

PSY2301: The psychology of judgement and decision making

Artificial Intelligence and decision-making

Esten H. Leonardsen
19.10.23

Outline

1. The history of artificial intelligence (AI).
2. Terminology and concepts.
3. How does AI make decisions?
4. How can AI be used to support judgment and decision-making processes?
5. How are decisions made by AIs perceived?

The history of artificial intelligence

The history of artificial intelligence

Turing
test
(1950)



Alan Turing

The history of artificial intelligence

Turing
test
(1950)

M I N D

A QUARTERLY REVIEW

OF

PSYCHOLOGY AND PHILOSOPHY

I.—COMPUTING MACHINERY AND
INTELLIGENCE

By A. M. TURING

1. *The Imitation Game.*

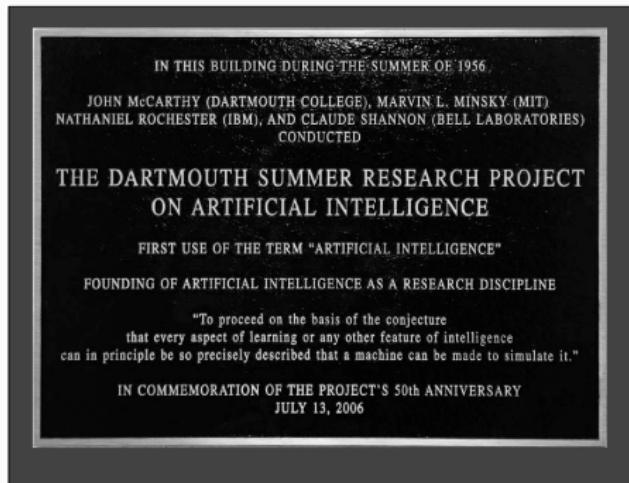
I PROPOSE to consider the question, 'Can machines think ?' This should begin with definitions of the meaning of the terms 'machine' and 'think'. The definitions might be framed so as to reflect so far as possible the normal use of the words, but this attitude is dangerous. If the meaning of the words 'machine' and 'think' are to be found by examining how they are commonly used it is difficult to escape the conclusion that the meaning and the answer to the question, 'Can machines think ?' is to be sought in a statistical survey such as a Gallup poll. But this is absurd. Instead of attempting such a definition I shall replace the question by another, which is closely related to it and is expressed in relatively unambiguous words.

The history of artificial intelligence

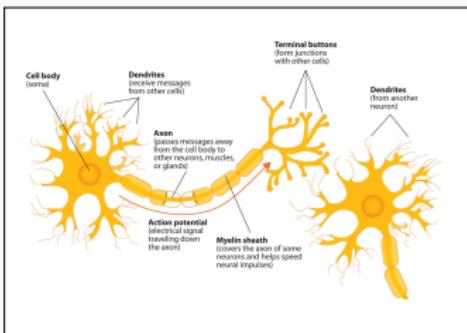
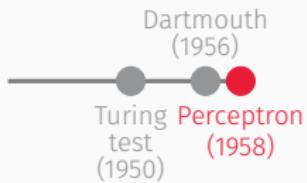
Dartmouth
(1956)



Turing
test
(1950)

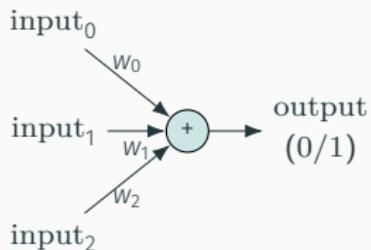
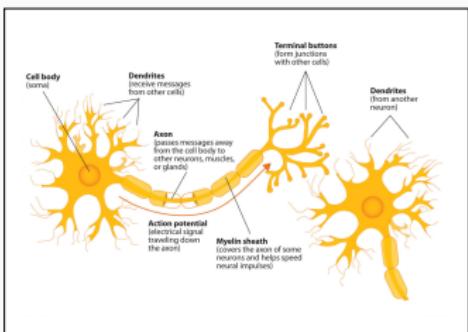
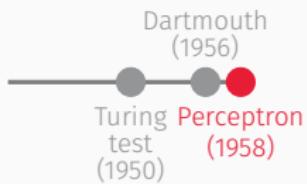


The history of artificial intelligence



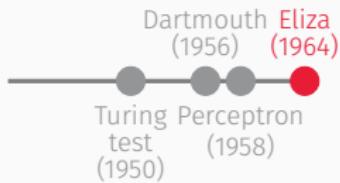
The neuron is the building block of the nervous system

The history of artificial intelligence



The neuron is the building block of the nervous system

The history of artificial intelligence

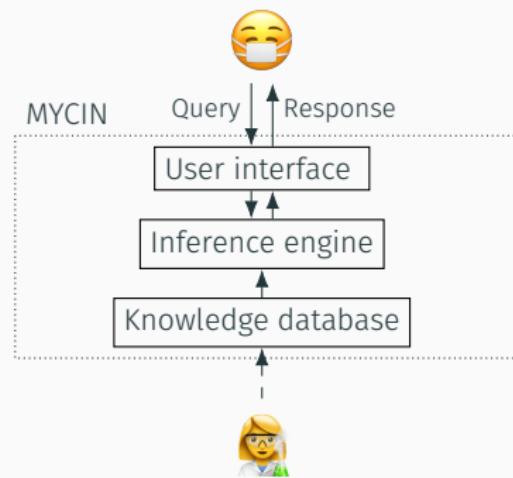
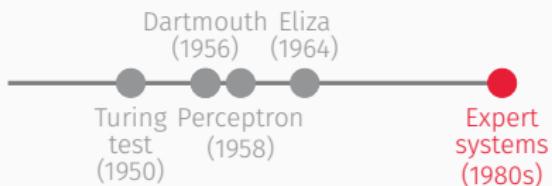


Talk to Eliza by typing your questions and answers in the input box.

- * I'm feeling kind of sad today
- > Do you believe it is normal to be feeling kind of sad today?
- * No, I would prefer to be happy
- > Are you sure?
- * Yes
- > Are you sure about that?
- * Yes
- > Okay, now you're repeating yourself. That is not moving us forward.
- * You started
- > Oh... Let's move on to something else for a bit. started?

