PSY2301: Psychology of judgement and decision-making

Artificial Intelligence and decisions

Esten H. Leonardsen

10.11.25

UNIVERSITY
OF OSLO

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

UNIVERSITY
OF OSLO

The history of artificial intelligence

Talos
(∼300 BC)

Image source: ChatGPT 5

The history of artificial intelligence

Talos
(∼300 BC)

Golem
(∼1600)

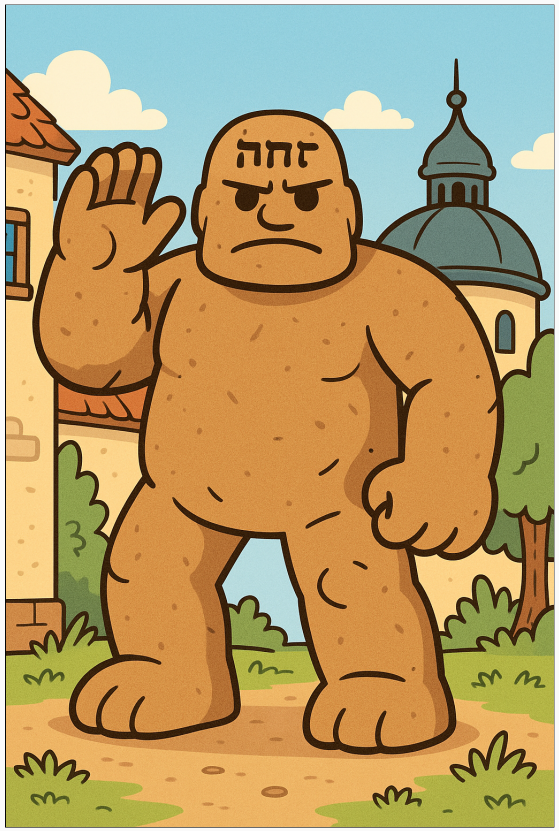


Image source: ChatGPT 5

The history of artificial intelligence

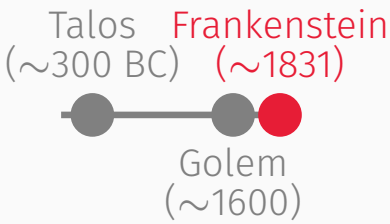
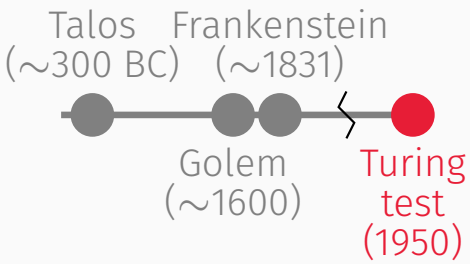


