



Lecture 04 September 2025

Philosophical fundamentals of AI

Philip J. Nickel

Part 1

Practitioner foundations

Two kinds of reasoning

Trustworthy AI

Part 2

Philosophical foundations for
intelligent machines

Conditions on intelligence

Conditions on personhood

Part I: Practitioner foundations (Chapter 1 of CHLEP)

AI = (designed) learning + reasoning machines

Learning = getting better at a task through experience

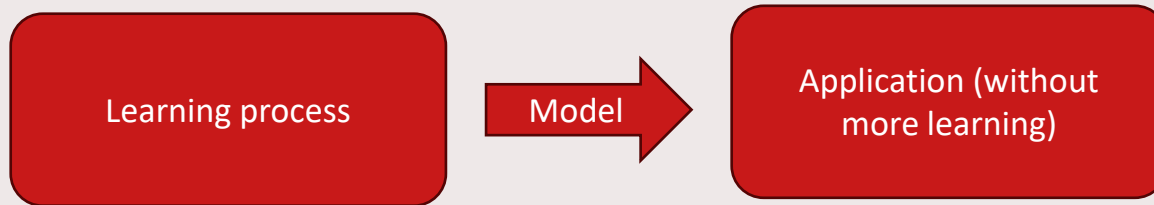
- Requires a metric that measures improvement (better/ worse)
- Is always relative to a specific task

Reasoning = logical deduction or probabilistic inference



Learning from experience

But a truly intelligent being must be able to improve *from any experience*, not merely apply a process that *resulted* from learning in the past.



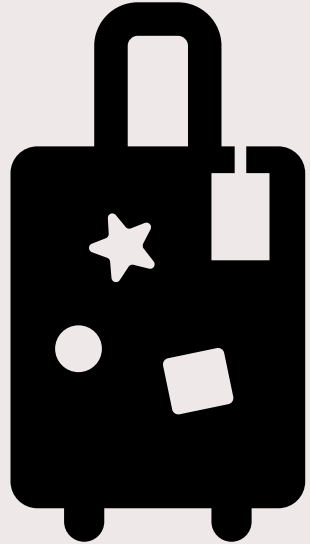


Reasoning as logical deduction

Suppose I'm chatting with an AI chatbot about travel plans. The chatbot suggests that I go to the Great Barrier Reef. After that, I say the following:

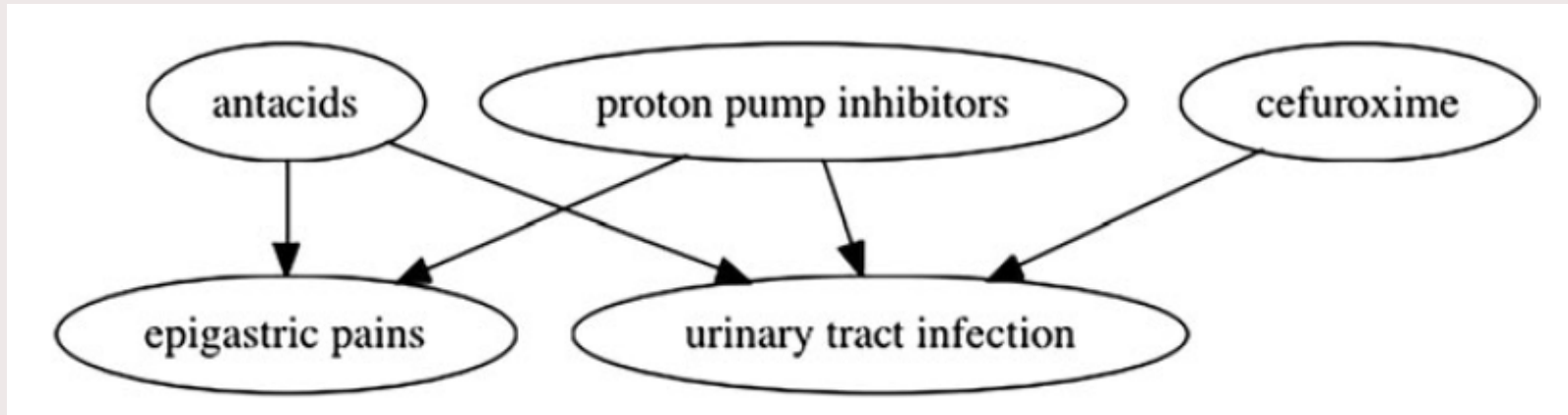
"I'll be traveling with my mother, who is 83. I don't think it's safe for people over the age of 80 to go diving."