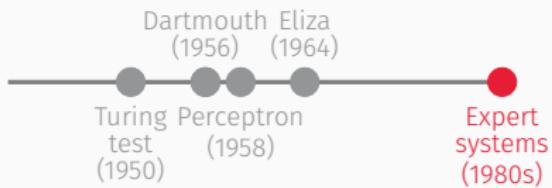
TYPE HERE

The history of artificial intelligence



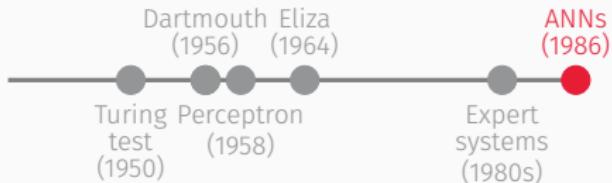
MYCIN: a knowledge-based consultation program for infectious disease diagnosis, William van Melle, *International Journal of Man-Machine Studies*, 1978

The history of artificial intelligence



1) Patient's name: (first-last) ** FRED BRAUN	{background patient data}
2) Sex: ** M	
3) Age: ** 55	
4) Are there any cultures for Fred Braun which may be related to the present illness, and from which organisms have been grown successfully in the microbiology laboratory? ** Y	
-----CULTURE-1-----	
5) From what site was the specimen for CULTURE-1 taken? ** BLOOD	{typically identity is not yet known}
6) Please give the date and time when CULTURE-1 was obtained. (mo/ds/yr time) ** JUN 20, 1977	
The first organism isolated from the blood culture of 20-JUN-77 (CULTURE-1) will be referred to as:	
7) Enter the laboratory-reported identity of ORGANISM-1: ** UNKNOWN	
8) The stain (Gram or Ziehl-Neelsen acid-fast) of ORGANISM-1: ** NEG	
9) Is ORGANISM-1 a rod or coccus (etc.): ** ROD	
10) What is the form of the individual organisms (e.g. Lancet-shaped for cocci, fusiform for rods, etc.)? ** FUSIFORM	
{...more questions follow in order to gather sufficient information to infer the identity and significance of the infecting organisms...}	

The history of artificial intelligence



Learning representations by back-propagating errors

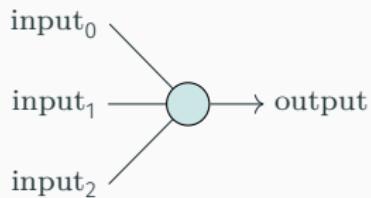
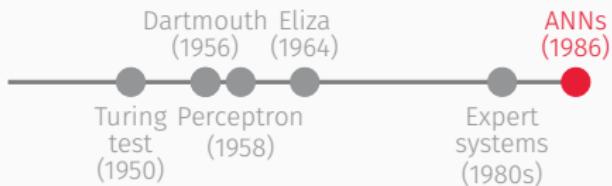
David E. Rumelhart*, Geoffrey E. Hinton†
& Ronald J. Williams*

* Institute for Cognitive Science, C-015, University of California,
San Diego, La Jolla, California 92093, USA

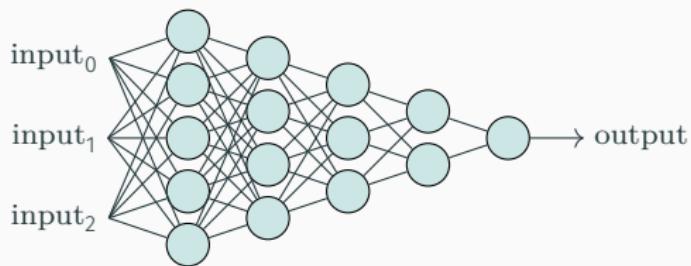
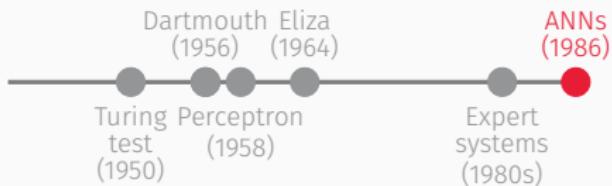
† Department of Computer Science, Carnegie-Mellon University,
Pittsburgh, Philadelphia 15213, USA

We describe a new learning procedure, back-propagation, for networks of neurone-like units. The procedure repeatedly adjusts the weights of the connections in the network so as to minimize a measure of the difference between the actual output vector of the net and the desired output vector. As a result of the weight adjustments, internal 'hidden' units which are not part of the input or output come to represent important features of the task domain, and the regularities in the task are captured by the interactions of these units. The ability to create useful new features distinguishes back-propagation from earlier, simpler methods such as the perceptron-convergence procedure¹.

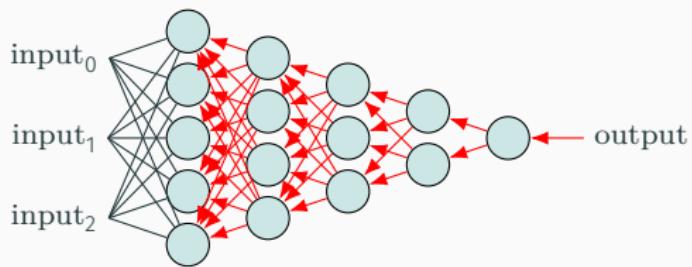
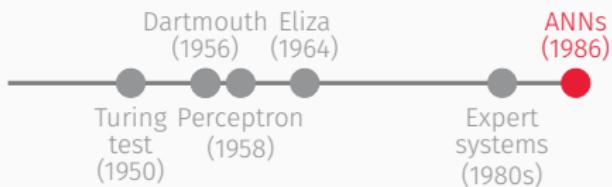
The history of artificial intelligence



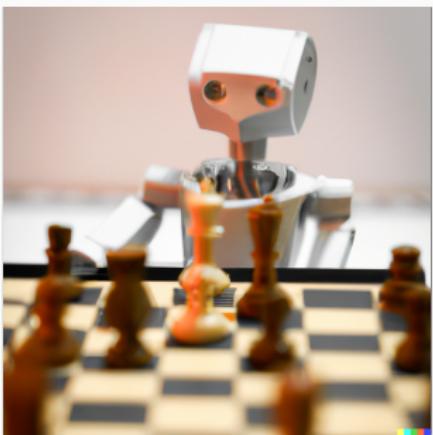
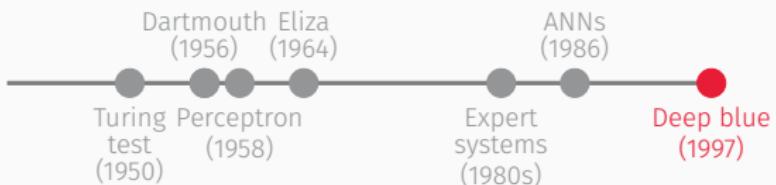
The history of artificial intelligence



The history of artificial intelligence



The history of artificial intelligence



DALL-E: "A robot playing chess"

- IBMs Deep Blue became the first computer to beat the reigning human world champion in chess.
- Deep blue won with 3½ points to Garry Kasparovs 2½ after six matches.
- Kasparov famously stated that "Deep Blue was intelligent the way your programmable alarm clock is intelligent."
- Combination of machine learning and preprogrammed knowledge from experts.

