Philosophy and AI
Lecture 1: Introduction

Marco Degano

4 February 2025

Outline

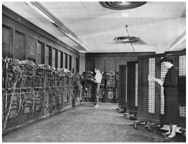
1. General Information About the Course
2. What is Philosophy?
3. What is AI?
4. Different Domains of Philosophy

Course Team

- **Lecturer:**
  Marco Degano `m.degano@uva.nl`
- **Senior TAs:**
  Frank Wildenburg `f.c.l.wildenburg@uva.nl`
  Emmeke Veltmeijer `e.a.veltmeijer@uva.nl`
- **TAs:**
  Storm Hartkamp, Kirti Singh, Guido van der Knaap, Jesse Wonnink
- The emails of the TAs are available on Canvas under the 'Contact' page.

Aim of the course

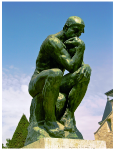
AI and computing technology is progressing at an astounding pace.



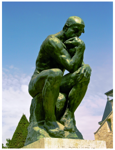
ENIAC digital computer (1945)

NVIDIA H100 Tensor Core GPU

The fundamental questions of philosophy remain constant (with novel challenges posed by new technologies, like AI)



The Thinker, Auguste Rodin



The Thinker, Auguste Rodin

Aim of the course

Reflecting on general goals, uses and ethical challenges of AI is today of fundamental importance.

Content:

- the most important philosophical concepts and tools
- their applications to foundational questions of AI.

Skills:

- write well (clear, concise, analytic, soundly argued)
- critically reflect on foundational issues of AI.

Course Format

- 12 Lectures. Lectures will be recorded and made available via Canvas. Attendance to the lecture is encouraged.
- 6 Tutorials: Attendance to tutorials is **mandatory**. You are allowed to skip only 1 of the tutorials without providing a reason. Please see the group assignment on Datanose or Canvas.

Assessment:

- 2 Homework assignments in the first half of the course, submission before Tutorial 2 and 3 (10%)
- 2 Exams: Midterm (30%) + Final Exam (30%)
- Final Essay (30%)

Please see document **General Information** on Canvas.

Schedule and Course Content

Preliminary schedule of the course:

- **Week 1:** Introduction and Epistemology
- **Week 2:** Epistemology and Philosophy of Science
- **Week 3:** Philosophy of Mind
- **Week 4:** Midterm Examination
- **Week 5:** Ethics
- **Week 6:** Philosophy of Language
- **Week 7:** Selected Topics
- **Week 8:** Final Examination

Course Materials

- Clark Glymour, *Thinking Things Through: An Introduction to Philosophical Issues and Achievements*
- These slides and class content.
- A collection of articles

The list of required readings is available on Canvas. Readings will also be listed on each lecture's slides. Additional optional readings and supporting resources may also be provided. Optional materials will not be included in any examinations.

A note on these slides

- The slides will be primarily based on the textbook (Glymour 2015) and assigned readings.
- But they may also contain additional content and concepts that you are expected to study.
- The slides are aimed at being mostly self-contained (to help if you miss a lecture or when reviewing the material), prioritizing completeness over presentation style.

Today's Readings

Required: none
Optional:

- Arkoudas, Konstantine, and Selmer Bringsjord. "Philosophical Foundations." *The Cambridge Handbook of Artificial Intelligence*. Ed. Keith Frankish and William M. Ramsey. Cambridge: Cambridge University Press, 2014. pp. 34–63.

Outline

1. General Information About the Course
2. What is Philosophy?
3. What is AI?
4. Different Domains of Philosophy

What is Philosophy?

- Philosophy is the study of fundamental questions such as:
  - What is knowledge, and how do we acquire it?
  - Is there free will, or is the world deterministic?
  - What constitutes valid reasoning? What is rationality?
  - What is the relationship between language and the world?
  - How should I live?
  - . . . and countless other questions.

Philosophy and Science

Philosophy provides the opportunity for reflection on a science:

- Simply put: philosophy of X means studying and questioning the methods, assumptions, concepts, etc., of the science X.
- Thus, it asks fundamental questions that are important to practitioners of X but (strictly speaking) are not part of X.
- Many philosophical domains can be useful in answering such questions (not just philosophy of science!).