Basic reasoning skills in conversation would lead a human speaker to make a *deduction* here, to a conclusion that will guide the conversation.





Probabilistic inference



Explainability and Trustworthiness

The report “White paper on AI: A European approach to excellence and trust” strongly influenced EU’s AI Act and other policy on AI.

This report (linkable from your reading) identifies trustworthy AI as:

- The goal of European AI policy; and,
- Necessarily connected with robustness, legality, and ethical standards.

The ethical standards are the **Four Principles Plus One** that we identified in the lecture last week. “Plus One” refers to explainability.

Robust AI

Robustness is the property of an AI model, such that its sensitivity/specificity will continue to obtain for new data that is similar (“weak robustness”), or for new data that is different (“strong robustness”).

For example, suppose a model is trained to identify a given kind of cancer from an image (e.g., of adult scans).

- Will it show the same accuracy for new images of the same kind, from the clinical setting?
- Will it show the same (or sufficient) accuracy for images that are known to be different (e.g, of children)?
- How easy is it to create adversarial images or queries?

Interlude: Reasoning and arguments

- An **argument** is a set of propositions containing at least one premise and one conclusion. Premises are intended to support the conclusion (i.e., to show that the conclusion is true).
- 1) Either I am talking too fast, or it's likely I won't get through all my slides during the lecture.
 - 2) I am not talking too fast.
 - 3) Therefore it's likely I won't get through all my slides during the lecture.

Deductive Arguments and Validity

- **Validity** is a desired feature of a deductive argument. Validity is that property of arguments such that **if the premises are true, then the conclusion is guaranteed to be true.**
 - 1) All toads are mammals.
 - 2) Socrates was a toad.
 - 3) Therefore, Socrates was a mammal.

Deductive Arguments and Soundness

- A second essential feature of an argument is that its **premises are true**.
 - When an argument is valid and has all true premises, then we call it **sound**.
- 1) All humans are mammals.
 - 2) Socrates was a human.
 - 3) Therefore, Socrates was a mammal.

Is-Ought Gaps

“No ought-judgment may be correctly inferred from a set of premises expressed only in terms of ‘is’.”

(“Hume’s Moral Philosophy,” *Stanford Encyclopedia of Philosophy*, Accessed 4 September 2016)

To illustrate:

- 1) Robot type A is square.**
- 2) Therefore, we ought to build a robot of type A.**

Simple normative arguments

- Stepping on a dog's tail for fun causes harm.
- It is wrong to cause harm, other things equal.
- Therefore, it is wrong to step on a dog's tail for fun.

Inductive Arguments

- “Valid arguments are risk-free. Inductive logic studies risky arguments” (Hacking 2001, p. 11).
- 1) These four oranges taken at random from the box are good.
 - 2) Therefore, the oranges in this box are good.

Arguments using probabilistic reasoning (e.g., Bayesian reasoning) and then stating a non-probabilistic conclusion are risky in this way.

Normative argumentation involving rights

- EU's AI Act fails to set gold standard for human rights - European Digital Rights (EDRI)

Example: “The AI Act falls short of civil society’s demand to ensure that EU-based AI providers whose systems impact people outside of the EU are subject to the same requirements as those inside the EU. The Act does not stop EU-based companies from exporting AI systems which are banned in the EU, therefore creating a huge risk of violating rights of people in non-EU countries by EU-made technologies that are essentially incompatible with human rights. Additionally, the Act does not require exported high-risk systems to follow the technical, transparency or other safeguards otherwise required when AI systems are intended for use within the EU, again risking the violation of rights of people outside of the EU by EU-made technologies.”

Extraction of argumentation from text

- Policy P in jurisdiction 1 leaves in place a significant risk that human rights will be violated in jurisdiction 2, compared to alternative policy P*
- (Other things equal, policies should reduce the risk of human rights violations in extraterritorial jurisdictions, relative to alternatives.)
- Therefore, policy P is inadequate in jurisdiction 1.