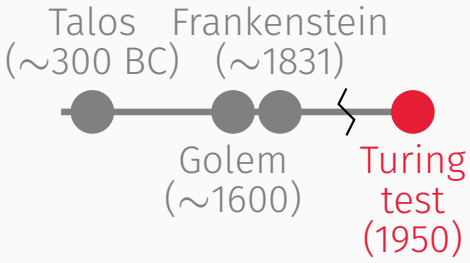
Image source: ChatGPT 5

The history of artificial intelligence

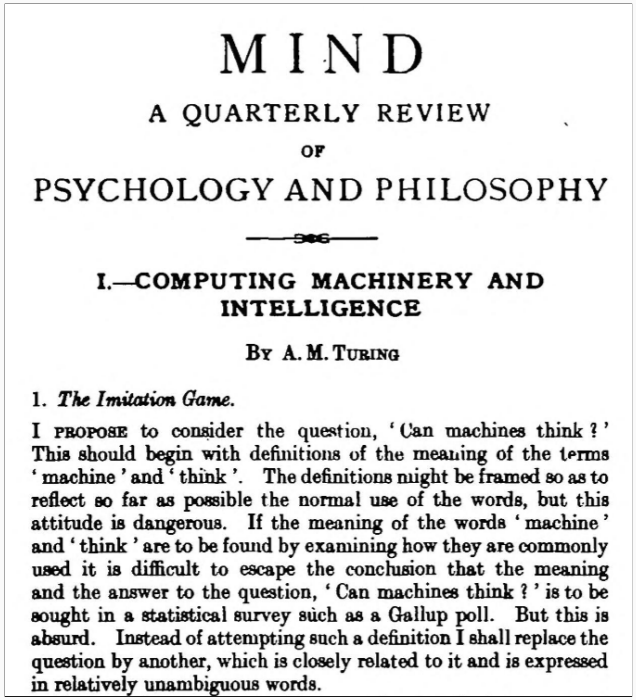


Alan Turing

The history of artificial intelligence



Alan Turing



MIND

A QUARTERLY REVIEW

OF

PSYCHOLOGY AND PHILOSOPHY

I.—COMPUTING MACHINERY AND INTELLIGENCE

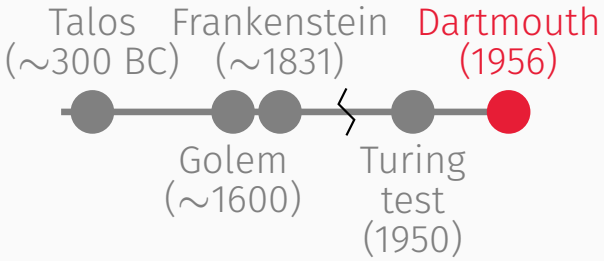
By A. M. Turing

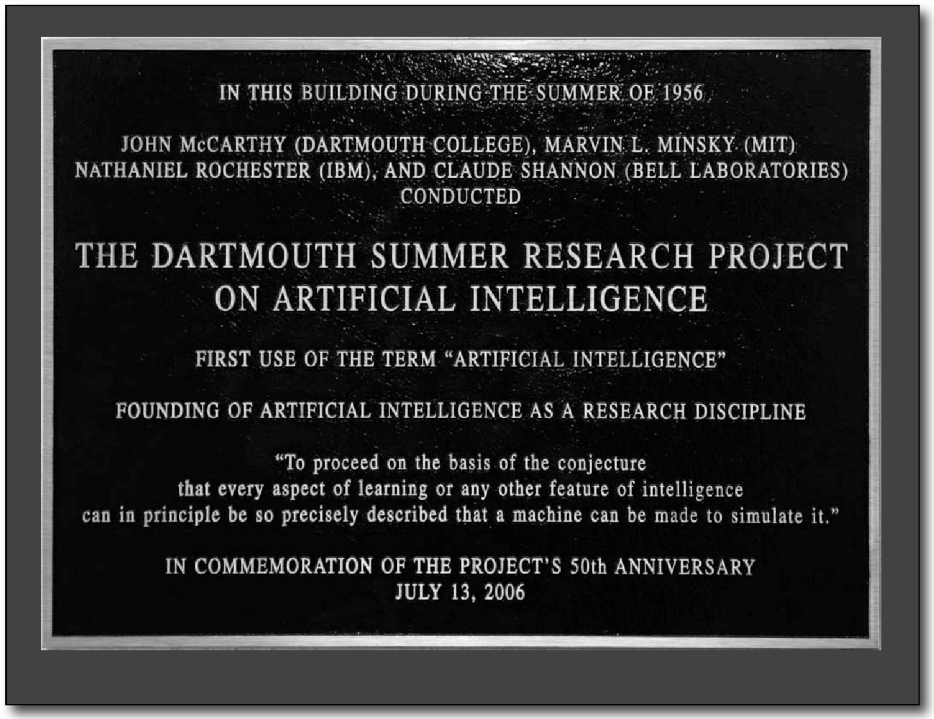
1. *The Imitation Game.*

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.

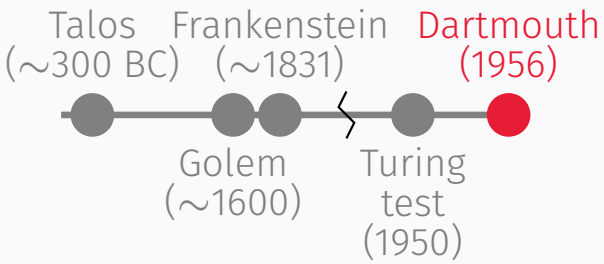
Computing Machinery and Intelligence, A. M. Turing, *Mind*, 1950

The history of artificial intelligence





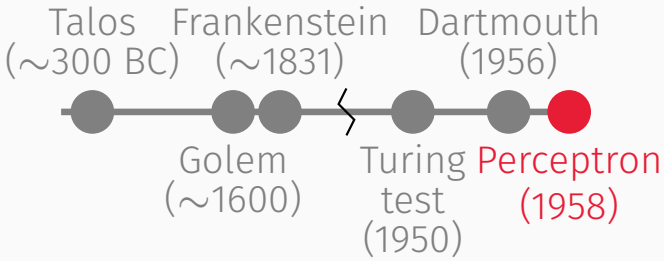
The history of artificial intelligence

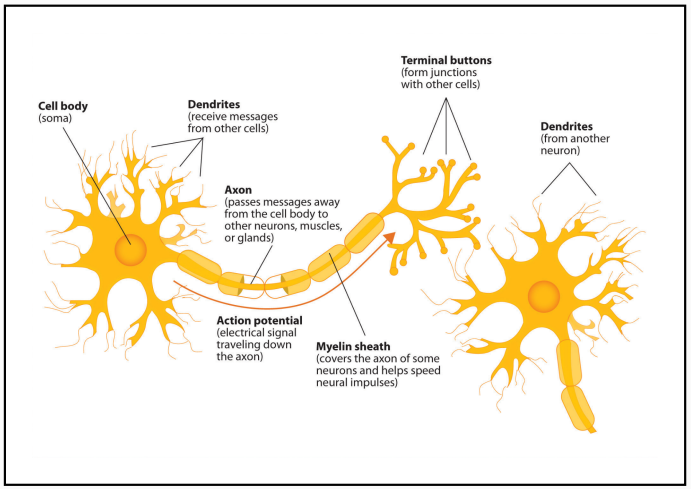


”We propose that a 2-month, 10-man study of artificial intelligence be carried out [...]. An attempt will be made to find how to make machines use language, form abstractions and concepts, solve kinds of problems now reserved for humans, and improve themselves. We think that a significant advance can be made in [...] a summer.”

- Proposal, Dartmouth summer school (1956)

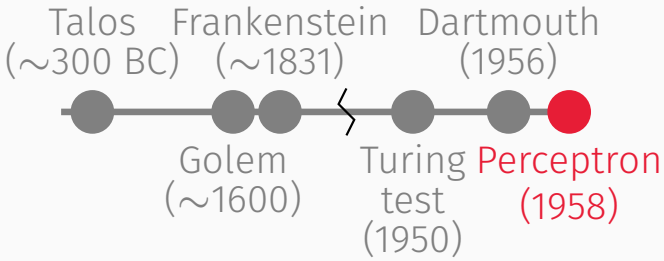
The history of artificial intelligence

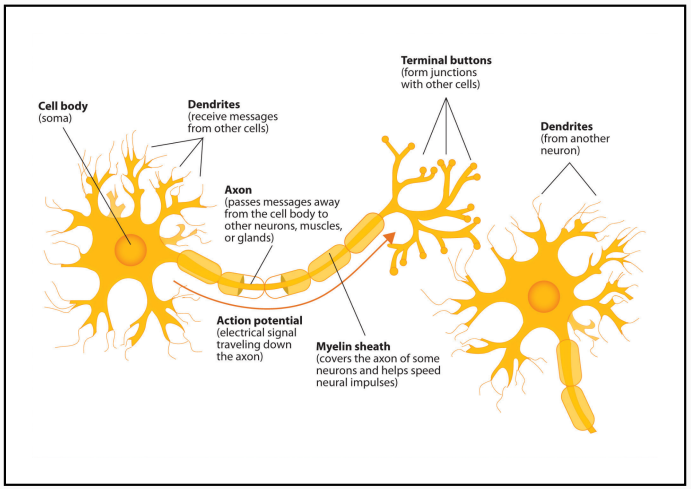


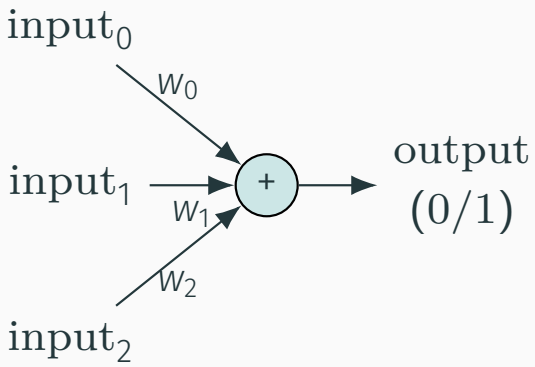


The neuron is the building block of the nervous system, Stangor, C. and Walinga, J, *Introduction to psychology*, 2024