Are Philosophical Questions Scientific Questions? (Glymour 2015, chapter 1)

Consider the following examples:

- How can we know there are particles too small to observe?
- How can anyone know there are other minds?
- What constitutes a scientific explanation?
- How do we know that the process of science leads to the truth, whatever the truth may be?
- What is meant by “truth”?
- Does what is true depend on what is believed?

Are Philosophical Questions Scientific Questions?

- What facts determine whether a person at one moment of time is the same person as a person at another moment of time?
- What are the limits of knowledge?
- How can anyone know whether she is following a rule?
- What is a proof?
- What does “impossible” mean?
- What is required for beliefs to be rational?
- What is the best way to conduct inquiry?
- What is a computation?
- How should people behave?
- How should social and political institutions be organized and ruled?

Are Philosophical Questions Scientific Questions?

- They are *somewhat* related to science but seem too fundamental to belong to science.
- They are also important: the answers influence how we think about science and how we practice science.
- Philosophy, despite all its limitations, has indeed made progress on some of these questions.

Philosophy and AI

- Many of the major problems in AI are actually old philosophical questions:
  - How do we go from input (human: senses, AI: data) to knowledge about the world? Is the input sufficient, or are further assumptions necessary (think of inductive bias)?
  - How do we go from probabilistic information to knowledge/decisions?
- Answers from philosophy are often relevant and useful for AI.
- But also vice versa: AI allows philosophical questions to have practical significance and to be tested.
- Perhaps AI researchers are the philosophers of today...

Outline

1. General Information About the Course
2. What is Philosophy?
3. What is AI?
4. Different Domains of Philosophy

Defining AI

You are likely already familiar with the history and basic distinctions of AI. But let's set the basic terminology we will employ.

What is *artificial intelligence*?

- system that *thinks* like humans?
- system that *acts* like humans?
- system that *functions* like humans?
- system that thinks/acts/functions rationally (Hobbes: "ratiocination is computation")
- ...

Weak vs Strong AI

- **Weak/Narrow AI:** Machines that simulate intelligence to perform tasks but do not possess true understanding.
- **Strong/General AI:** Machines with a mind that demonstrate genuine cognition and self-awareness.
- Notions like *mind*, *intelligence*, and *self-awareness* should be critically examined. We will revisit these concepts throughout the course.

Can Machines Think?

- What would convince us that a machine can truly think and demonstrate understanding?
- In 2022, the Cognitive Science Society held a contest for 5-minute videos addressing the question *Can Machines Think?*:
  1. Ariadne Letrou and Asher Liftin: Watch here
  2. Ryan Rhodes: Watch here
  3. Kristyn Sommer and William Bingley: Watch here
- These videos, though insightful, highlight how quickly AI research evolves-perspectives from 2022 already seem outdated in 2025.

Defining AI

What is the aim of AI?

- **Technological Aim:** Building systems that exhibit intelligent behavior.
- **Scientific Aim:** Understanding the principles of intelligence.

How is AI implemented?

- **Virtual:** Software-based agents (e.g., chatbots).
- **Physical:** Embodied systems (e.g., robots).

Symbolic vs Subsymbolic

Symbolic AI:

- Utilizes explicit rules and logic to represent knowledge.
- Examples: Expert systems (e.g., MYCIN, DENDRAL), theorem provers, rule-based chatbots[1].
- Strengths: Well-suited for tasks requiring clear reasoning and structured problem-solving.
- Limitations: Struggles with ambiguity and learning from unstructured data.

Subsymbolic AI:

- Based on statistical models and large-scale data-driven learning.
- Strengths: Excels at pattern recognition and learning from data.
- Limitations: Often considered a ‘black box’ due to lack of interpretability.

