Introduction to Artificial Intelligence

Lecture 13: Fundamental Philosophical Issues

Nina Gierasimczuk

Outline

Important figures: back to Lecture 8

Is Artificial Intelligence possible?

Important figures: back to Lecture 8

Is Artificial Intelligence possible?

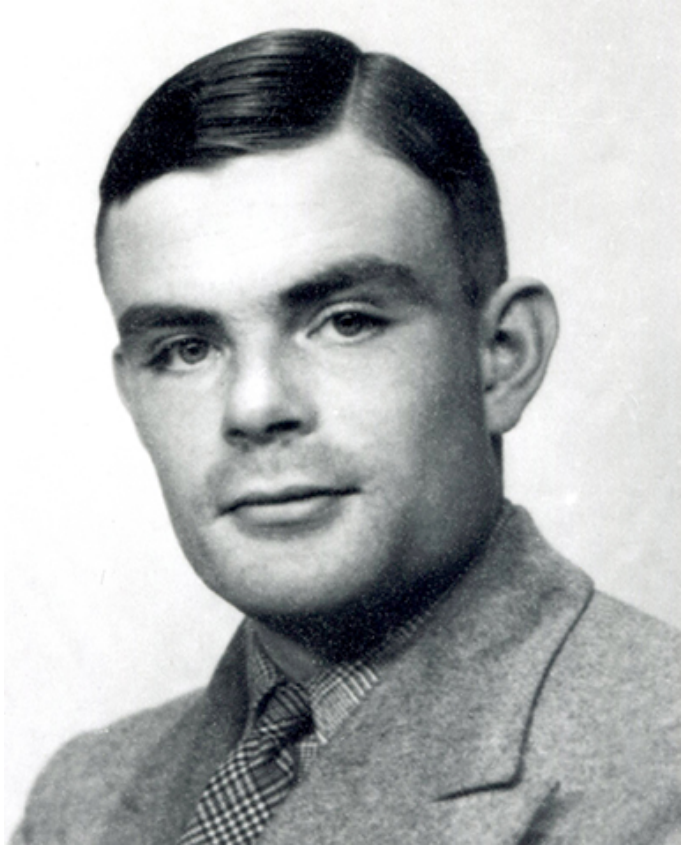


Figure: Alan Turing

Turing Test

HARD QUESTION:
Can machines think?

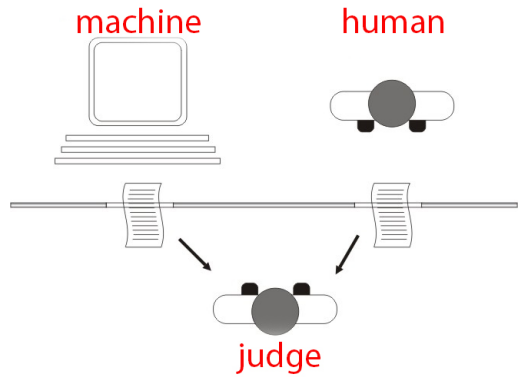
Turing Test

HARD QUESTION:
Can machines think?

Instead:

Can machines pass a behavioural intelligence test?

Alan Turing, 'Computing Machinery and Intelligence' (1950)



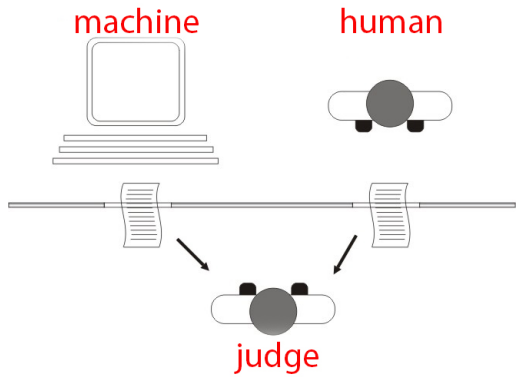
Turing Test

HARD QUESTION:
Can machines think?

Instead:

Can machines pass a behavioural intelligence test?

Alan Turing, 'Computing Machinery and Intelligence' (1950)



The start of chatbot technology.