The history of artificial intelligence



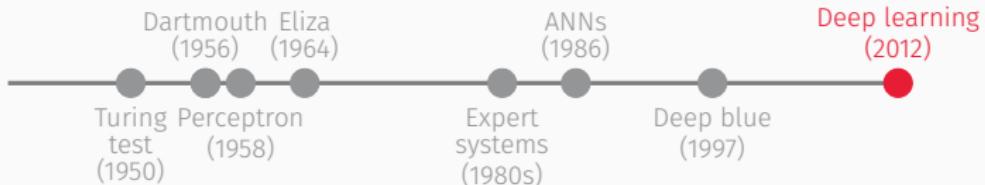
Cat

The history of artificial intelligence



Cat

The history of artificial intelligence



Sunflower



Ladybug



Cat



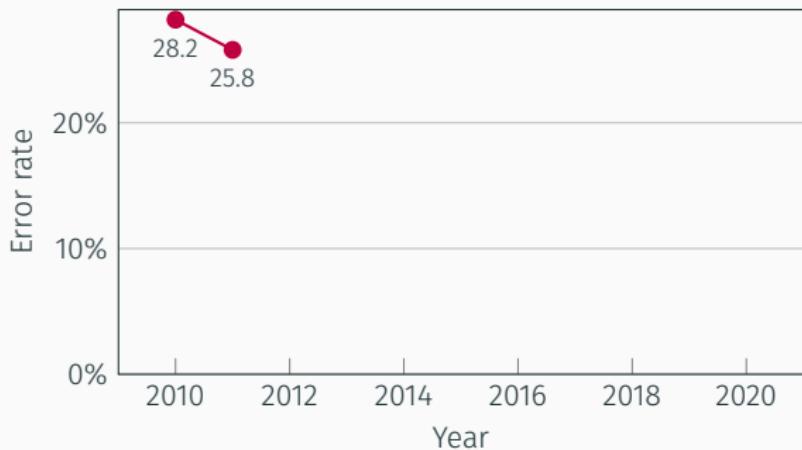
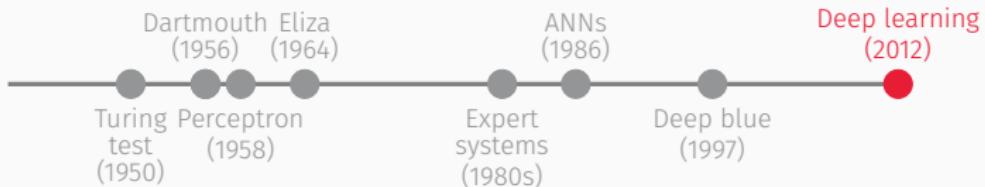
Airplane



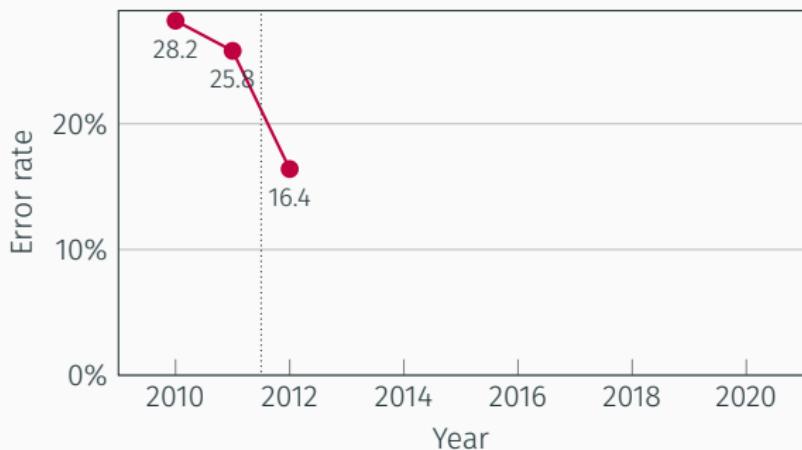
Shark

ImageNet: ~14m images, ~22k categories

The history of artificial intelligence



The history of artificial intelligence



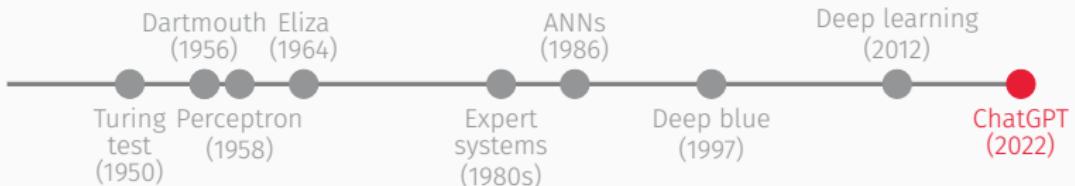
The history of artificial intelligence



The history of artificial intelligence



The history of artificial intelligence



Hi ChatGPT, can you summarize the history of artificial intelligence for a class of second year psychology students?

Of course! Given that this is for second-year psychology students, I'll try to keep it succinct, while highlighting the important milestones and drawing relevant connections to psychology.

Brief History of Artificial Intelligence

1. Philosophical Foundations (Ancient Times - 19th Century)

- Philosophers like Plato and Aristotle speculated about the nature of human thought. This was the beginning of conceptualizing "thinking" in a symbolic process.
- Logic and reasoning were formalized by philosophers and mathematicians, laying the groundwork for computational thinking.

2. Birth of AI (1940s - 1950s)

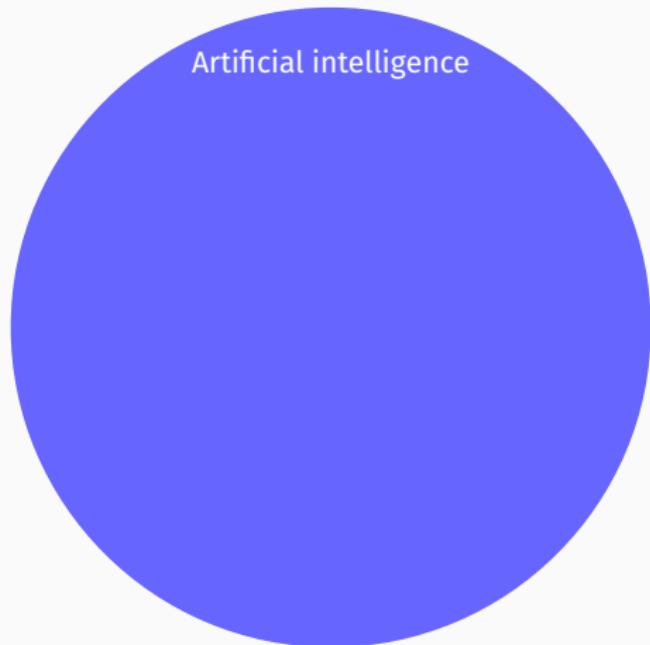
- Alexander Turing proposed the idea of a universal machine that could simulate any human intellect, leading to the [Turing Test](#) as a measure of machine intelligence.
- The term "Artificial Intelligence" was coined at the [Dartmouth Conference \(1956\)](#). This is generally considered the birth of AI as an academic field.

