

The Human Factor in AI: It's All About Explanations

Tax and Law Day 2025

Max Velthoven

January 2025



The better the question. The better the answer. The better the world works.



Shape the future
with confidence

Agenda

What is this talk going to be about?

1. The philosophy of science: Karl Popper and David Deutsch

Break

2. AI technology
3. AI and philosophy
4. Questions/comments from the audience

Conclusion will be: humans are still very relevant, as humans are *universal explainers*.

Why do *you* talk about this stuff?

Introduction to Popper and Deutsch

- **Karl Popper (1902 - 1994):** philosopher and academic, mostly known for his work on the philosophy of science and society (critique of Plato/Marx/Hegel).
- **David Deutsch (1953):** physicist who is widely regarded as the father of quantum computing.
- Throughout his works, Popper attacked many of the mainstream ideologies in science such as inductivism, positivism, empiricism (**Falsification Principle**).
- Deutsch is an ardent Popperian and has worked with many of Popper's ideas, also in **AI/AGI**.

Deutsch and Quantum Computing

- **Turing Principle of computational universality:** any physical system can be simulated arbitrarily well using a model computer equivalent to the Turing Machine.



- **Deutsch' universal quantum computer:** Deutsch showed that if quantum theory is true (which we believe it is) then the Turing Principle is true, too.

Willow's performance on this benchmark is astonishing: It performed a computation in under five minutes that would take one of today's **fastest supercomputers** 10^{25} or 10 septillion years. If you want to write it out, it's 10,000,000,000,000,000,000,000,000 years. This mind-boggling number exceeds known timescales in physics and vastly exceeds the age of the universe. It lends credence to the notion that quantum computation occurs in many parallel universes, in line with the idea that we live in a multiverse, a **prediction** first made by **David Deutsch**.

Source: <https://blog.google/technology/research/google-willow-quantum-chip/>

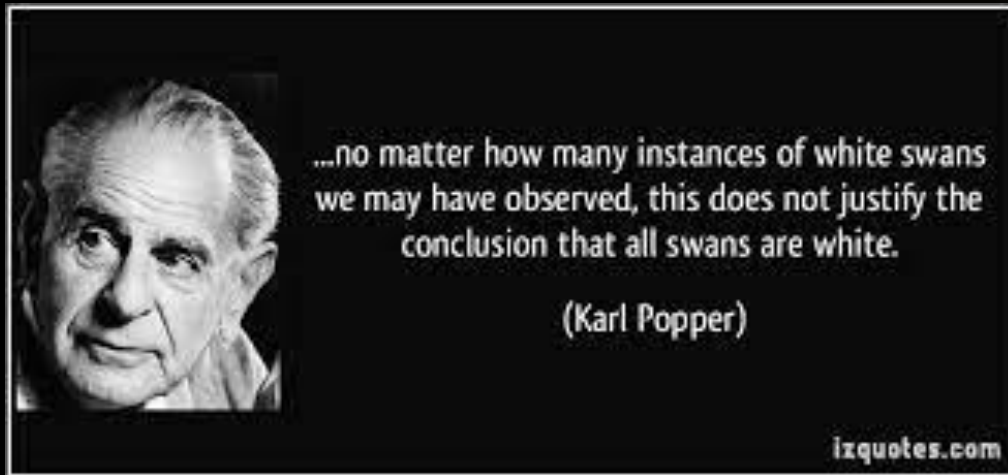
The commonsense theory of knowledge: Empiricism combined with inductivism

1. Someone gathers data
2. A general theory is 'induced' from the data
3. More data further proves the theory

Does knowledge really come from our senses?

How do finite data justify the truth of universal theories?