[1] ELIZA, one of the early rule-based chatbots, recently ‘reanimated’ in December 2024: `https://sites.google.com/view/elizagen-org/blog/eliza-reanimated`

Outline

1. General Information About the Course

2. What is Philosophy?

3. What is AI?

4. Different Domains of Philosophy

Domains within Philosophy and Their Questions

- Philosophy is divided into subdomains, each addressing key questions about existence, knowledge, ethics, reasoning, ... This is the approach we will mostly take in this course.
- One could also approach it differently.
- For instance, historically (e.g., Aristotle → Descartes → Kant → AI).
- Or by method (continental/analytical/formal/experimental).

Epistemology (Theory of Knowledge) I

- What is knowledge? Is it justified true belief?
- How do we acquire knowledge? Through perception, reasoning, or experience?
- What is the difference between a priori (knowledge independent of experience) and a posteriori (knowledge based on experience)?
- How do we deal with errors, illusions or biases?

Epistemology (Theory of Knowledge) II

Connections to AI: Symbolic AI

- **Knowledge Representation and Reasoning:** How can we model knowledge explicitly in a way machines can process?
- **Epistemic Logic:** Understanding how AI systems can reason about what they (and others) know.
- **AI verification:** Ensuring systems behave consistently with encoded knowledge.

Dennett 1979 (‘Artificial Intelligence as Philosophy and as Psychology’): *AI shares with epistemology the abstract question: how is knowledge possible?*

Epistemology (Theory of Knowledge) III

Connections to AI: Sub-symbolic AI

- **Deep Neural Networks (DNNs):**
  - How do DNNs “know” or “learn”?
  - Do they truly “understand” or do they just approximate patterns in data?
- **Key Challenges:**
  - **Domain Mismatch:**
    - Models trained on data distribution X might fail on distribution Y.
    - *Example:* Facial recognition trained on specific demographics but failing on others.
  - **Adversarial Attacks:**
    - Can we trust DNNs when small changes to input (e.g., adding noise) can cause large errors?
  - **Explainability (XAI):**
    - Should we trust models without understanding how they make decisions?
    - What kinds of explanations (mathematical, causal, intuitive) are useful for humans?

Philosophy of Science

- What makes a scientific explanation valid?
- How do we verify or falsify theories?
- What is the role of causality in science?

Connections to AI:

- **Explainable AI:** What constitutes a good explanation for AI predictions?
- **Scientific Method in ML:** How do we design algorithms to explore hypotheses (explore vs exploit in RL)? Inductive biases in models: What assumptions guide learning?
- **Causal Inference in AI**: How can we go beyond correlations to uncover true causal relationships?

Philosophy of Mind

- What is the nature of the mind?
- Is the mind reducible to physical processes, or is it distinct?
- What is consciousness, and can it be replicated in machines?
- Is language necessary for thought?
- Are mental states computations?

Connections to AI:

- **Strong AI vs Weak AI:** Can machines truly replicate the human mind generally (Strong AI)? Or are they just tools for specific tasks (Weak AI)?
- **Embodied AI:** Intelligence grounded in physical interactions (e.g., robotics).
- **Neuroscience-Inspired AI:** Neural networks inspired by the structure of the human brain.
- **Language and Thought:** Role of large language models (e.g., GPT) in replicating aspects of human cognition.

Ethics

- What is the nature of morality?
- How can we distinguish good actions from bad ones?
- What is responsibility?
- Are ethical principles universal or relative?

Connections to AI:

- **Value Alignment Problem:** ensuring AI systems align with human ethical principles (e.g., trolley problem in self-driving cars)
- **Bias in AI:** How do we address biases in datasets and algorithms?
- **Applications of AI and Ethics:** Privacy (e.g., surveillance systems); Autonomous weapons; Mental health impacts of social media algorithms.

Metaphysics

- What is the fundamental nature of reality?
- What exists (numbers, time, possibilities)?
- Is time an illusion? What about causality?

Connections to AI:

- **Simulation Hypothesis:** Are we living in a simulation, and can AI help detect it?
- **Representation Learning:** Can AI find a universal representation of the world, or is all representation anthropocentric?
- **Ontology in AI:** How do we encode 'what exists' in a machine-readable way?

Aesthetics

- What is beauty? Is it subjective or objective?
- What makes something a work of art?
- How do we interpret and appreciate artistic expressions?