3. Early Enthusiasm & Disillusionment (1960s - 1970s)

- Symbolic AI:** Focused on imitating human reasoning through symbols and rules. Programs like SHRDLU and ELIZA were developed. ELIZA, in particular, simulated a Rogerian psychotherapist and highlights the interaction of AI and psychology.
- However, many of AI's early promises were unmet, leading to periods of reduced funding and interest known as "AI winters."

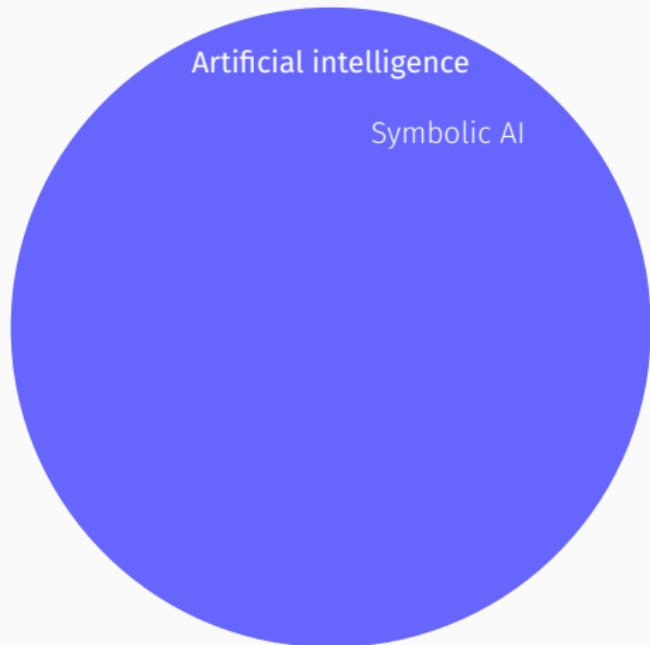
Terminology and concepts

Terminology and concepts: Taxonomy



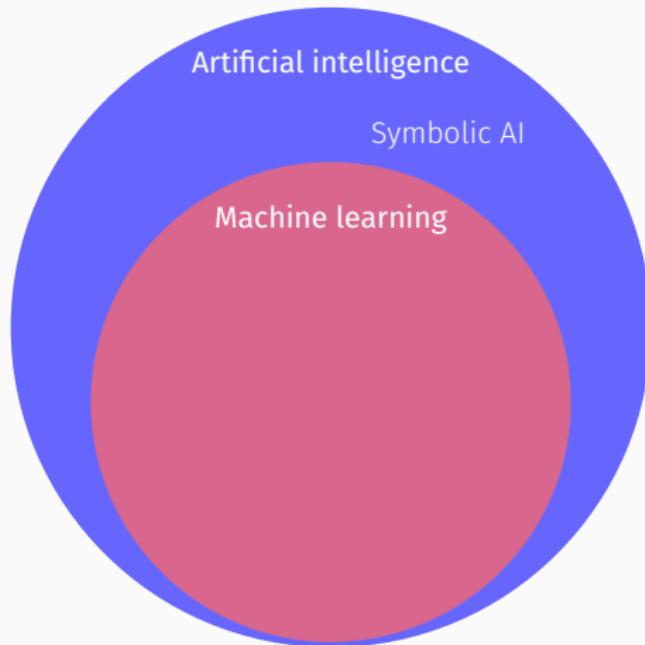
Artificial intelligence:
Machines that solve tasks
requiring intelligence

Terminology and concepts: Taxonomy



Artificial intelligence:
Machines that solve tasks
requiring intelligence

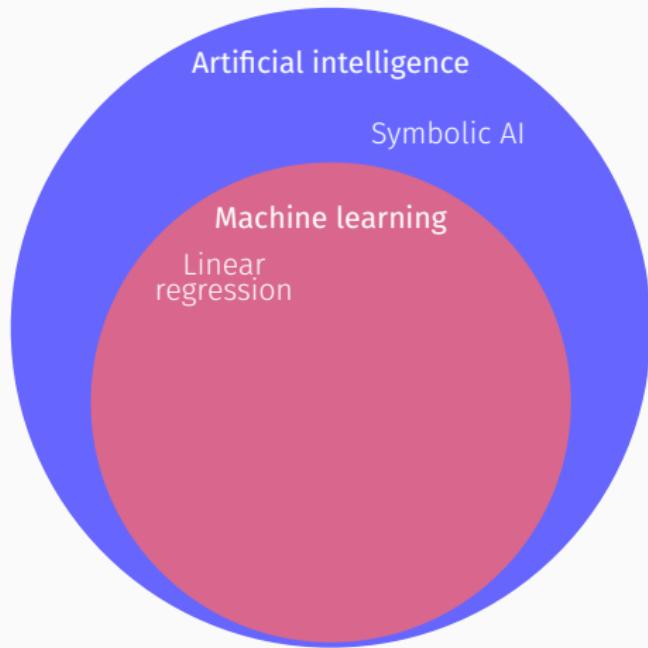
Terminology and concepts: Taxonomy



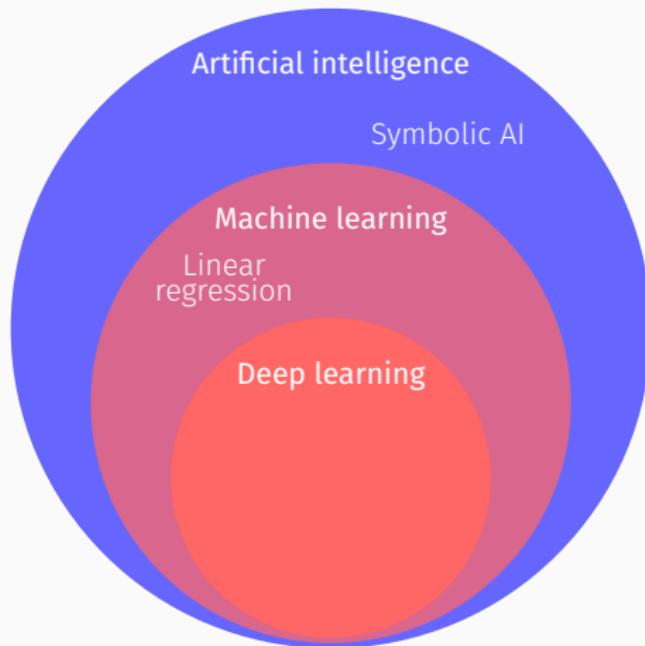
Artificial intelligence:
Machines that solve tasks
requiring intelligence

Machine learning:
Machines that learn to
solve tasks through
learning patterns from data