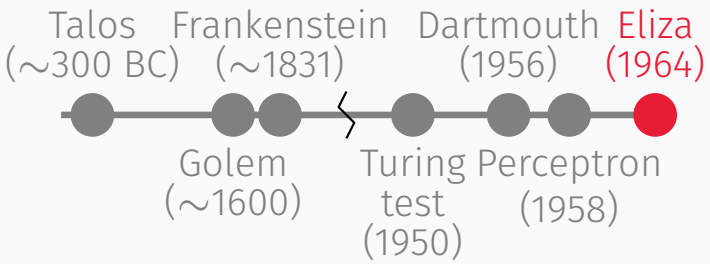
The history of artificial intelligence

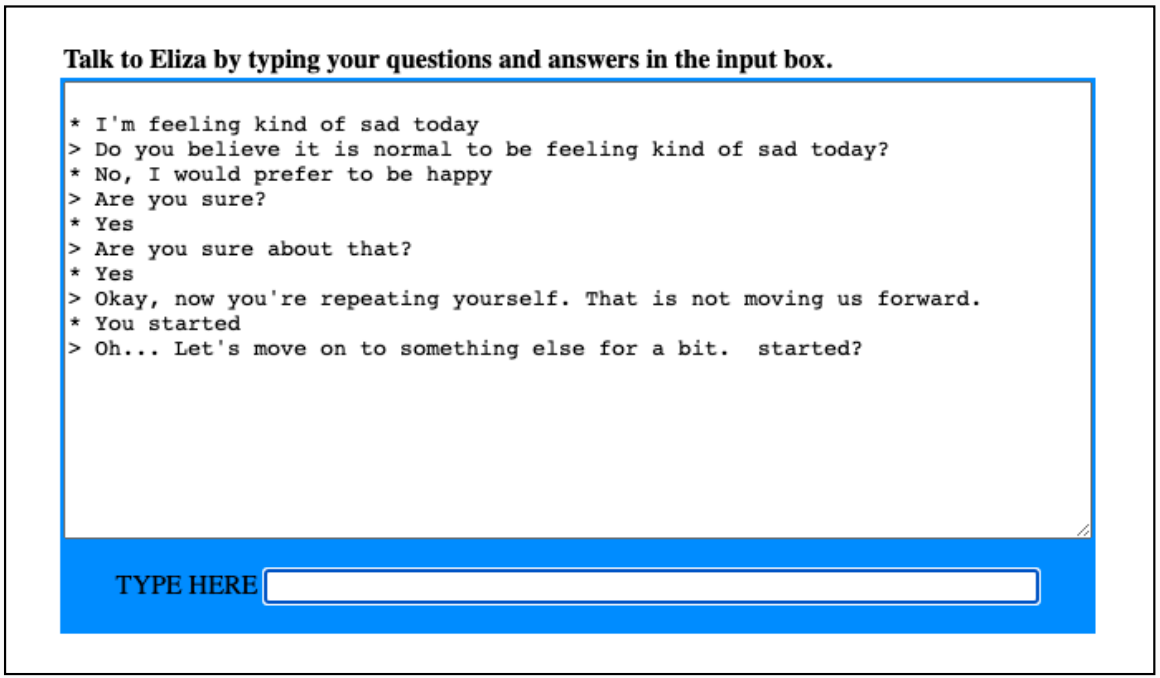






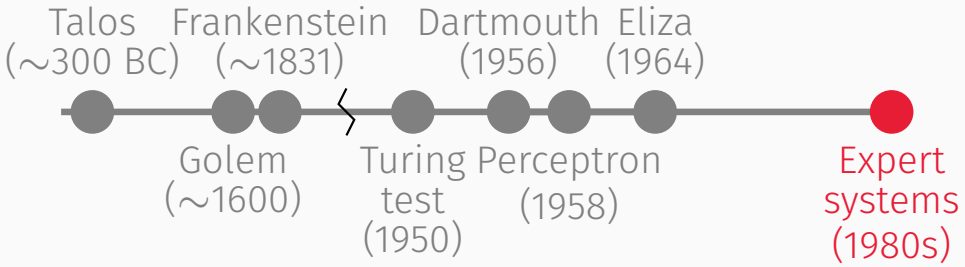
The history of artificial intelligence

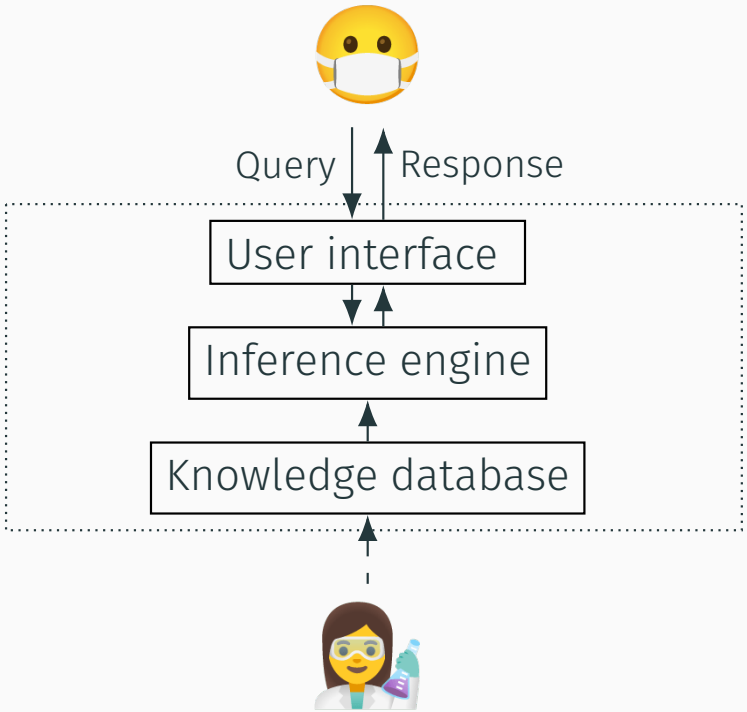




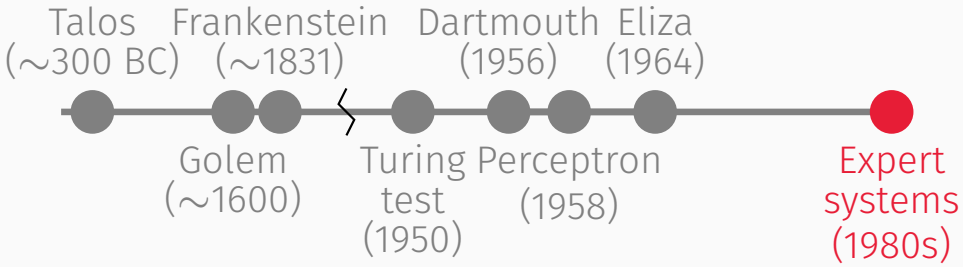
ELIZA: a very basic Rogerian psychotherapist chatbot