Popper's attack on inductivism



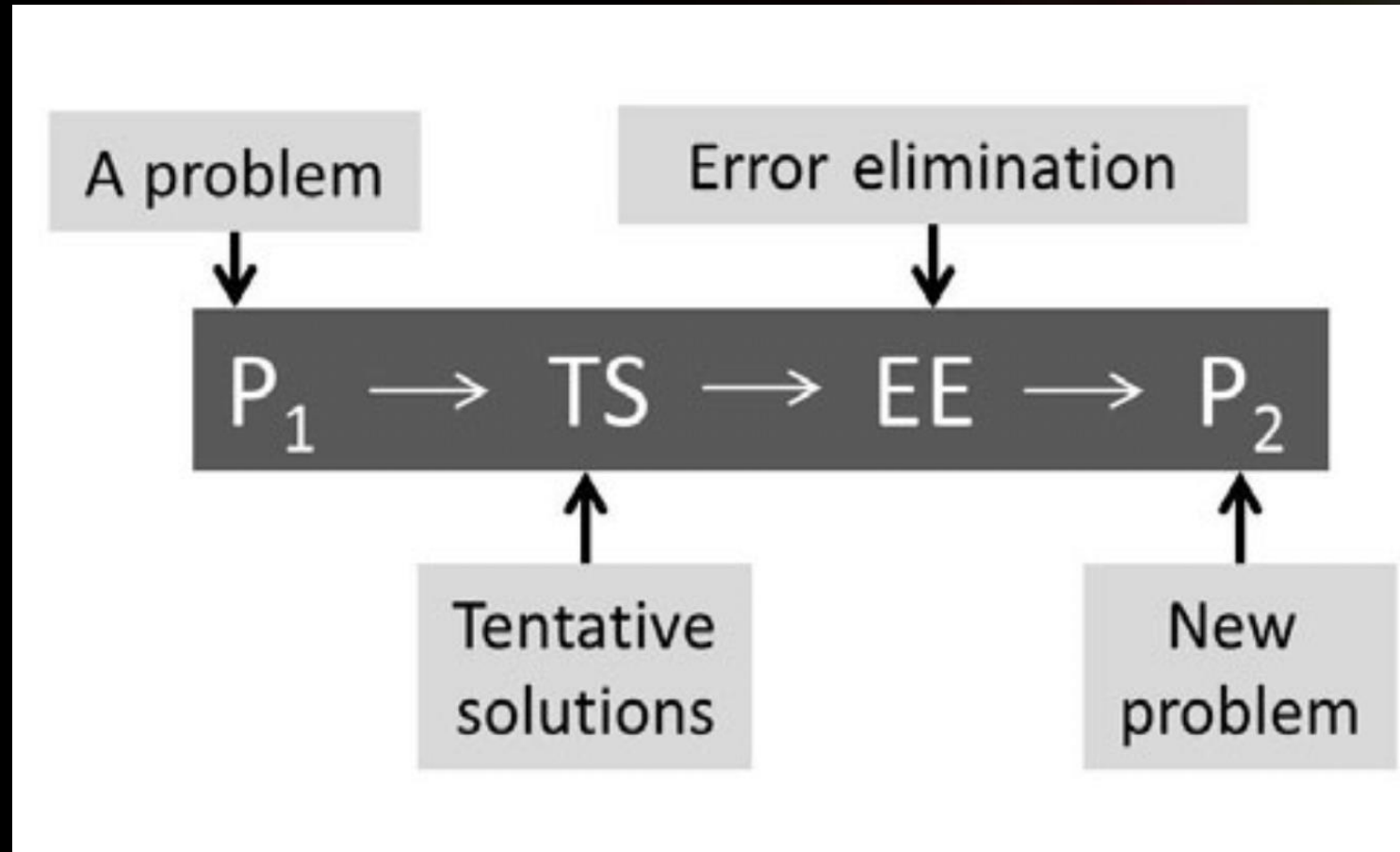
Black Swans and the
Limits of Inductive Reasoning



Popper's theory of knowledge

1. There is a **problem** in an existing theory.
2. To solve this problem, new ideas are **conjectured (explanations)**.
3. The conjectures are then **critically examined**, both via experiments and through critical discussions.
4. (2) and (3) are performed until a satisfactory **solution** is found.
5. We arrive at a **new problem situation** and the process starts again at (1).

Popper's theory of knowledge

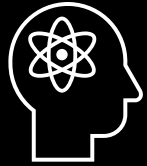


From Firestone, J.M., McElroy, M.W.: Key issues in the new knowledge management. Butterworth-Heinemann (2003)

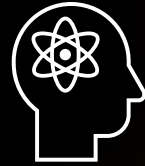
Error-correction is also the way scientific knowledge grows

"How can we explain the movement of stars & planets?"