Connections to AI:

- **Generative AI and Creativity:** AI-generated art, music, and poetry. Can machine-created art evoke the same emotions as human-made art? Can AI ever truly replicate the *experience* of creating art?
- **Subjectivity of Aesthetics:** How do we formalize subjective judgments (beauty, taste) for AI systems?
- **AI as a Collaborative Tool:** AI assisting human creativity (e.g., suggesting designs, generating ideas). Are AI tools collaborators or just advanced instruments?
- **Ethics in Aesthetics:** Plagiarism and originality and intellectual property? Is AI art a reflection of the data it was trained on, or something fundamentally new, creative?

Philosophy of Language

- How does language refer to the world?
- What is meaning, and how is it conveyed?
- What is truth?
- Is there an ideal or universal language?

Connections to AI:

- **NLP:** How do we teach machines to understand and generate human language?
- **Grounding:** How do AI systems link abstract symbols to real-world concepts?
- **Language Models and Reality:** Challenges in AI understanding nuance, context, and ambiguity.

Logic

- What is valid reasoning?
- Can logic serve as the foundation for mathematics, science, and language?
- What are the limits of computability and provability?

Connections to AI:

- Symbolic AI (e.g., Prolog programming for logical inference)
- Algorithms and Formal Languages
- What problems can AI solve, and what lies beyond its reach?

Some Remarks

This division is useful as a rough outline, but not more than that:

- Not exhaustive: There are more domains, and you can also divide the domains differently.
- Not exclusive: Philosophical questions and debates often span across several fields.
- For example, what knowledge is (epistemology) is important to what a good scientific explanation is (philosophy of science): viz., one that produces knowledge.
- Or questions like: Do the distinctions ‘a priori vs a posteriori’ (epistemology), ‘analytic vs synthetic’ (philosophy of language), and ‘necessary vs contingent’ (metaphysics) all coincide?

Content of the course

- For reasons of time, we cannot cover all subfields. So we cover the ones most relevant to AI:
- The first half (weeks 1–3): epistemology, philosophy science and philosophy of mind. (focus on history and theories of philosophy)
- The second half (weeks 5–7): ethics, philosophy of language and selected topics. (Focus on applications and current methods in AI.)

How philosophy is different I

- Warning: When coming from an ‘exact’ science (AI, computer science, mathematics, physics, engineering, etc.) it is easy to dismiss philosophical questions as pointless:
    - There are no concrete results, unlike programs, theorems, experiments.
    - There are no ‘difficult looking things’: complicated code, long proofs, elaborate labs, large datasets.
- Rather: The difficulty of philosophy is to take a messy, elusive, but important concept and try to understand more and more aspects of it precisely until one arrives at a satisfying formal definition of the concept. (e.g., the notion of ‘computation’.)

How philosophy is different II

- Slogan: While formal concepts are the starting point of exact sciences, they are the ending point of philosophy.
- This ‘conceptual work’ is a very different kind of difficulty than the difficulties found in exact sciences.
- But it cannot be dismissed as simple or pointless: without it many of the present exact science wouldn’t exist (again, think, e.g., of computation).

How philosophy is different III

- So tackling philosophical problems requires a very different methodology:
- For example, careful argumentation in natural language, thought experiments, analysis of intuition, etc.
- But it doesn't exclude the use of formal tools: a characteristic method of modern analytic (or scientific or mathematical) philosophy is
  1. propose a formalization of the given intuitive concept,
  2. discuss which aspects of it are captured, which not
  3. based on that fine-tune the formalization,
  4. start the circle again.

How philosophy is different IV

- Often one is tempted to say: ‘where is the problem, why don’t you simply say . . . ?’
- But usually one then can appreciate that there is problem after all: often found by keeping on asking ‘why?’
- Of course, one can (and will) still like some philosophical questions more than others. And one also can work on the concrete problem of the exact science, while putting the philosophical issues aside.
- But the course aims to showcase the benefits of taking serious these foundational questions.
- In short: Just keep an open mind, to appreciate the different kind of difficulty posed by philosophy.

Exercises

1. Think of some big philosophical questions. To which subfield would you associate them?
2. Can you think of more examples of links between AI and the big questions of the various subfields of philosophy?