The history of artificial intelligence





The history of artificial intelligence



1) Patient's name: (first-last)
** *FRED BRAUN*
2) Sex:
** *M* *{background patient data}*
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*

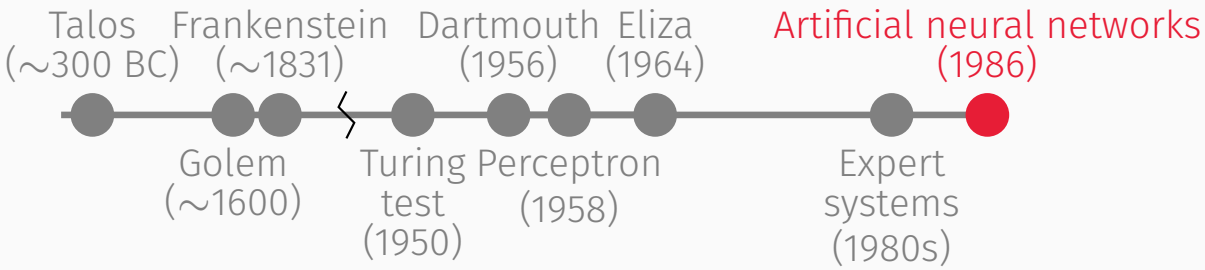
--------CULTURE-1--------
5) From what site was the specimen for CULTURE-1 taken?
** *BLOD*
=BLOOD
6) Please give the date and time when CULTURE-1 was obtained. (mo/da/yr time)
** *JUN 20, 1977*
The first organism isolated from the blood culture of 20-JUN-77 (CULTURE-1) will be
referred to as:

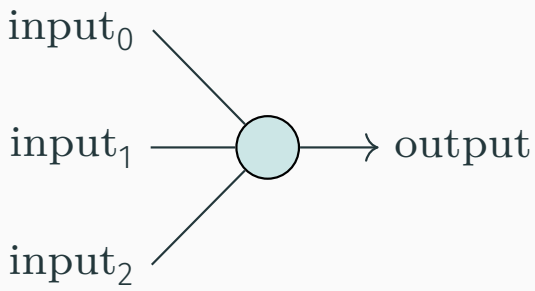
--------ORGANISM-1--------
7) Enter the laboratory-reported identity of ORGANISM-1:
** *UNKNOWN* *{typically identity is not yet known}*
8) The stain (Gram or Ziehl-Neelsen acid-fast) of ORGANISM-1:
** *NEG* *{preliminary lab results give some clues}*
9) Is ORGANISM-1 a rod or coccus (etc.):
** *ROD*
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...}*

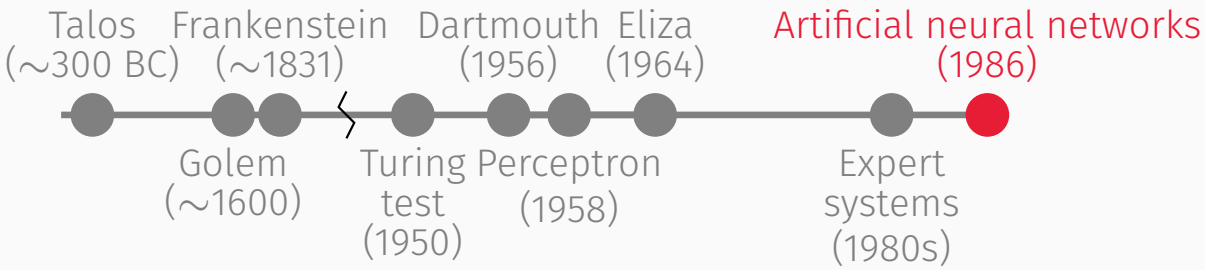
MYCIN, William van Melle, *International Journal of Man-Machine Studies*, 1978