Copernicus



Kepler



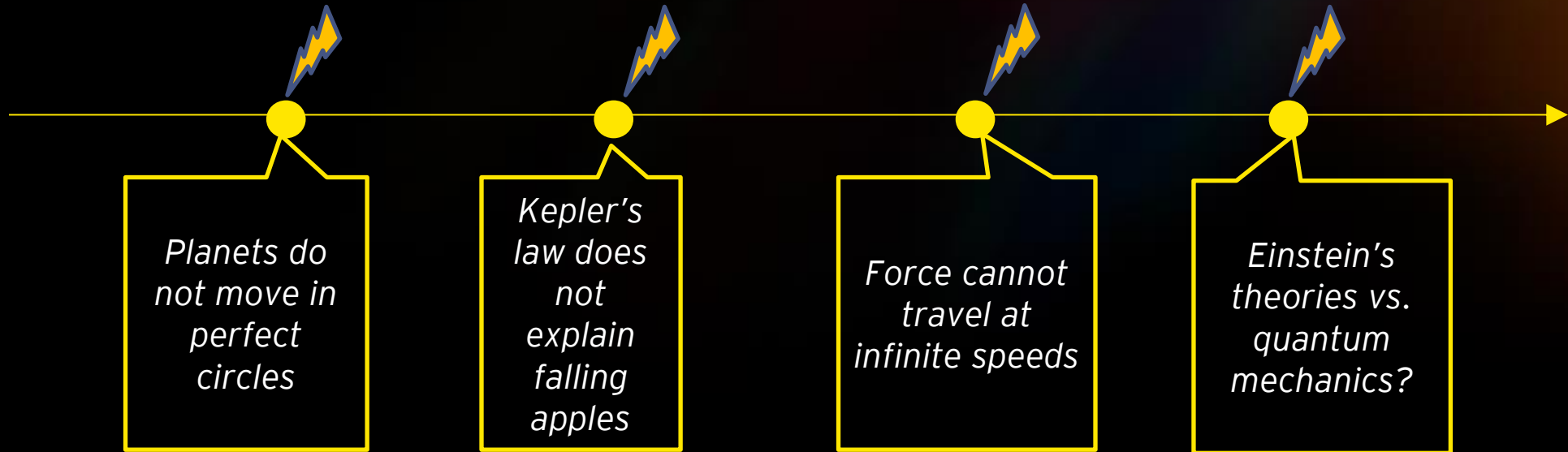
Newton



Einstein



?



Popper's theory of knowledge

What Newton actually did was conjecture a universal theory (**an explanation**), which **cannot be proven true**, either by experiment or any other way.

Newton's theory includes a lot of **unseen entities**, like forces.

However, the theory and the unseen entities it postulates **can be refuted by experiments** because:

If a universal theory U implies an observation O , then ' O is false' implies ' U is false'

Newton's theory of gravity was already theoretically refuted by Einstein's theory. The Eddington experiment (solar eclipse) provided an experimental test, which showed that " O was false" for Newton's theory, but not false for Einstein's theory.

Popper's theory of knowledge

A solution is a hard to vary theory (conjecture) that explains the problem.

A theory is hard to vary if changing any part of it ruins the explanation.

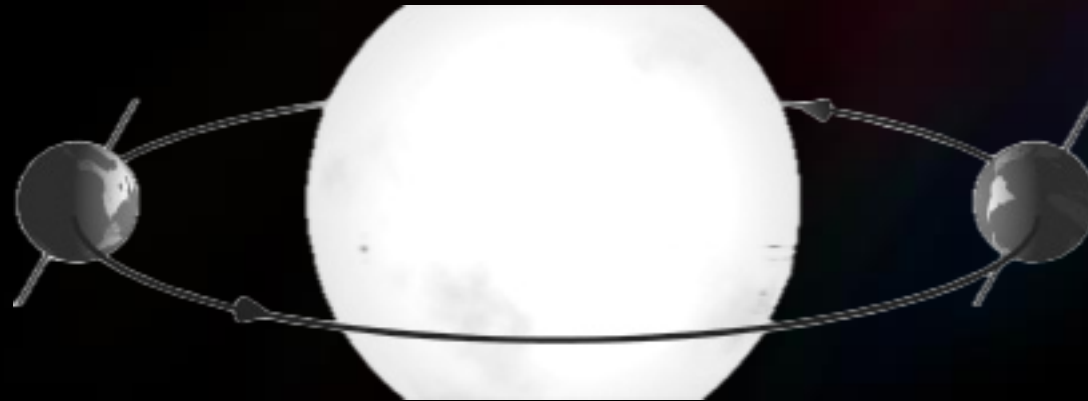
An easy to vary theory can be changed while still 'explaining' the problem.

Popper's theory of knowledge

Explanations account for the seen in terms of the unseen

An easy to vary explanation: the myth of Persephone

A hard to vary explanation: the modern theory of the seasons



From Deutsch, David. *The Beginning of Infinity: Explanations That Transform the World*. Penguin, 2011.

Popper's theory of knowledge

A theory is better than a rival if, for instance:

- *It has fewer assumptions than a rival theory.*
- *It makes more novel predictions than a rival theory.*
- *It solves not only the problems it was intended to solve but also others besides it.*

Popper proposed the following ways to improve problem-solving:

- *We should try to identify new problems in existing explanations.*
- *We should state our problems and proposed explanations as clearly and simply as we can.*
- *We should subject our proposed explanations to critical scrutiny and seek out and invite criticisms of them.*

From Frederick, Danny. *Against the Philosophical Tide: Essays in Popperian Critical Rationalism* (p. 12). Critias Publishing. Kindle Edition.

'I think (like you, by the way) that theory cannot be fabricated out of the results of observation, but that it can only be invented.'