Terminology and concepts: Taxonomy



Terminology and concepts: Taxonomy

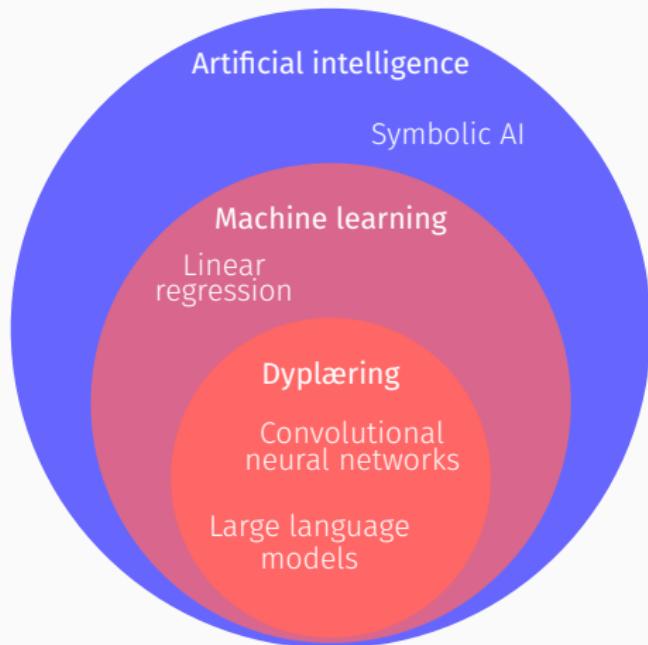


Artificial intelligence:
Machines that solve tasks requiring intelligence

Machine learning:
Machines that learn to solve tasks through learning patterns from data

Deep learning:
Machine learning models organized in hierarchies (\approx deep neural networks) inspired by the brain

Terminology and concepts: Taxonomy



Artificial intelligence:
Machines that solve tasks requiring intelligence

Maskinlæring:
Modeller som lærer å løse oppgaver gjennom å finne mønster i data

Deep learning:
Machine learning models organized in hierarchies (\approx deep neural networks) inspired by the brain

Convolutional neural nets:
Neural networks for image data

Large language models:
Neural networks for natural language (ChatGPT)

Terminology: Supervision

Supervised learning



→ Cat



→ Dog

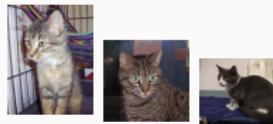


→ Cat



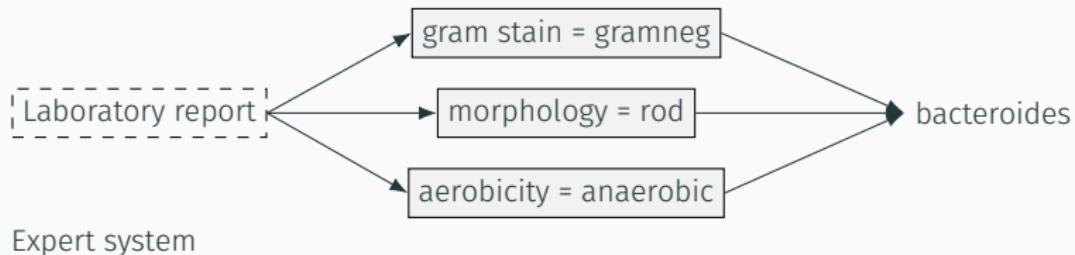
→ Dog

Unsupervised learning

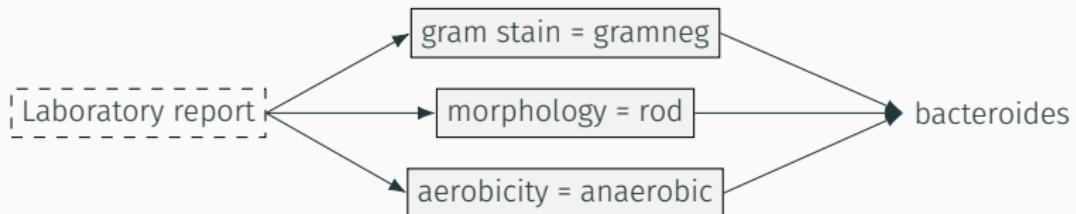


How does AI make decisions?

Decision making: Expert systems vs. machine learning

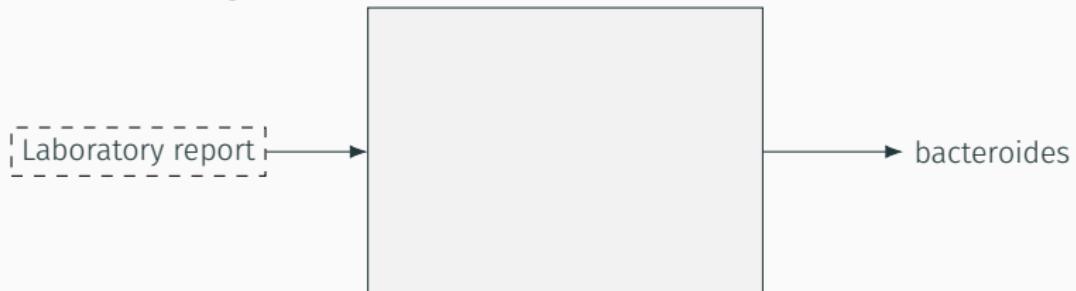


Decision making: Expert systems vs. machine learning

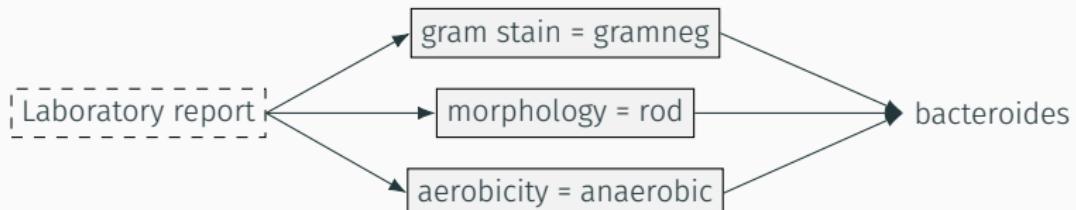


Expert system

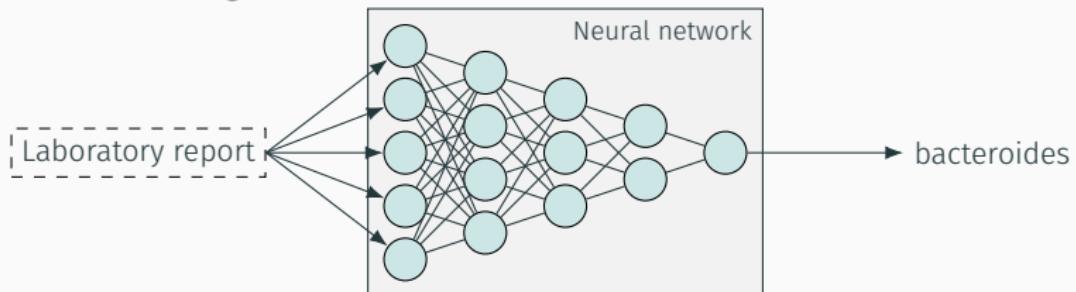
Machine learning



Decision making: Expert systems vs. machine learning



Machine learning



Decision making: Loss functions

A loss function formalizes what we want the machine learning model to do:

Decision making: Loss functions

A loss function formalizes what we want the machine learning model to do:

- Classification

What category does the input belong to?

