into technical AI safety, we examine the 'alignment 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.

Req. Reading Nick Bostrom. 2017/2014. 'The Control Problem', Ch. 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* 1804.04268: https://arxiv.org/abs/1804.04268.

Week 11 **Fairness and Bias**

Topic Overview

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).

Req. 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.

Gabbrielle M. Johnson. 2020. ‘Algorithmic bias: on the implicit biases of social’, Forthcoming in *Synthese*. 1-21.

Opt. 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).

Week 12 **Legal and Societal Impacts**

Topic Overview

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.

Req. 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. ‘Misuses of AI’, Ch. 4 in *Human Compatible: Artificial Intelligence and the Problem of Control*. New York: Viking. 103-131.

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*.

Opt. Reading

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*.

Week 13 **Concluding (and Further Philosophical Questions)**

Topic Overview 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.

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

José Hernéndez-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.