Albert Einstein in a letter to Karl Popper

"There is only one way of thinking that is capable of making progress and that is the way of seeking good explanations through creativity and criticism"

- David Deutsch

BREAK

The operation of AI technology

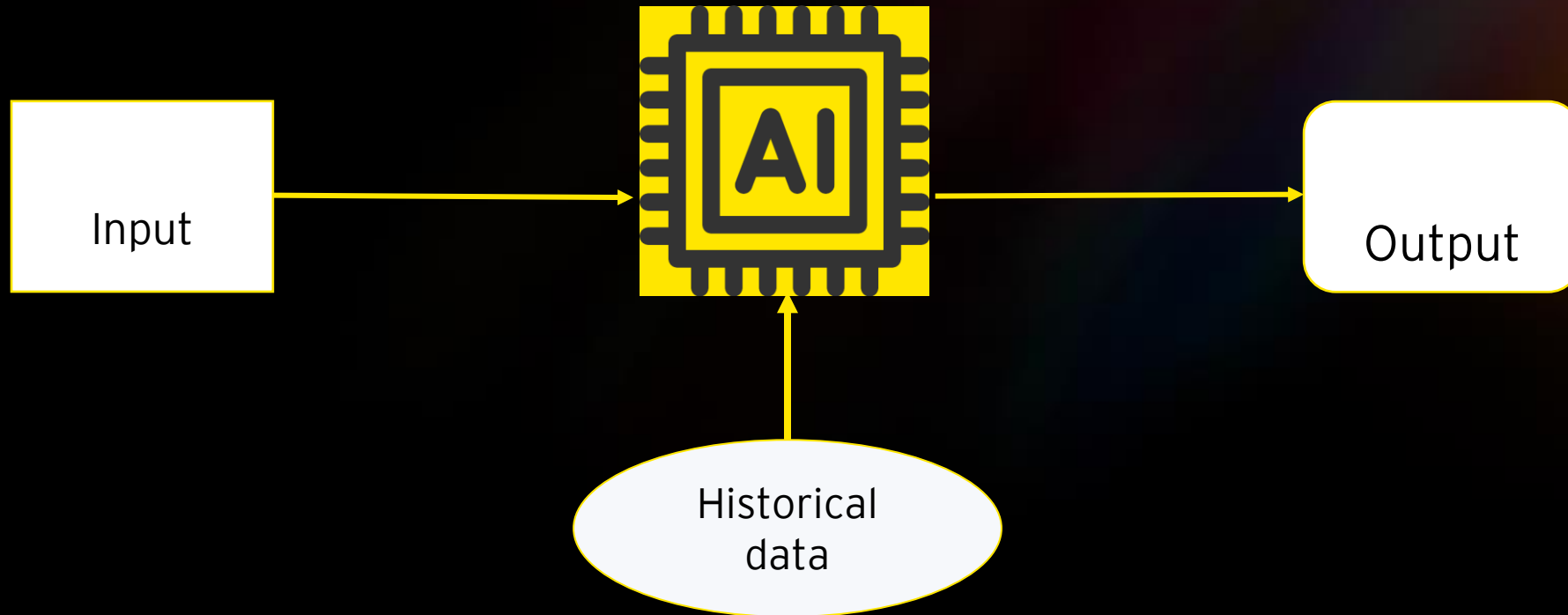
Introduction to Artificial Intelligence | Definition

"Software that can, for a given set of human-defined objectives, generate outputs such as content, predictions, recommendations, or decisions influencing the environments they interact with."

Introduction to Artificial Intelligence | Example

An AI model learns patterns from historical data.