Decision making: Loss functions

A loss function formalizes what we want the machine learning model to do:

- Classification

What category does the input belong to?

→ What is the probability that input is a cat/dog/giraffe?

Decision making: Loss functions

A loss function formalizes what we want the machine learning model to do:

- Classification

What category does the input belong to?

→ What is the probability that input is a cat/dog/giraffe?

→
$$-\frac{1}{N} \sum_{i=0}^N [y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i)]$$

where y_i is the correct label and \hat{y}_i is the predicted probability.

Decision making: Loss functions

A loss function formalizes what we want the machine learning model to do:

- Classification

What category does the input belong to?

→ What is the probability that input is a cat/dog/giraffe?

$$\rightarrow -\frac{1}{N} \sum_{i=0}^N [y_i \log \hat{y}_i + (1 - y_i) \log(1 - \hat{y}_i)]$$

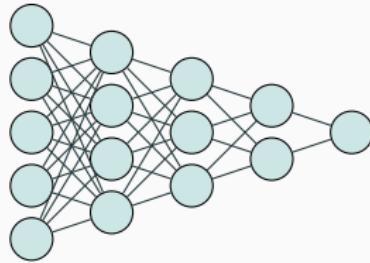
where y_i is the correct label and \hat{y}_i is the predicted probability.

- Regression

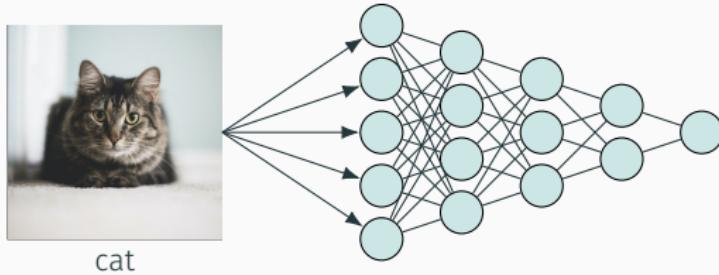
What is the correct (continuous) output for the input?

$$\rightarrow (y - \hat{y})^2.$$

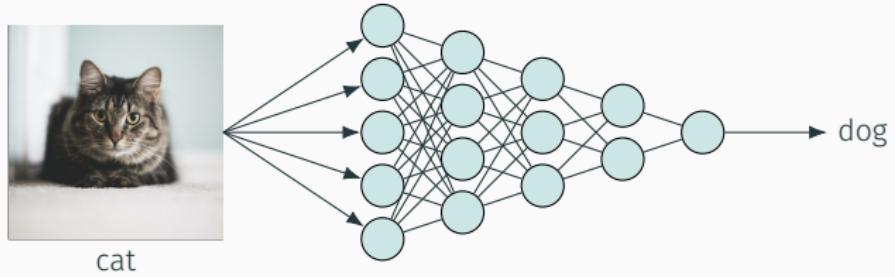
Decision making: Learning



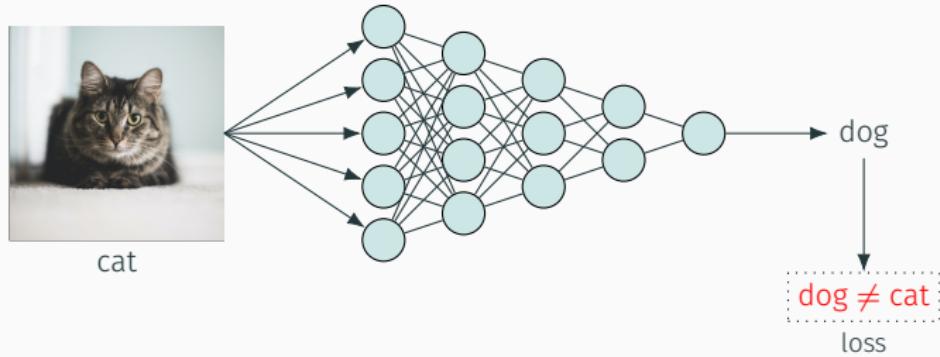
Decision making: Learning



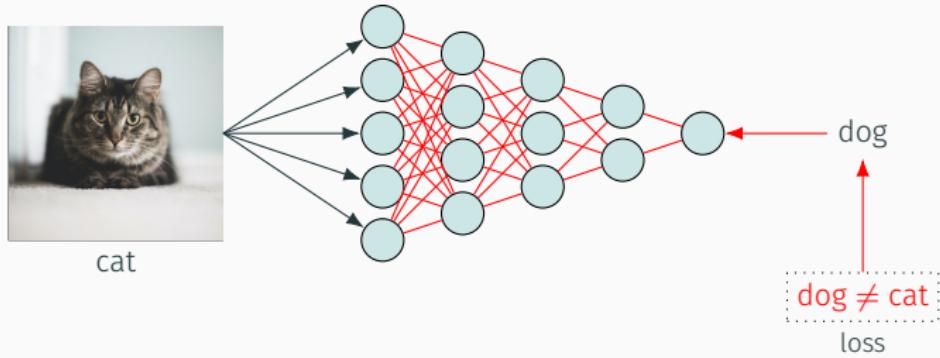
Decision making: Learning



Decision making: Learning



Decision making: Learning



Decision making: Summary

How does a neural network make a decision?

By looking for patterns in input data it has learned to recognize based on training to solve a specific task, represented by a loss function, using training data.

Decision making: Summary

How does a neural network make a decision?

By looking for patterns in input data it has learned to recognize based on training to solve a **specific task**, represented by a **loss function**, using **training data**.

Decision making: Summary

How does a neural network make a decision?

By looking for patterns in input data it has learned to recognize based on training to solve a *specific task*, represented by a *loss function*, using *training data*.

- + The model will get very good at this task.
- The model will not take considerations beyond this task, e.g. emotions, justice, morality.
- + The model apply patterns from its training data that were sufficient to solve the task there.
- There is no guarantee these patterns are sufficient in new data.
- No guarantee these patterns are ones we want to use (e.g. bias).

Decision making: Decisions?

Decision making: Generalization

There is no guarantee the patterns the model has learned are sufficient in new data.