Part II: Philosophical foundations for intelligent machines

Philosophical fundamentals: terminology and agenda-setting

Necessary and sufficient conditions

Conditions on intelligence

Conditions on personhood

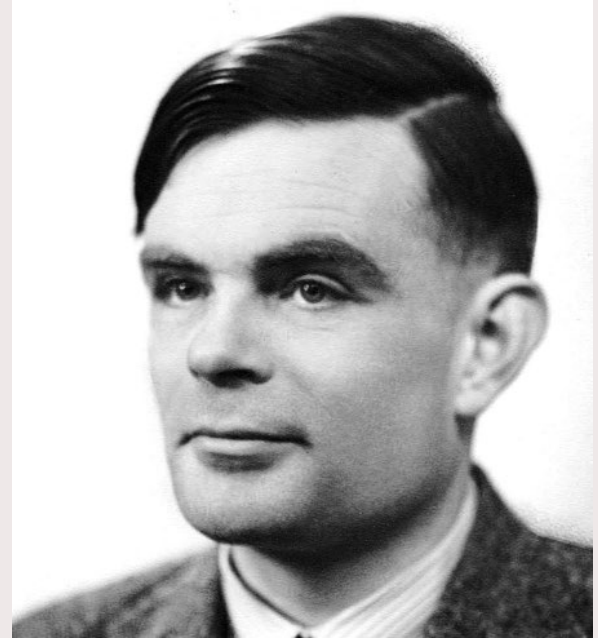


Some foundational aims

Classical AI aims to build artificial machines that have intelligence. It holds that any feature of (human) intelligence can be built into a machine.

In addition, it holds that a machine so built would help us understand (human) intelligence.

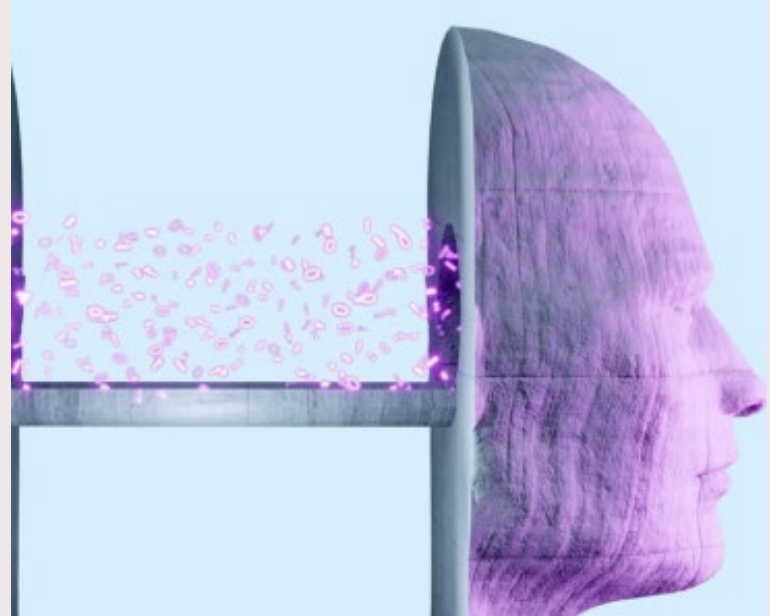
Strong AI actually has a mind and cognition.