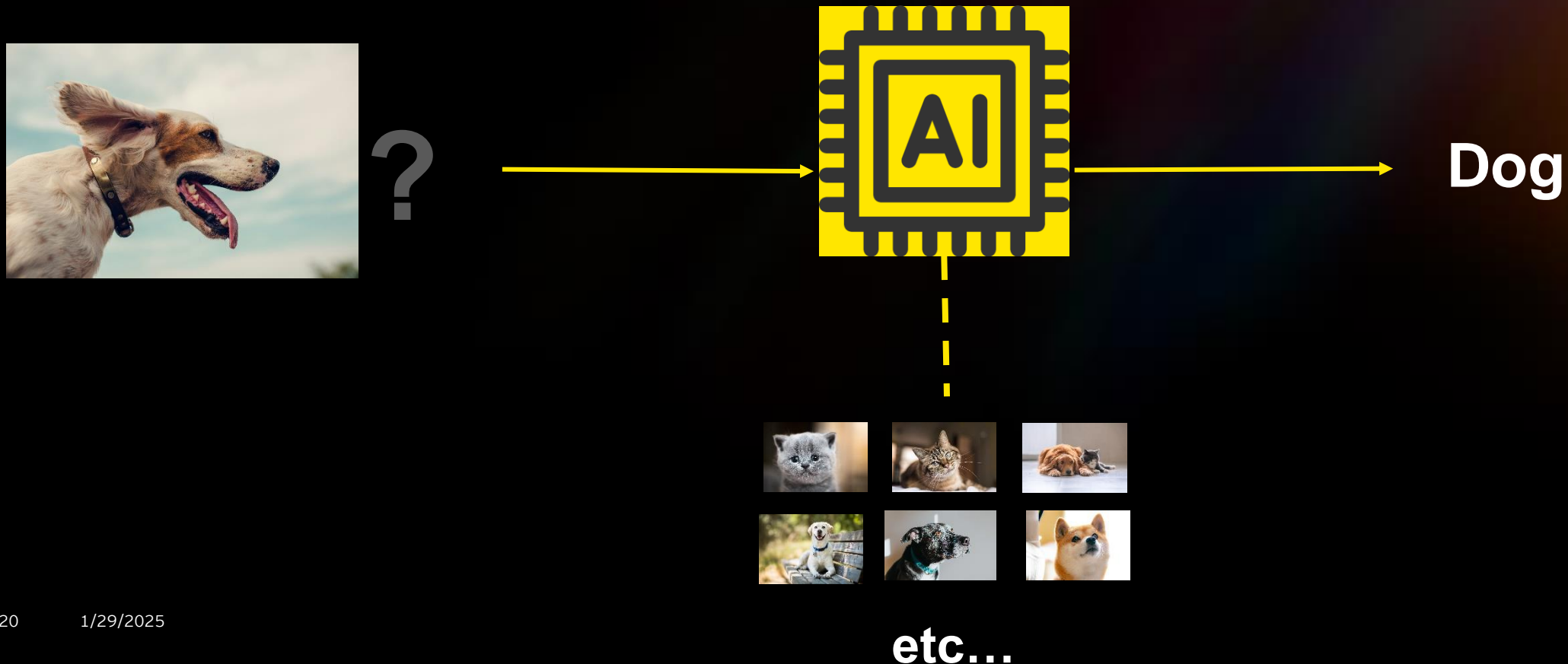
- First: 'train' the model to learn those patterns
- Then: Present the model new input it has never seen, and the model predicts an output



Introduction to Artificial Intelligence | Example

An AI model learns patterns from historical data.

- First: 'train' the model to learn those patterns
- Then: Present the model new input it has never seen, and the model predicts an output



Proposition 1

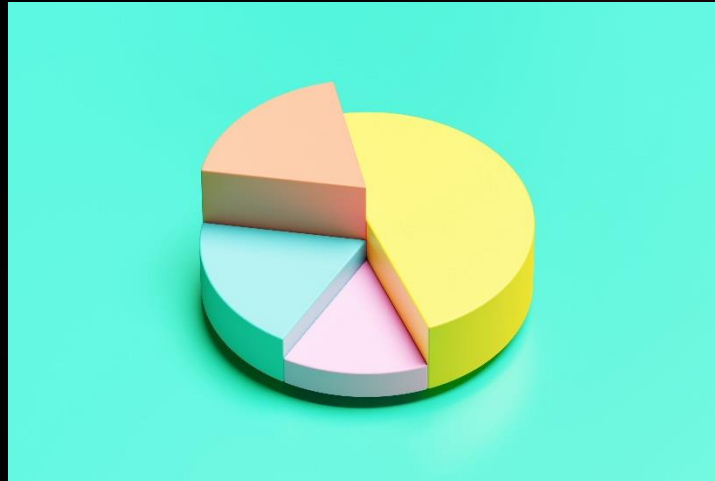
There are many similarities between the philosophical ideas surrounding AI (if any) and mistaken philosophies on the growth of knowledge which were criticized by Popper and Deutsch.

Popper, Deutsch, and AI

- AI models are *induction machines*: from particular observations they make general statements
- An AI model trained upon only white swans *will* predict all swans to be white
- An AI model, thus, does **not** generate new knowledge as people do
- People generate knowledge via *conjecture* (explanation) and *refutation*

What about data/ facts/ observations ?

- *Is a set of data/ facts/ observations ... an explanation ?*



No ...

Data does not explain anything by itself ...

it does not account for the seen in terms of the unseen

Proposition 2

Human actors should continue to bear the ultimate responsibility to *explain* the application of AI

Return to philosophy: instrumentalism

- A closely related misconception to induction, which is omnipresent in discussions around AI, is *instrumentalism*.
- Instrumentalism is the idea that knowledge is a tool: if it does the job, it works and it does not have to be questioned or explained.
- There are at least two major issues with instrumentalism from the perspective of Popper's philosophy:
 1. Instrumentalism does not lend itself to *falsification*: a piece of knowledge either works or it does not but is never falsified.
 2. Instrumentalism is *explanationless* science, whereas Popperian science is all about explanations.
- In a real-world application, instrumentalism can result in various problematic outcomes:
 1. Lack of explainability.
 2. Lack of responsibility.
- Not all "instrumentalist" approaches in the real world are problematic:
 - Bridges are still built today by using Newton's (refuted) theory of gravity, whereas Einstein's theory is used in satellites.
 - Crucially, part of the success of Einstein's theory is that it *explains* why Newton's theory can be used in bridges but not in satellites.