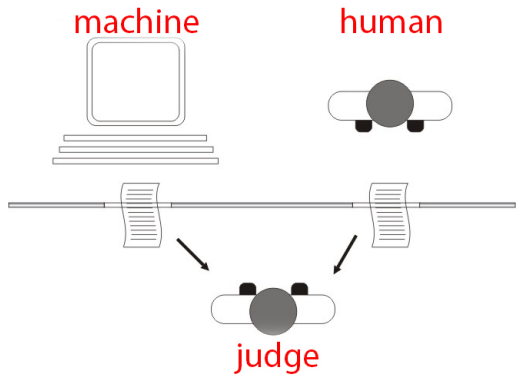
Turing Test

HARD QUESTION:
Can machines think?

Instead:

Can machines pass a behavioural intelligence test?

Alan Turing, 'Computing Machinery and Intelligence' (1950)



The start of chatbot technology.

WEAK AI: machines can behave **as if** they are intelligent

Objection: the argument from disability

For an ability X: a machine will never be able to do X.

Objection: the argument from disability

For an ability X: a machine will never be able to do X.

> *Be kind, resourceful, beautiful, friendly, have initiative, have a sense of humor, tell right from wrong, make mistakes, fall in love, enjoy strawberries and cream, make someone fall in love with it, learn from experience, use words properly, be the subject of its own thought, have as much diversity of behaviour as man, do something really new.*
>
> *(Alan Turing, Ibid.)*

Objection: the argument from disability

For an ability X: a machine will never be able to do X.

> *Be kind, resourceful, beautiful, friendly, have initiative, have a sense of humor, tell right from wrong, make mistakes, fall in love, enjoy strawberries and cream, make someone fall in love with it, learn from experience, use words properly, be the subject of its own thought, have as much diversity of behaviour as man, do something really new.*
>
> *(Alan Turing, Ibid.)*

> *We do not wish to penalise the machine for its inability to shine in beauty competitions, nor to penalise a man for losing in a race against an aeroplane. The conditions of our game make these disabilities irrelevant.*
>
> *(Alan Turing, Ibid.)*

Objection: the argument from disability

For an ability X: a machine will never be able to do X.

> *Be kind, resourceful, beautiful, friendly, have initiative, have a sense of humor, tell right from wrong, make mistakes, fall in love, enjoy strawberries and cream, make someone fall in love with it, learn from experience, use words properly, be the subject of its own thought, have as much diversity of behaviour as man, do something really new.*
>
> *(Alan Turing, Ibid.)*

> *We do not wish to penalise the machine for its inability to shine in beauty competitions, nor to penalise a man for losing in a race against an aeroplane. The conditions of our game make these disabilities irrelevant.*
>
> *(Alan Turing, Ibid.)*

Nowadays: isolated attempts to show the advantage of machines over humans in certain domains.

Objection: the argument from informality

It is not possible to produce a set of rules purporting to describe what a man should do in every conceivable set of circumstances (...) if each man had a definite set of rules of conduct by which he regulated his life he would be no better than a machine. But there are no such rules, so men cannot be machines. *(Alan Turing, Ibid.)*

Objection: the argument from informality

It is not possible to produce a set of rules purporting to describe what a man should do in every conceivable set of circumstances (...) if each man had a definite set of rules of conduct by which he regulated his life he would be no better than a machine. But there are no such rules, so men cannot be machines. *(Alan Turing, Ibid.)*

We cannot so easily convince ourselves of the absence of complete laws of behaviour as of complete rules of conduct.

(Alan Turing, Ibid.)

Objection: the argument from informality

> *It is not possible to produce a set of rules purporting to describe what a man should do in every conceivable set of circumstances (...) if each man had a definite set of rules of conduct by which he regulated his life he would be no better than a machine. But there are no such rules, so men cannot be machines.* *(Alan Turing, Ibid.)*

> *We cannot so easily convince ourselves of the absence of complete laws of behaviour as of complete rules of conduct.*
>
> *(Alan Turing, Ibid.)*

Nowadays: situated agents rather than disembodied logical inference engines; focus on embodied cognition and environment.

Objection: Gödel's Incompleteness

For any formal axiomatic system F powerful enough to do arithmetic, it is possible to construct a so-called Gödel sentence $G(F)$ with the following properties:

- $G(F)$ is a sentence of F, but cannot be proved in F;
- If F is consistent, then $G(F)$ is true.