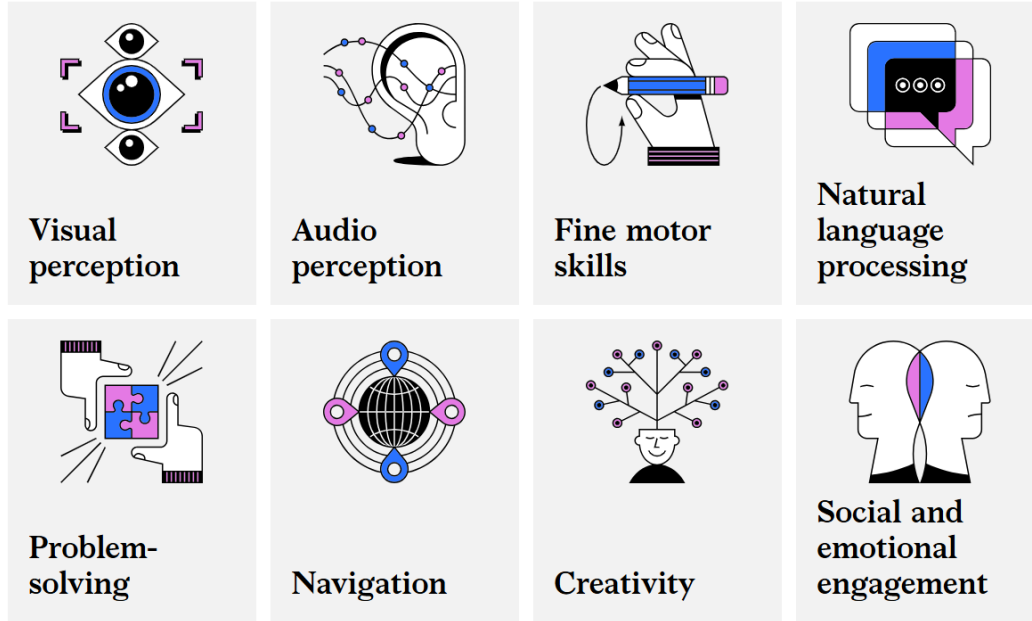
Artificial General Intelligence

We can have the ambition to recreate human or superhuman intelligence, while maintaining the current methodology of Technical AI (with which you are familiar or are becoming familiar), a set of computer-science methods for recreating the component capacities of human cognition.

Image: [What is Artificial General Intelligence \(AGI\)? | McKinsey](#)



McKinsey claims these skills are needed for AGI



Philosophy of AI

What does it take for computation to be intelligent (“to think”)? AI has prompted philosophers, cognitive scientists, and computer scientists to consider carefully what we mean by such terms.

Arguments that intelligence is computation-plus

The 4E argument: only embodied computation that interacts with the environment can be intelligent

The Chinese room argument: only computation with intentionality can be intelligent

Chinese room thought experiment visualized



[The Chinese Room Experiment](#) | [The Hunt for AI](#) | [BBC Studios](#) -

[YouTube](#) [Title]

Structure of argumentation

Necessary conditions/ modus tollens

1. E is strongly intelligent only if E can learn and reason. (Necessary condition for intelligence)
2. AI does not have such-and-such a capacity.
3. Therefore, AI is not strongly intelligent.

Structure of argumentation

Sufficient conditions/ modus ponens

1. If E has such-and-such a capacity, E is a person (Sufficient condition for personhood).
2. AI has such-and-such a capacity.
3. Therefore, AI is a person.


Candidate necessary conditions of personhood

Consciousness

Selfhood

Responsibility

Goal-setting



If any one of these is a necessary condition of personhood, but impossible for an intelligent, silicon-based machine to possess, then that machine cannot be a person. Usually, being a person is considered a basic kind of moral status (high degree of moral value requiring respect/ possessing rights).

Some questions to keep in mind for this course:

Can every human activity be broken down into tasks? What are ethics tasks?

Are there any metrics we can use to measure success at ethics tasks?

Is being good at ethics a component part of high intelligence?

Is deductive reasoning a necessary part of high intelligence?

That's it! Now on to some practical things...

Client consult meetings

First round of coaching meetings: format

Group and topic sign-up

Client consult meetings

10 minute presentation or introduction from the client

20 minutes for any student questions about the topic

Possible questions you have might concern:

- Target context (literature or website suggestions)
- Data sharing
- Prompting and technical suggestions
- Ethical issues

First Round of Coaching Meetings: Format

Timeframe and participants: 30 minutes total, two groups, 8 students + instructor

Prepare to answer questions about the concrete results of literature search, initial exploration of technology options, initial exploration of target context and projected user. Each individual student should be prepared to say something about what they have been doing to get ready for the submission of Phase 1.

Give initial answers to the following questions:

- What is the problem or need to which the AI tool is meant to respond?
- What frameworks or concepts can be used to understand the problem and structure a response to it?
- What technical options appear available to respond to the problem or need?

Self-assessment for coaching meetings

Basic	Intermediate	Advanced
Problem or need comes from imagination or brainstorm	Problem or need comes from literature or real life	Problem or need has been refined on the basis of literature or real-life sources
Framework/ concept is not present or merely intuitive	Framework/ concept comes from literature	Framework/ concept has been refined on the basis of literature
Technical options are explored at a layperson level	Technical options are explored at a scientifically informed level	Scientifically-informed options are explored in a systematic way

Sign-up for groups and topics