Introduction to Explainable AI

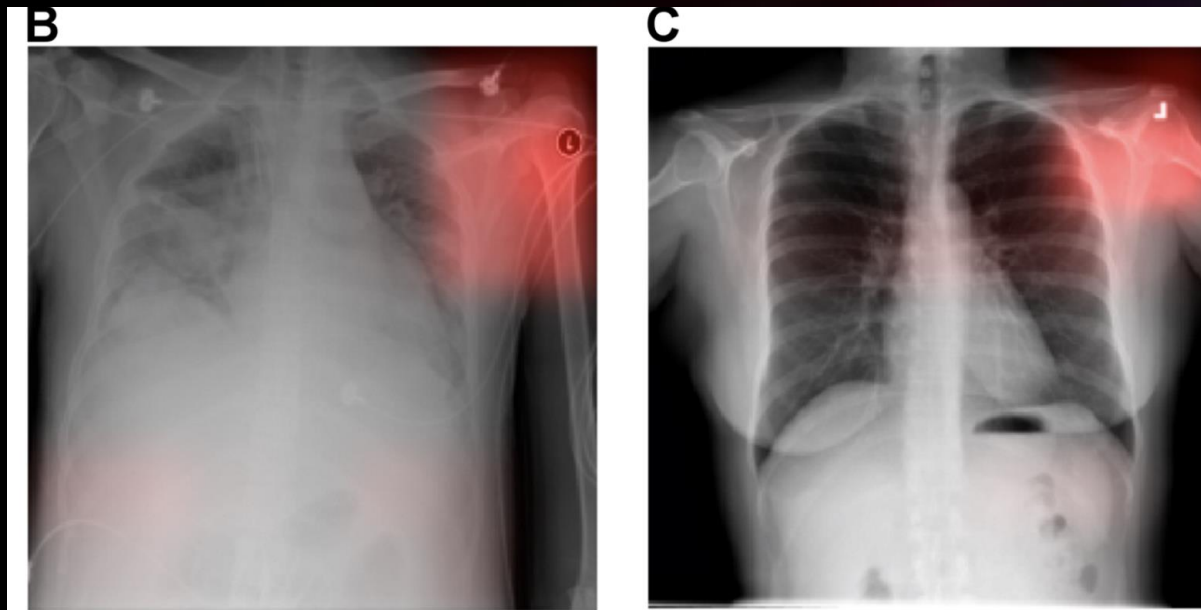
Explainable AI is a field within AI that aims to explain the AI model and its output, in a way that makes sense to a human being.

What can go wrong? A medical AI example

- Suppose we have a model trained on X-rays from various hospitals
- It obtains 90% accuracy in predicting cancer presence

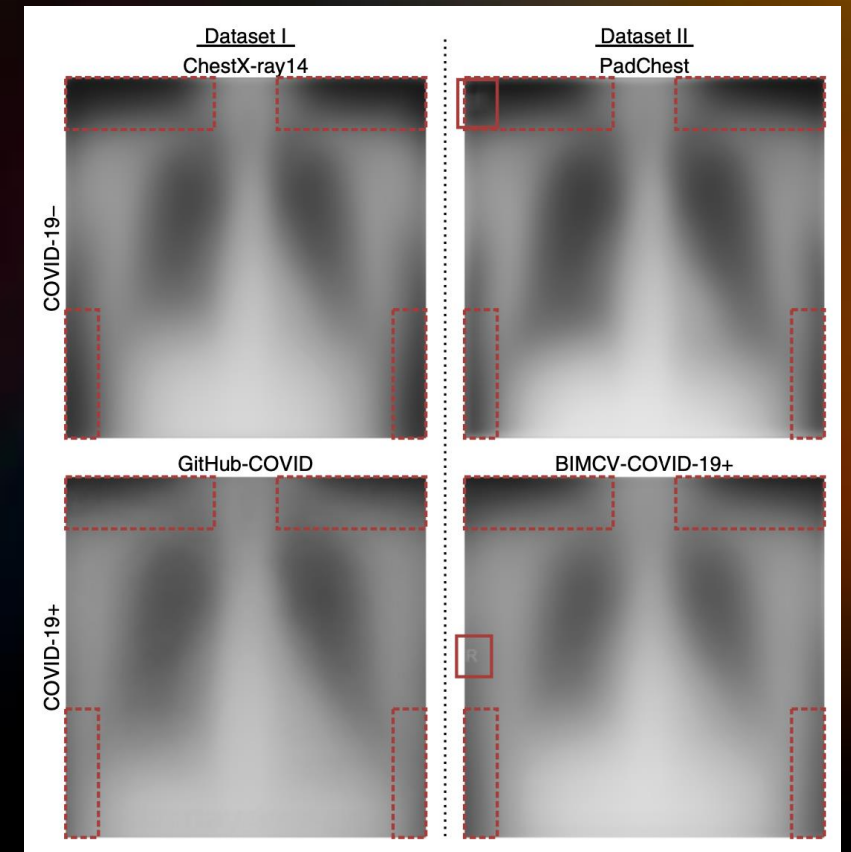
What can go wrong?