Objection: Gödel's Incompleteness

For any formal axiomatic system F powerful enough to do arithmetic, it is possible to construct a so-called Gödel sentence $G(F)$ with the following properties:

- $G(F)$ is a sentence of F, but cannot be proved in F;
- If F is consistent, then $G(F)$ is true.

Argument: Gödel's incompleteness indicates that machines are 'mentally' inferior to humans, because machines are formal systems that are limited by the incompleteness theorem—they cannot establish the truth of their own Gödel sentence—while humans have no such limitations. (J.R. Lucas, 1961)

Objection: Gödel's Incompleteness

For any formal axiomatic system F powerful enough to do arithmetic, it is possible to construct a so-called Gödel sentence $G(F)$ with the following properties:

- $G(F)$ is a sentence of F, but cannot be proved in F;
- If F is consistent, then $G(F)$ is true.

Argument: Gödel's incompleteness indicates that machines are 'mentally' inferior to humans, because machines are formal systems that are limited by the incompleteness theorem—they cannot establish the truth of their own Gödel sentence—while humans have no such limitations. (J.R. Lucas, 1961)

> *We too often give wrong answers to questions ourselves to be justified in being very pleased at such evidence of fallibility on the part of the machines.* *(Alan Turing, Ibid.)*

Strong AI

BUT STILL:

Can machines really think?

Strong AI

BUT STILL:
Can machines really think?

'Really thinking' includes: consciousness, phenomenology, intentionality, etc.

Strong AI

BUT STILL:
Can machines really think?

'Really thinking' includes: consciousness, phenomenology, intentionality, etc.

> *According to the most extreme form of this view the only way by which one could be sure that machine thinks is to be the machine and to feel oneself thinking. Likewise according to this view the only way to know that a man thinks is to be that particular man. It is in fact the solipsist point of view.*
>
> *(Alan Turing, Ibid.)*

Strong AI

BUT STILL:
Can machines really think?

'Really thinking' includes: consciousness, phenomenology, intentionality, etc.

> *According to the most extreme form of this view the only way by which one could be sure that machine thinks is to be the machine and to feel oneself thinking. Likewise according to this view the only way to know that a man thinks is to be that particular man. It is in fact the solipsist point of view.*
>
> *(Alan Turing, Ibid.)*

Polite convention: let's politely agree that everybody thinks.

But what is it that humans have that machines do not?

Another HARD QUESTION:
What is a mind?

But what is it that humans have that machines do not?

Another HARD QUESTION:
What is a mind?

If we know what it means for a human to have a mind,
we should be able to decide if we can ascribe it to other entities,
in particular, machines!

But what is it that humans have that machines do not?

Another HARD QUESTION:
What is a mind?

If we know what it means for a human to have a mind,
we should be able to decide if we can ascribe it to other entities,
in particular, machines!

Mind-body problem:

what is the relationship between mind and body?

Dualism: they are two completely separate realms.

René Descartes, Meditation on First Philosophy, 1641

But then, how can the mind control the body?!

Figure: Descartes

But what is it that humans have that machines do not?

Another HARD QUESTION:
What is a mind?

If we know what it means for a human to have a mind,
we should be able to decide if we can ascribe it to other entities,
in particular, machines!

Mind-body problem:

what is the relationship between mind and body?

Dualism: they are two completely separate realms.

René Descartes, Meditation on First Philosophy, 1641

But then, how can the mind control the body?!

Figure: Descartes

But **monists** and **physicalists** also face problems, like **brains in a vat**.

Another option

Mental states is what happens between the input and the output.

Functionalism

Another option

Mental states is what happens between the input and the output.

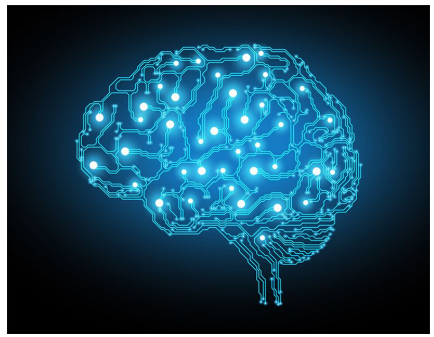
Your brain is gradually replaced with appropriate input-output circuits...
Would you retain consciousness?
Robotists are inclided to say: yes!
Philosophers, not necessarily...