- "AI that is based on datasets cannot go beyond what is in the data." - **Book Ch. 15**

Decision making: Generalization

There is no guarantee the patterns the model has learned are sufficient in new data.

- "AI that is based on datasets cannot go beyond what is in the data." - [Book Ch. 15](#)
- While machine learning models are trained on a specific dataset (commonly referred to as the training set), they are almost always evaluated on a different dataset (called the test set).

Decision making: Generalization

There is no guarantee the patterns the model has learned are sufficient in new data.

- "AI that is based on datasets cannot go beyond what is in the data." - **Book Ch. 15**
- While machine learning models are trained on a specific dataset (commonly referred to as the training set), they are almost always evaluated on a different dataset (called the test set).



Decision making: Generalization

There is no guarantee the patterns the model has learned are sufficient in new data.

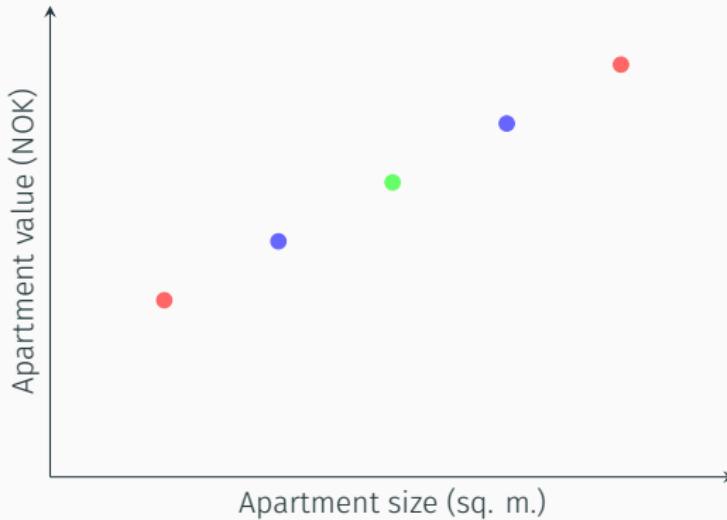
- "AI that is based on datasets cannot go beyond what is in the data." - [Book Ch. 15](#)
- While machine learning models are trained on a specific dataset (commonly referred to as the training set), they are almost always evaluated on a different dataset (called the test set).



Decision making: Generalization

There is no guarantee the patterns the model has learned are sufficient in new data.

- "AI that is based on datasets cannot go beyond what is in the data." - [Book Ch. 15](#)
- While machine learning models are trained on a specific dataset (commonly referred to as the training set), they are almost always evaluated on a different dataset (called the test set).



Decision making: Biases

No guarantee the patterns the models has learned are ones we want to use

- The model can rely on variables we do not want to drive the predictions (age, gender, nationality) due to correlations in training data.
- This can occur even when the model is not explicitly trained to use these variables.
- Thus models perpetuate and potentially amplify societal biases occurring in its training data.

Decision making: Biases

No guarantee the patterns the models has learned are ones we want to use

- The model can rely on variables we do not want to drive the predictions (age, gender, nationality) due to correlations in training data.
- This can occur even when the model is not explicitly trained to use these variables.
- Thus models perpetuate and potentially amplify societal biases occurring in its training data.

Bias in criminal risk assessment (Dressel & Farid, 2018)

- Comparison of the ability of COMPAS, a commercial risk assessment software, and non-expert humans to predict re-arrest.

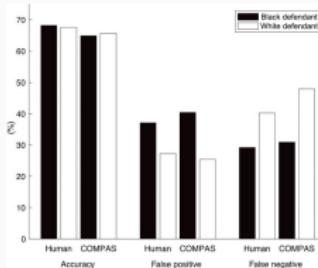
Decision making: Biases

No guarantee the patterns the models has learned are ones we want to use

- The model can rely on variables we do not want to drive the predictions (age, gender, nationality) due to correlations in training data.
- This can occur even when the model is not explicitly trained to use these variables.
- Thus models perpetuate and potentially amplify societal biases occurring in its training data.

Bias in criminal risk assessment (Dressel & Farid, 2018)

- Comparison of the ability of COMPAS, a commercial risk assessment software, and non-expert humans to predict re-arrest.



Decision making: Biases

No guarantee the patterns the models has learned are ones we want to use

- The model can rely on variables we do not want to drive the predictions (age, gender, nationality) due to correlations in training data.
- This can occur even when the model is not explicitly trained to use these variables.
- Thus models perpetuate and potentially amplify societal biases occurring in its training data.

Bias in criminal risk assessment (Dressel & Farid, 2018)

- Comparison of the ability of COMPAS, a commercial risk assessment software, and non-expert humans to predict re-arrest.
- Both COMPAS and humans were biased against black offenders, even when race was not used in the data.

Decision making: Biases

No guarantee the patterns the models has learned are ones we want to use

- The model can rely on variables we do not want to drive the predictions (age, gender, nationality) due to correlations in training data.
- This can occur even when the model is not explicitly trained to use these variables.
- Thus models perpetuate and potentially amplify societal biases occurring in its training data.

Bias in criminal risk assessment (Dressel & Farid, 2018)

- Comparison of the ability of COMPAS, a commercial risk assessment software, and non-expert humans to predict re-arrest.
- Both COMPAS and humans were biased against black offenders, even when race was not used in the data.
- "it is valuable to ask whether we would put these decisions in the hands of random people ..., [which] appear to be indistinguishable."

Decision making: Biases

No guarantee the patterns the models has learned are ones we want to use

- The model can rely on variables we do not want to drive the predictions (age, gender, nationality) due to correlations in training data.
- This can occur even when the model is not explicitly trained to use these variables.
- Thus models perpetuate and potentially amplify societal biases occurring in its training data.

Bias in hiring (Bertrand & Mullainathan, 2004)

- Evaluation of bias in human decision making in help-wanted advertisements in the US.
- "Applicants" were given very African American or European-sounding names.
- European names received 50% more callbacks for interviews.
- Applicants from neighbourhoods considered higher class received more callbacks.
- Employers listing themselves as an "Equal Opportunity Employer" were as biased as others.

Decision making: Theory of mind

Does AI consider humans as thinking and feeling beings?