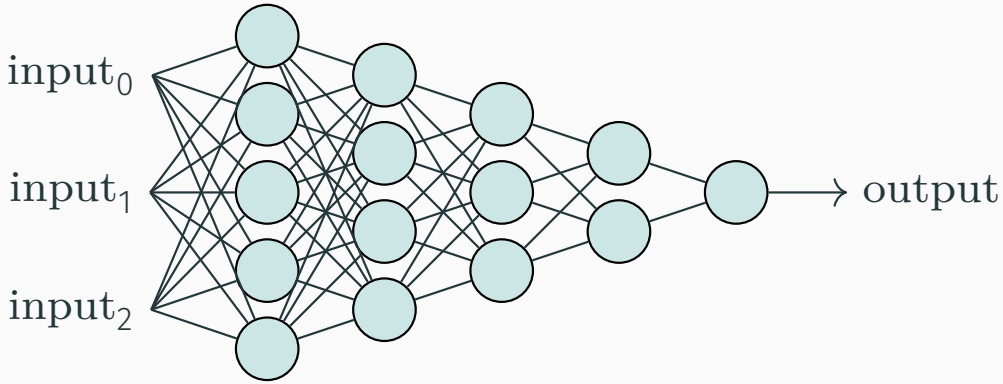
The history of artificial intelligence



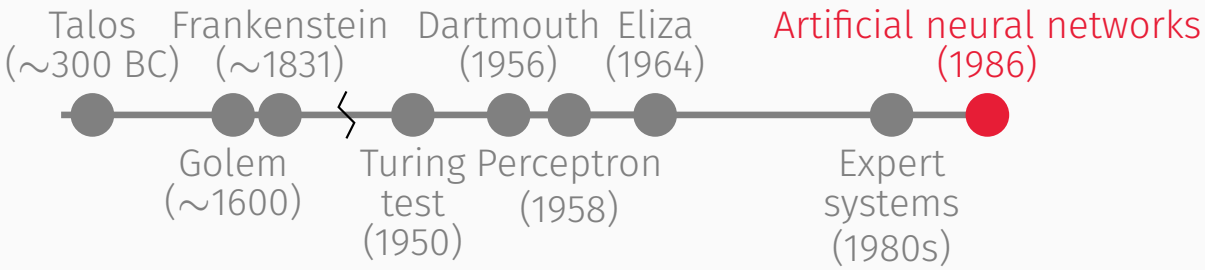


The history of artificial intelligence





The history of artificial intelligence



Learning representations by back-propagating errors

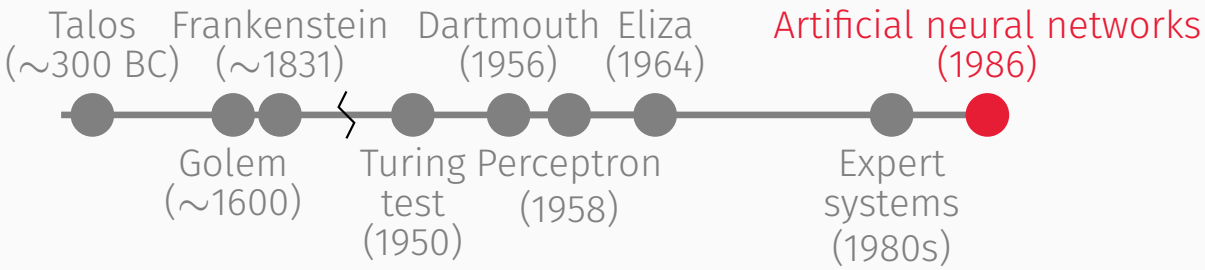
David E. Rumelhart*, Geoffrey E. Hinton† & Ronald J. Williams*

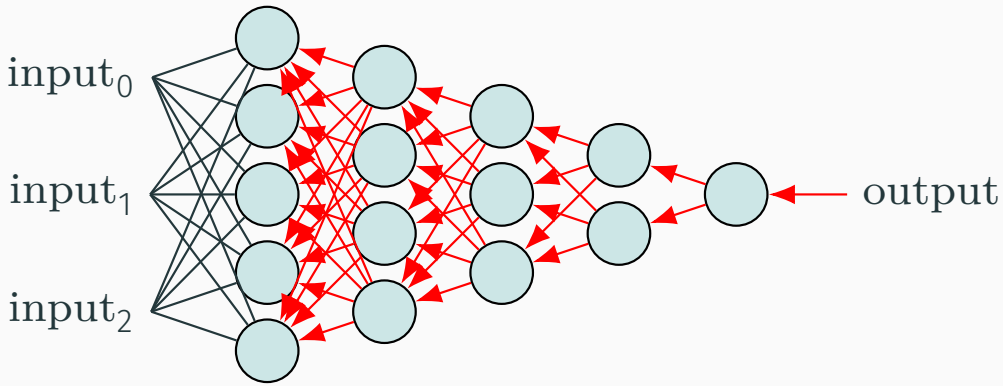
* Institute for Cognitive Science, C-015, University of California, San Diego, La Jolla, California 92093, USA
† Department of Computer Science, Carnegie-Mellon University, Pittsburgh, Philadelphia 15213, USA

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 procedure[1].

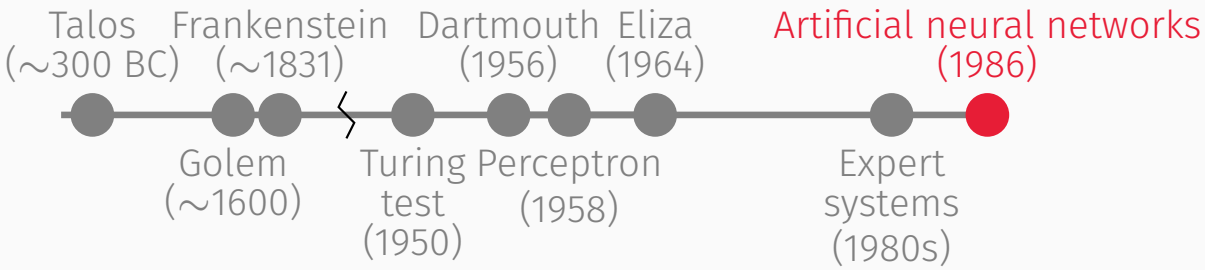
Learning representations by back-propagating errors, Rumelhart, D. et al., *Nature*, 1986

The history of artificial intelligence





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 procedure[1].

The history of artificial intelligence

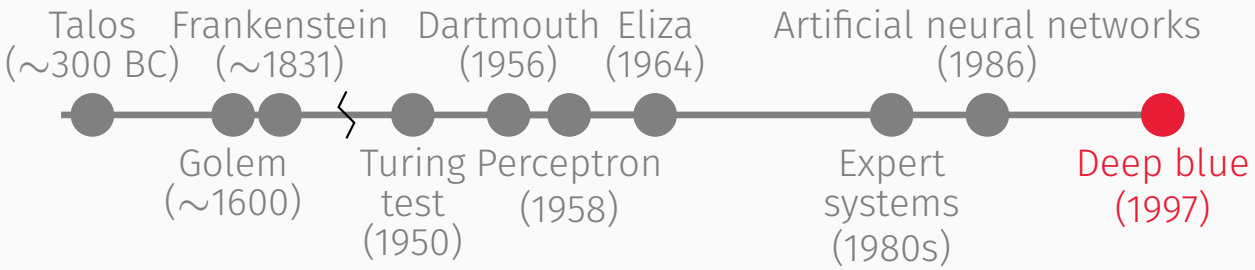
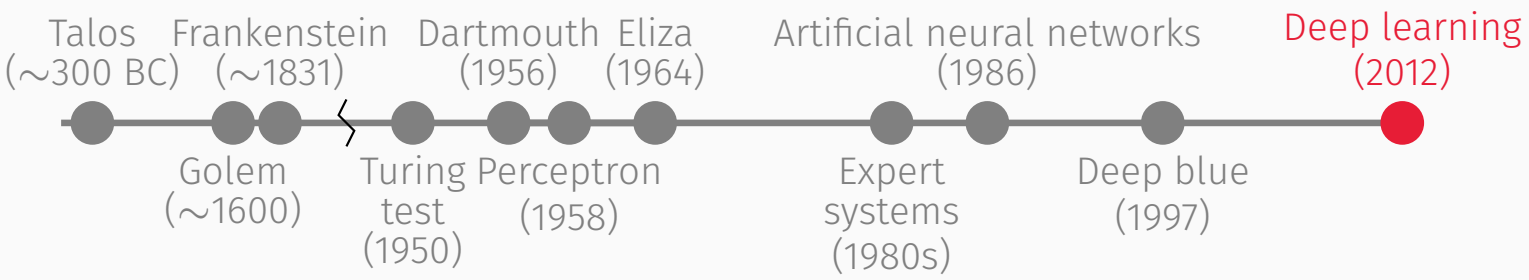