Functionalism

Another option

Mental states is what happens between the input and the output.

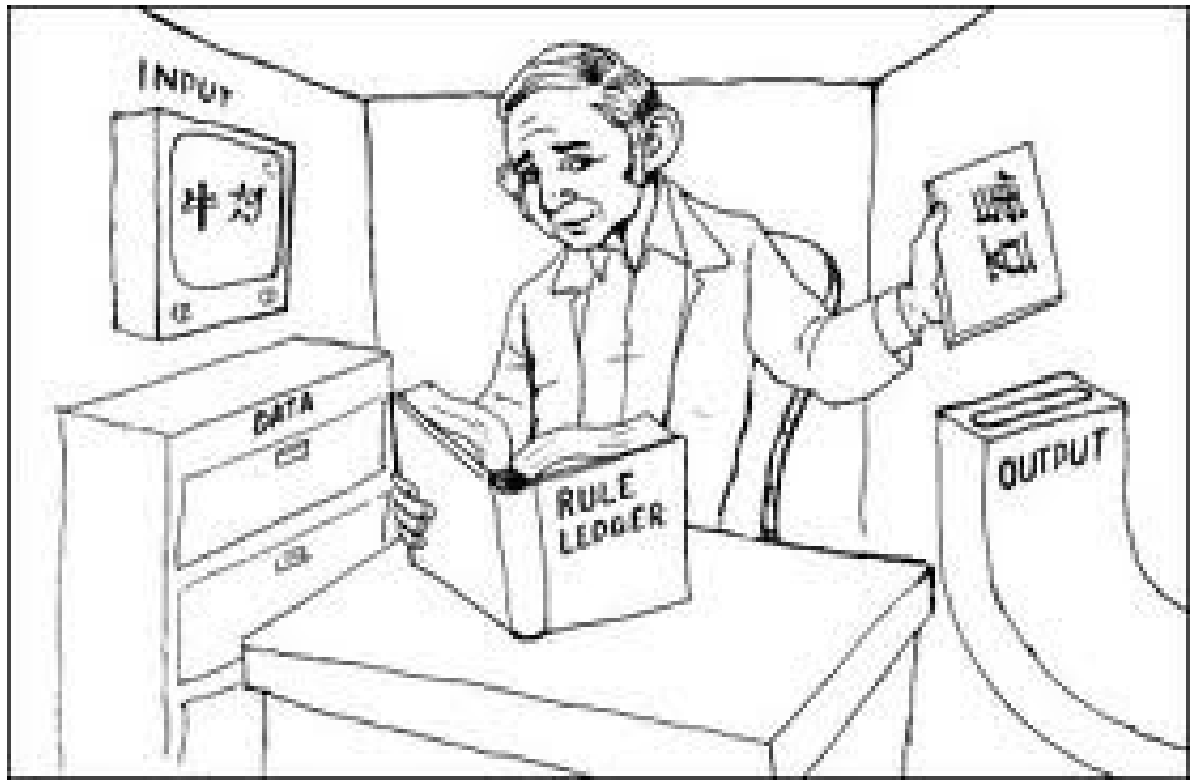
Your brain is gradually replaced with appropriate input-output circuits...
Would you retain consciousness?
Robotists are inclided to say: yes!
Philosophers, not necessarily...



> *You find, to your total amazement, that you are indeed losing control of your external behavior. You find, for example, that when doctors test your vision, you hear them say 'We are holding up a red object in front of you; please tell us what you see.' You want to cry out 'I can't see anything. I'm going totally blind.' But you hear your voice saying in a way that is completely out of your control, 'I see a red object in front of me.' (...) your conscious experience slowly shrinks to nothing, while your externally observable behavior remains the same.*
>
> *(Searle, 1992)*

Biological naturalism: Chinese Room

Chinese Room is a **thought experiment** by John Searle.



Where is **understanding**? Where is **consciousness**?