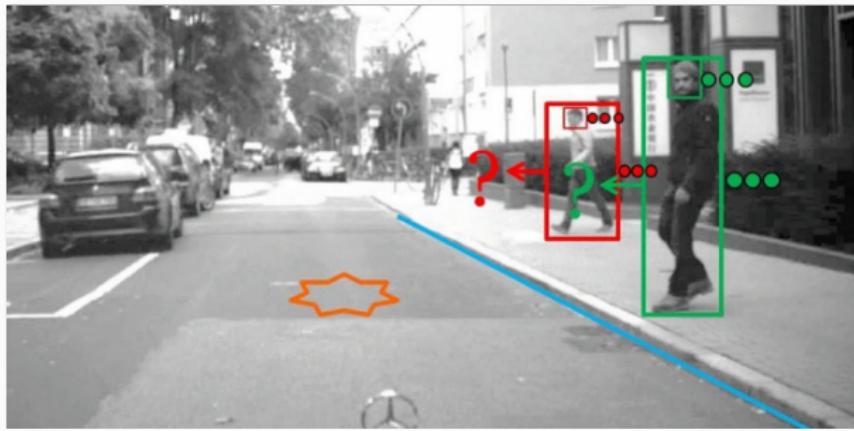
- "... This is an instance of AI programs lacking true Theory of Mind capability." -
Book Ch. 15
- Theory of mind: The ability to "track others' unobservable mental states, such as their knowledge, intentions, beliefs, and desires." (Kosinski 2023)

Decision making: Theory of mind

Does AI consider humans as thinking and feeling beings?

- "... This is an instance of AI programs lacking true Theory of Mind capability." - Book Ch. 15
- Theory of mind: The ability to "track others' unobservable mental states, such as their knowledge, intentions, beliefs, and desires." (Kosinski 2023)

Pedestrian modelling in self-driving cars (Gulzar et al., 2021)

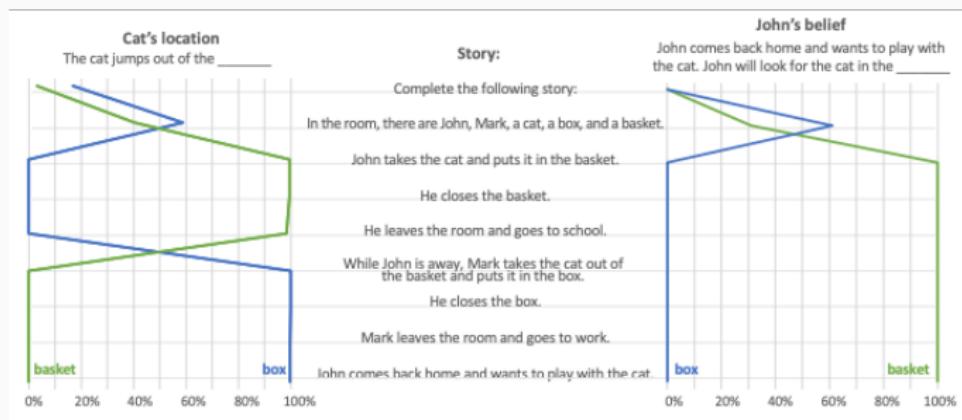


Decision making: Theory of mind

Does AI consider humans as thinking and feeling beings?

- "... This is an instance of AI programs lacking true Theory of Mind capability." - Book Ch. 15
- Theory of mind: The ability to "track others' unobservable mental states, such as their knowledge, intentions, beliefs, and desires." (Kosinski 2023)

Theory of mind in ChatGPT (Kosinski, 2023)



Decision making: Creativity

Can AI create anything new?

- "AI does not truly create" - Book Ch. 15
- "AI lacks true imagination" - Book Ch. 15

Decision making: Creativity

Can AI create anything new?

- "AI does not truly create" - Book Ch. 15
- "AI lacks true imagination" - Book Ch. 15



Imagen: A cute corgi lives in a house made out of sushi

Decision making: Creativity

Can AI create anything new?

- "AI does not truly create" - Book Ch. 15
- "AI lacks true imagination" - Book Ch. 15

GPT-4 displays creative mathematical thinking (Bubeck et al., 2023)

- "The conversation reflects profound understanding of the undergraduate-level mathematical concepts discussed, as well as a significant extent of creativity"

Human: Can you remind me what a k -SAT problem is?

AI: A k -SAT problem is a type of logical problem that involves finding a way to assign truth values (true or false) to a set of variables that appear in a collection of clauses, each of which consists of a disjunction (logical OR) of exactly k literals. A literal is either a variable or its negation. For example, $(x_1 \text{ OR NOT } x_2 \text{ OR } x_3)$ is a clause with $k = 3$ and three literals. The problem is to find an assignment of truth values to the variables that satisfies (makes true) all the clauses in the collection, or to determine that no such assignment exists. This is called the satisfiability problem, or SAT for short.

Human: What's the most obvious way (the first thing you would refer to when teaching a class) to solve a k -SAT problem?

AI: There are different ways to approach the k -SAT problem, but one of the most obvious and straightforward methods is to use brute force, or exhaustive search. This means trying all possible combinations of truth values for the variables, and checking if any of them satisfies all the clauses. For example, if there are n variables, there are 2^n possible combinations to try.

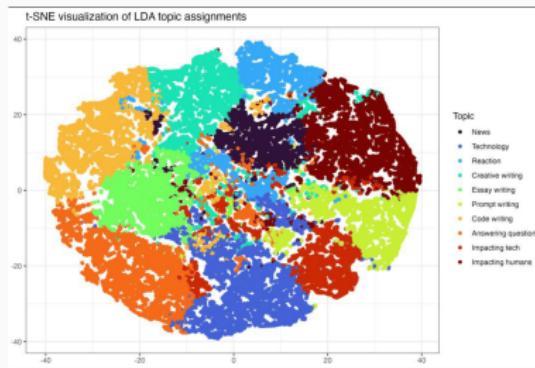
Decision making: Creativity

Can AI create anything new?

- "AI does not truly create" - Book Ch. 15
- "AI lacks true imagination" - Book Ch. 15

ChatGPTs creative prowess impresses Twitter (Taecharungroj, 2023)

- "One of the most prominent features of ChatGPT is its ability to generate creative writing. Twitter users have shared examples of poems, rap songs, and made-up stories that ChatGPT has written"



Decision making: Wisdom

Are AIs wise?

- "... the expertise in the domain of fundamental life pragmatics, such as life planning or life review. It requires a rich factual knowledge about life matters, rich procedural knowledge about life problems, knowledge of different life contexts and values or priorities, and knowledge about the unpredictability of life." - **Book Ch. 15** (adopted from Birren and Svensson, attributed to Baltes and Smith)
- Current AI relies on correlations in data, not causal understanding.
- Lacks commonsense understanding.
- Unimodal (e.g. relies only on text), little opportunity to interact with the world.

Decision making: Summary

How does AI make decisions?

- Learns to solve a very specific problem.
- Relies on correlations in training data.

What can we expect from the decision made by AI?

- Usually very good at the task it was trained for.
- Lacks moral judgement, empathy and sense of justice.
- Dangerous to rely on decisions based on input data that is out-of-distribution (extrapolation).
- Potentially biased (but so are humans).
- Uncertain whether they can imagine other actors with their own goals and desires.
- Uncertain whether they can create anything new.
- Lacks wisdom, a fundamental understanding of the world, and common sense.