Image source: DALL-E

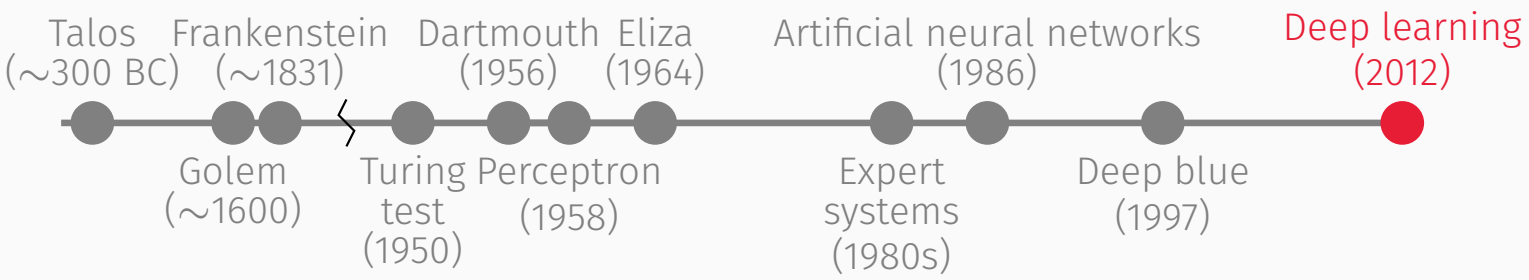
- IBMs Deep Blue became the first computer to beat the reigning human world champion in chess.
- Deep blue won with 3½ points to Garry Kasparovs 2½ after six matches.
- Kasparov famously stated that "Deep Blue was intelligent the way your programmable alarm clock is intelligent."