Consciousness

Consciousness can be thought of as understanding and self-awareness.

Consciousness

Consciousness can be thought of as understanding and self-awareness.

The problem of subjective experience (*qualia*):

the **inverted spectrum** argument.

Consciousness

Consciousness can be thought of as understanding and self-awareness.

The problem of subjective experience (*qualia*):

the **inverted spectrum** argument.

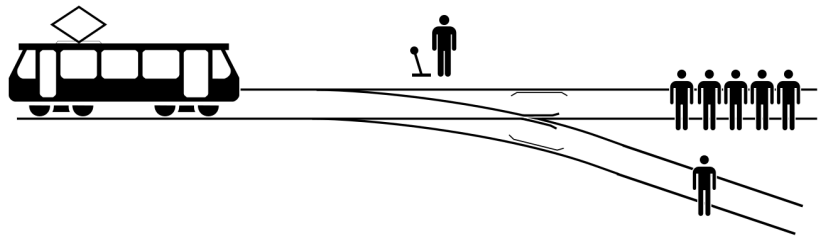
There is a methodological between scientific approach and subjective experiences, so called **explanatory gap**.

Outline

Ethics in AI

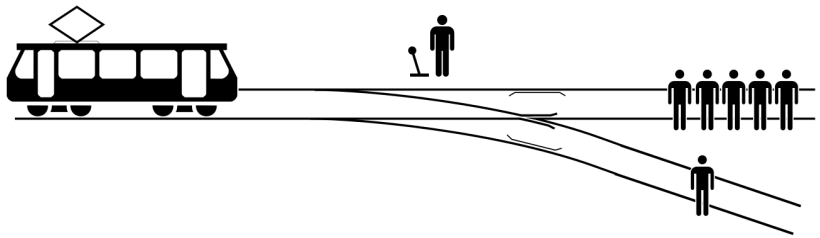
1. Robot rights.
2. Human dignity.
3. Transparency, accountability and open source.
4. Weaponisation.
5. Machine ethics.

Trolley Problem: Thought Experiment in Ethics



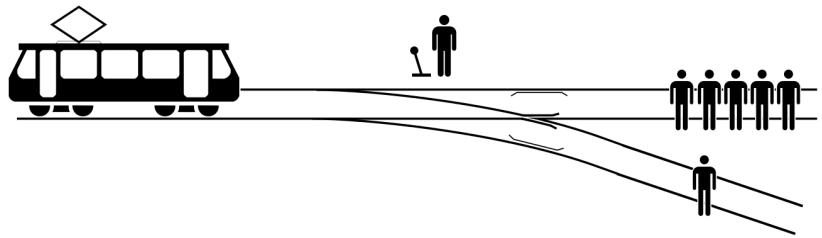
There is a runaway trolley barreling down the railway tracks. Ahead, on the tracks, there are five people tied up and unable to move. The trolley is headed straight for them. You are standing some distance off in the train yard, next to a lever. If you pull this lever, the trolley will switch to a different set of tracks. However, you notice that there is one person tied up on the side track.

Trolley Problem: Thought Experiment in Ethics



There is a runaway trolley barreling down the railway tracks. Ahead, on the tracks, there are five people tied up and unable to move. The trolley is headed straight for them. You are standing some distance off in the train yard, next to a lever. If you pull this lever, the trolley will switch to a different set of tracks. However, you notice that there is one person tied up on the side track. You have two options:

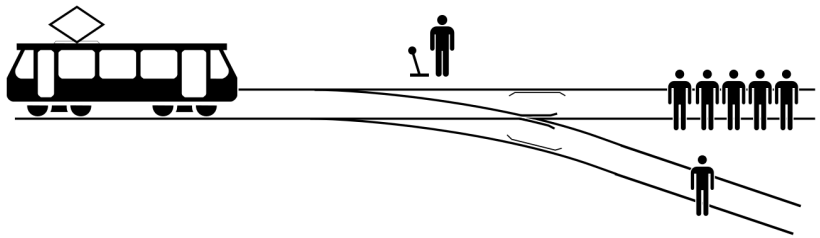
Trolley Problem: Thought Experiment in Ethics



There is a runaway trolley barreling down the railway tracks. Ahead, on the tracks, there are five people tied up and unable to move. The trolley is headed straight for them. You are standing some distance off in the train yard, next to a lever. If you pull this lever, the trolley will switch to a different set of tracks. However, you notice that there is one person tied up on the side track. You have two options:

1. Do nothing, and the trolley kills the five people on the main track.
2. Pull the lever, diverting the trolley onto the side track where it will kill one person.