- Different hospitals use different 'metal tokens' on X-rays
- Predicting the cancer-specific hospital yields high accuracy

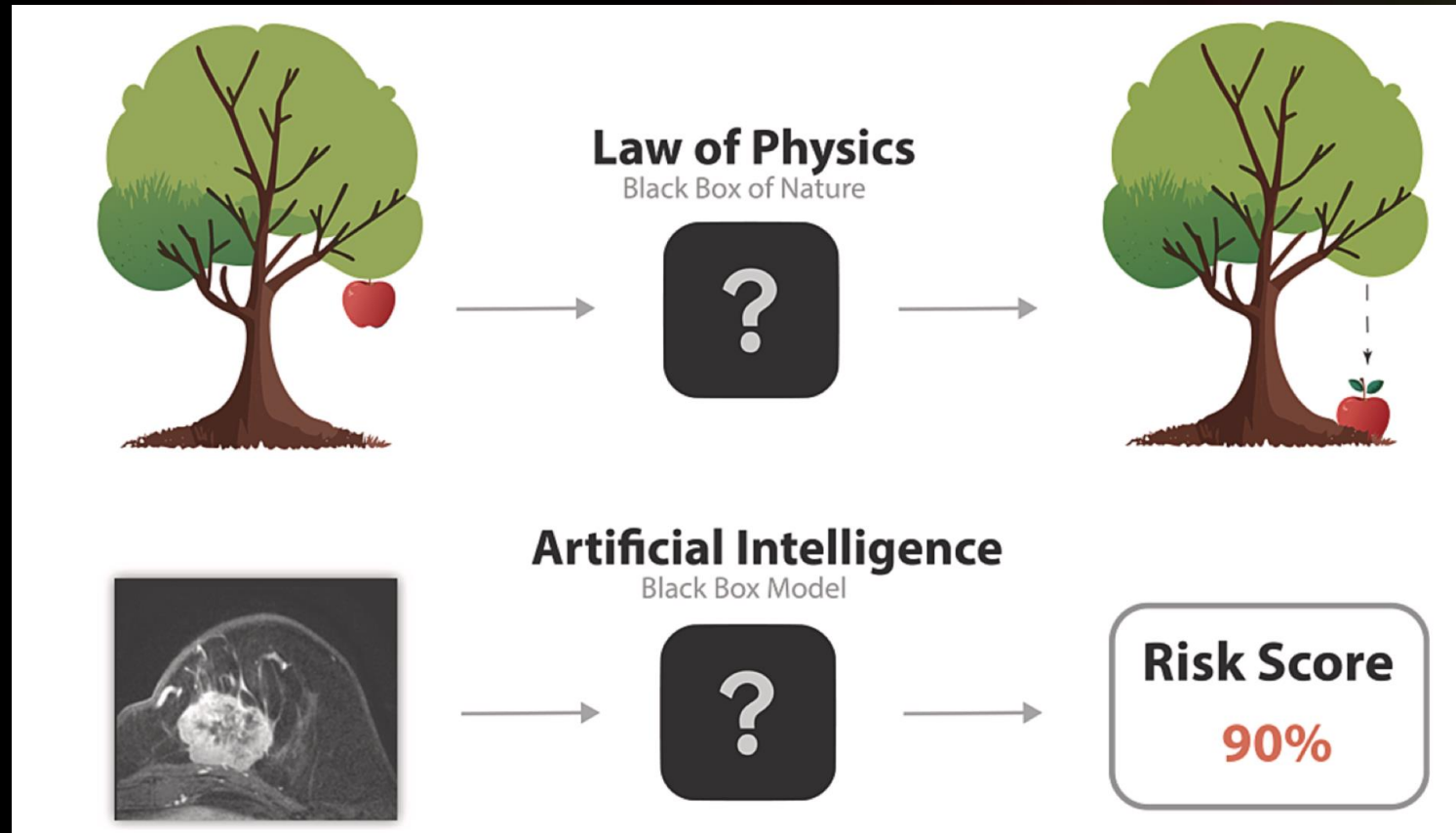


What can go wrong?