The history of artificial intelligence



Cat

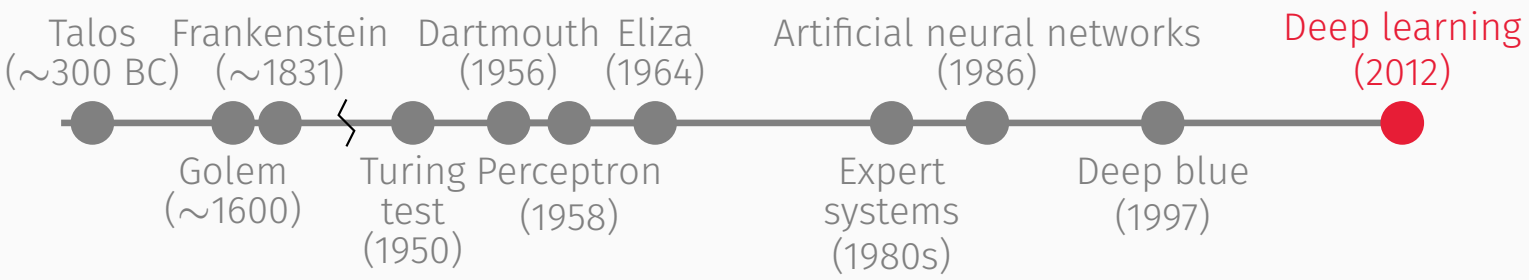
The history of artificial intelligence

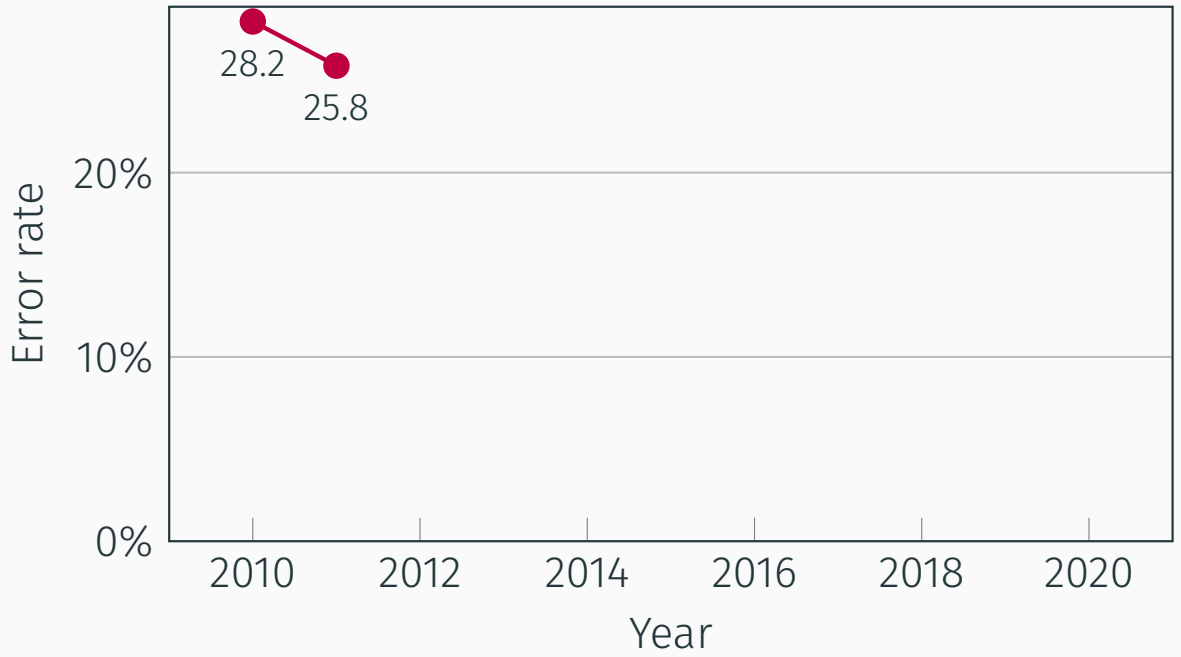


Sunflower Ladybug Cat Airplane Shark

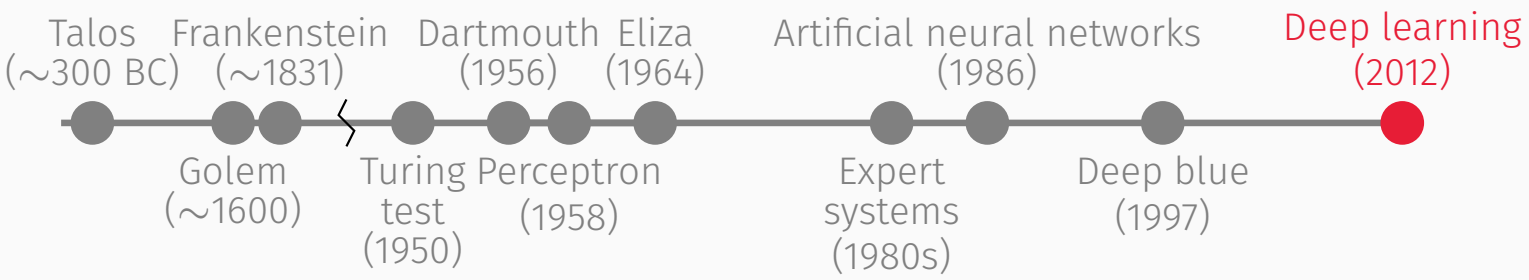
ImageNet: ∼14m images, ∼22k categories

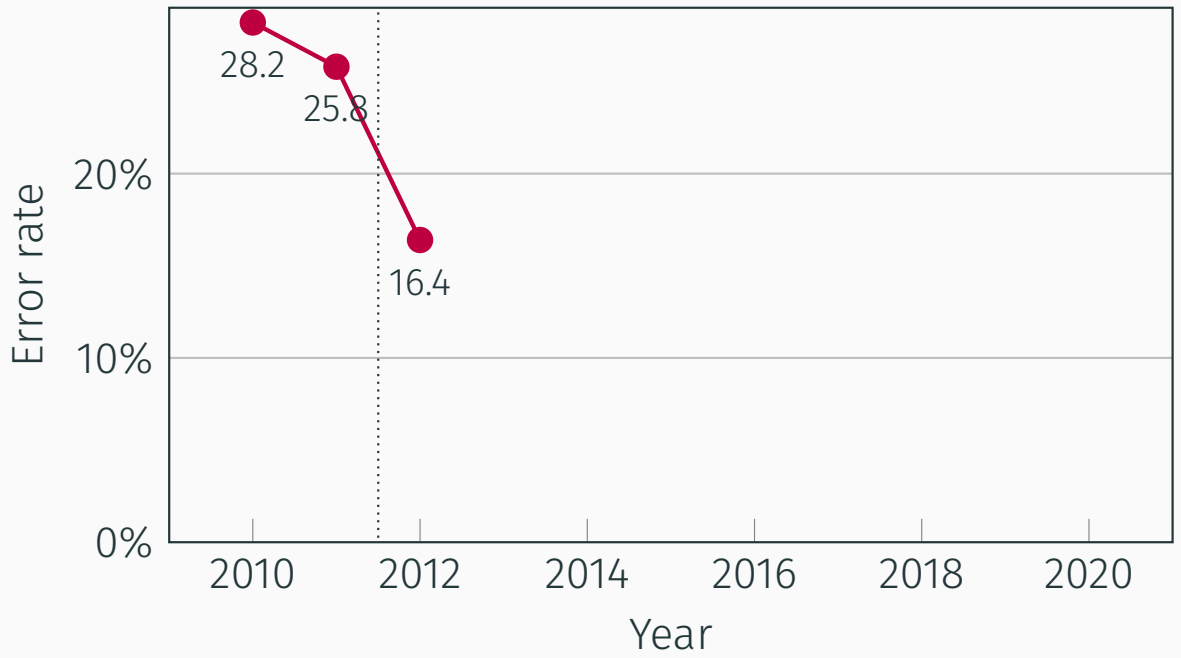