Trolley Problem: Thought Experiment in Ethics

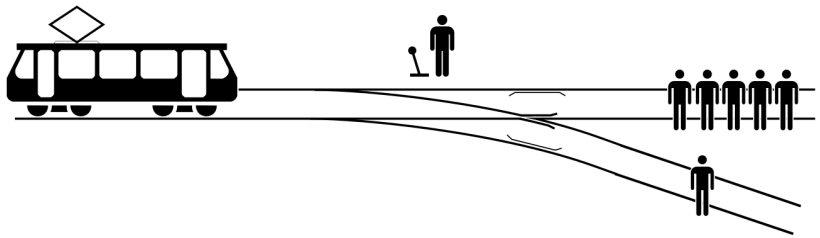


There is a runaway trolley barreling down the railway tracks. Ahead, on the tracks, there are five people tied up and unable to move. The trolley is headed straight for them. You are standing some distance off in the train yard, next to a lever. If you pull this lever, the trolley will switch to a different set of tracks. However, you notice that there is one person tied up on the side track. You have two options:

1. Do nothing, and the trolley kills the five people on the main track.
2. Pull the lever, diverting the trolley onto the side track where it will kill one person.

What do you do?

Trolley Problem: Thought Experiment in Ethics



There is a runaway trolley barreling down the railway tracks. Ahead, on the tracks, there are five people tied up and unable to move. The trolley is headed straight for them. You are standing some distance off in the train yard, next to a lever. If you pull this lever, the trolley will switch to a different set of tracks. However, you notice that there is one person tied up on the side track. You have two options:

1. Do nothing, and the trolley kills the five people on the main track.
2. Pull the lever, diverting the trolley onto the side track where it will kill one person.

What do you do?

Go to www.menti.com and use the code **2611 2140**.

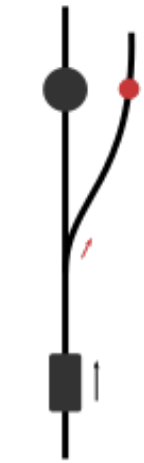
Some background

- The problem introduced by Philippa Foot in 1967.
- Since 2001 used extensively in empirical research on moral psychology.
- Relevant in the ethics of the design of **autonomous vehicles**.

Some traditional positions

- **utilitarian**: it is obligatory to steer to the track with one man on it
- **incommensurability** of human lives
- since moral wrongs are already in place in the situation, interfering is **participation in the moral wrong**
- or maybe being present and able to influence its outcome constitutes an **obligation to participate**
- the problem highlights the difference between **deontological** and **consequentialist** ethical systems

Different versions of the Trolley Problem



the switch
Foot, 1967

the fat man
Thomson, 1976

the fat villain

the loop
Costa, 1987

the man in the yard
Unger, 1992

Empirical research

- Joshua Greene et al. (2001): empirical investigation of **people's responses** to trolley problems.
- They used **functional magnetic resonance imaging** (FMRI), and
- showed that "personal" dilemmas (like pushing a man off a footbridge) engage brain regions of emotion, whereas "impersonal" dilemmas (like diverting the trolley by flipping a switch) engaged regions of controlled reasoning.
- This led to **dual-process account of moral decision-making**.
- Since then influence of stress, emotional state, different types of brain damage, physiological arousal, different neurotransmitters, and genetic factors, language and cultural difference were studied.

Trolley Problem and autonomous cars

- During a potential crash scenario the software decides between multiple courses of action, all of which may cause harm.
- MIT Media Lab made Moral Machine to gather info about public opinion.
- VR is used by others to test behavior in experimental settings.
- Since 2016 Germany has an ethical commission, which defined 20 rules for autonomous and connected driving, obligatory for upcoming laws.

Let's test the moral faculties of the class!

MIT Moral Machine

End of Lecture 13