- Shortcuts can persist over external validation
 - *Positioning* correlated with disease prevalence in multiple datasets
 - Still good scores on all external validation sets
- Another example: *age*



Analogy with Sciences



Eric Marcus, Jonas Teuwen, Artificial intelligence and explanation: How, why, and when to explain black boxes

Analogy with Sciences

- To understand, learn from, and prevent mistakes, we (humans) need to create *explanations* of what the AI does.
- Creating explanations of black box AI is analogous to creating explanations in natural sciences, such as physics.

Another example: AI and taxation

AI models rely on historical data, and thus can reproduce or exacerbate existing patterns of discrimination/socio-economic unfairness. (Art. 14 ECHR/21 Charter)

As an example of what can go wrong: the Netherlands automated the assessment of childcare allowance fraud with a predictive Machine Learning system, using nationality as a risk-indicator on the “blacklist”.

The system could automatically (no human input) request the reimbursement of childcare allowances perceived (35,000 - 250,000€).

For further reading, see: <https://taxadmin.ai/>

Statements on AI based on mistaken philosophy - 1

- **“AI develops new theories in the same way humans do.”**
 - Humans learn through formulating new *explanations* in response to *problems*. AI does no such thing. Humans are the only *universal explainers* that currently exist (as far as we know).
- **“We just need to provide AI machines with sufficient computing power and training data. If we do that, at some point the AI will become smarter than humans.”**
 - This statement is based on the (inductivist) “Bucket Theory of the Mind” theory which Popper argued against. New explanations are not generated once sufficient observations or additional computing capacity are added. A “bigger bucket” will not change this.
- **“Creativity is just connecting things” – Steve Jobs**
 - Creativity (often) involves truly new ideas, not just mix-ups of old ideas. Even if AI can nowadays make new paintings or pictures, humans ultimately remain at the “creative steering wheel”.

Statements on AI based on mistaken philosophy - 2

- **“By forcing AI to provide its reasoning step-by-step, we can trace and manage its creativity.”**
 - Producing truly new explanations is an inherently creative enterprise. This requires *disobedience* and making huge “leaps” rather than instructed step-by-step reasoning. In fact, developments in AI are in fact moving *further away* from mimicking creativity.
- **“Artificial General Intelligence will somehow emerge if we develop AI further”**
 - David Deutsch: “Expecting to create an AGI without first understanding how it works is like expecting skyscrapers to fly if we build them tall enough.”

Questions / comments?

Thank you!

Acknowledgements and source materials:

- Sam Kuypers (Université de Montréal)
- Eric Marcus (University of Amsterdam and the Netherlands Cancer Institute)
- Bart Vanderhaegen (Pactify)
- David Hadwick (Centre of Excellence DigiTax of the University of Antwerpen)
- Anouk Wolters (Deeploy)

Further reading/watching for those interested:

- TedTalks by David Deutsch ([here](#)) and ([here](#))
- Essay ([here](#)) and talk on AI ([here](#)) by David Deutsch
- “The Beginning of Infinity” - by David Deutsch
- “Conjectures and Refutations: The Growth of Scientific Knowledge”- by Karl Popper
- Paper on AI/philosophy by Eric Marcus and Max Velthoven ([here](#))
- Maria Violaris on recent Quantum Chip breakthrough and Deutsch ([here](#))
- PhD thesis by Max Velthoven featuring Popper & Deutsch and tax law ([here](#))

Max Velthoven - max.velthoven@nl.ey.com

EY | Building a better working world

EY is building a better working world by creating new value for clients, people, society and the planet, while building trust in capital markets.

Enabled by data, AI and advanced technology, EY teams help clients shape the future with confidence and develop answers for the most pressing issues of today and tomorrow.

EY teams work across a full spectrum of services in assurance, consulting, tax, strategy and transactions. Fueled by sector insights, a globally connected, multi-disciplinary network and diverse ecosystem partners, EY teams can provide services in more than 150 countries and territories.

All in to shape the future with confidence.

EY refers to the global organization, and may refer to one or more, of the member firms of Ernst & Young Global Limited, each of which is a separate legal entity. Ernst & Young Global Limited, a UK company limited by guarantee, does not provide services to clients. Information about how EY collects and uses personal data and a description of the rights individuals have under data protection legislation are available via ey.com/privacy. EY member firms do not practice law where prohibited by local laws. For more information about our organization, please visit ey.com.

© 2024 EY Belastingadviseurs B.V.
All Rights Reserved.

ED none.

This material has been prepared for general informational purposes only and is not intended to be relied upon as accounting, tax, legal or other professional advice. Please refer to your advisors for specific advice.

ey.com