The history of artificial intelligence



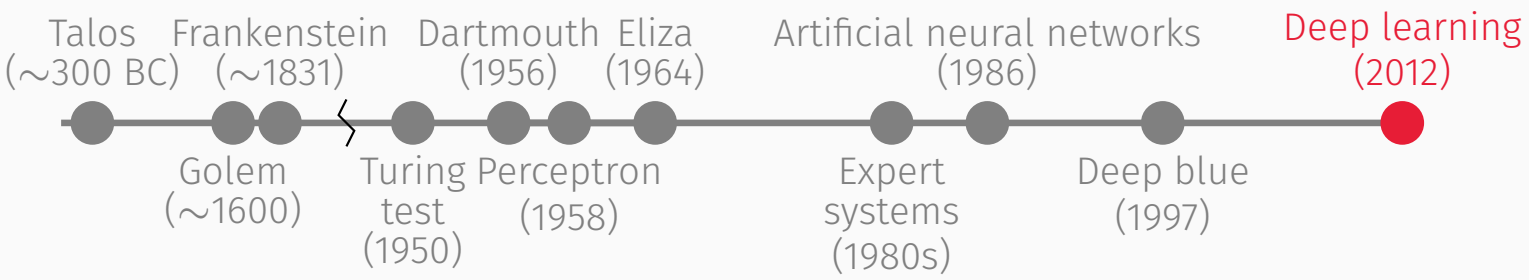


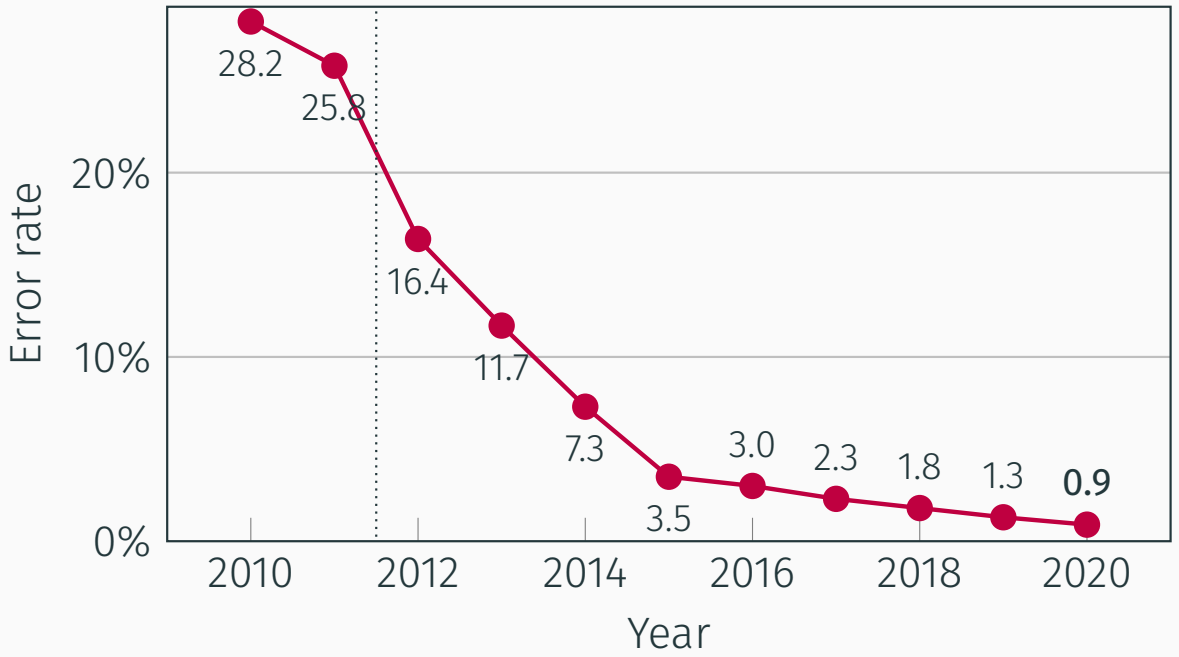
The history of artificial intelligence



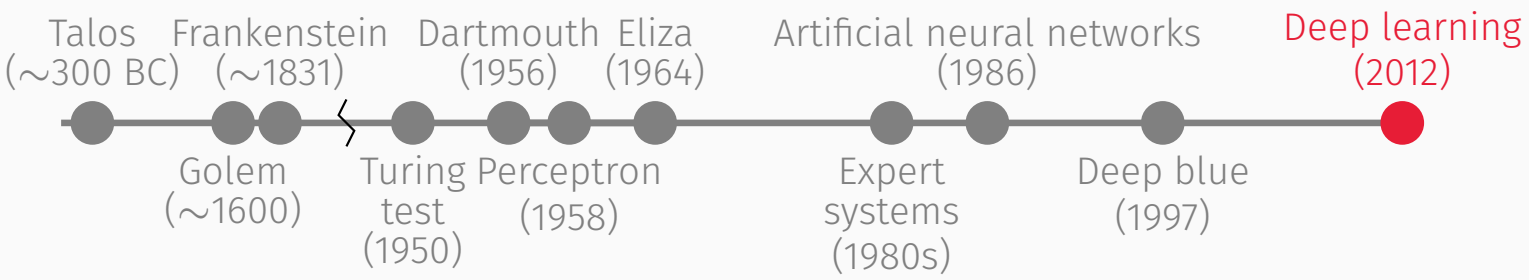


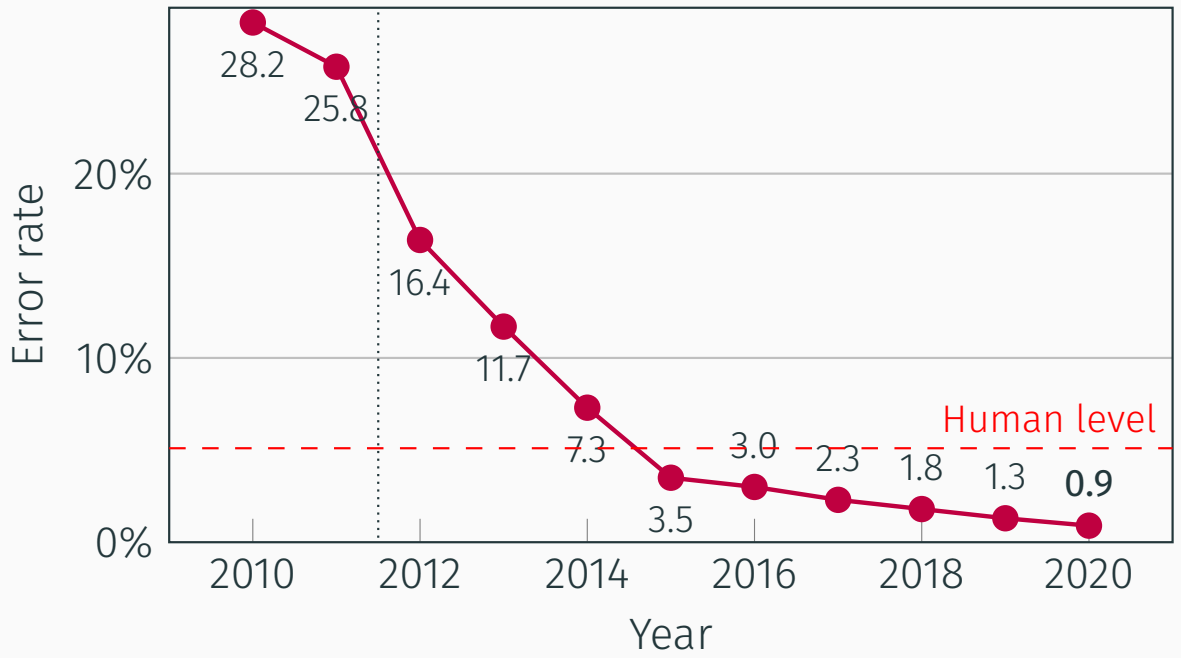
The history of artificial intelligence



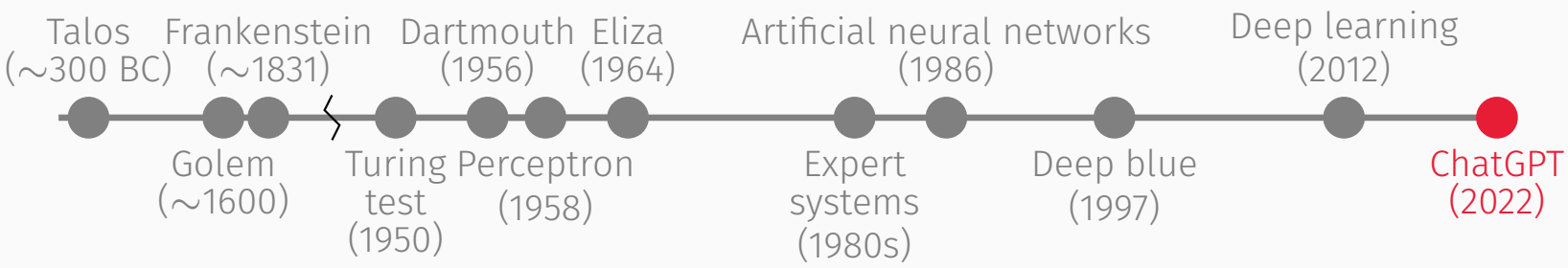


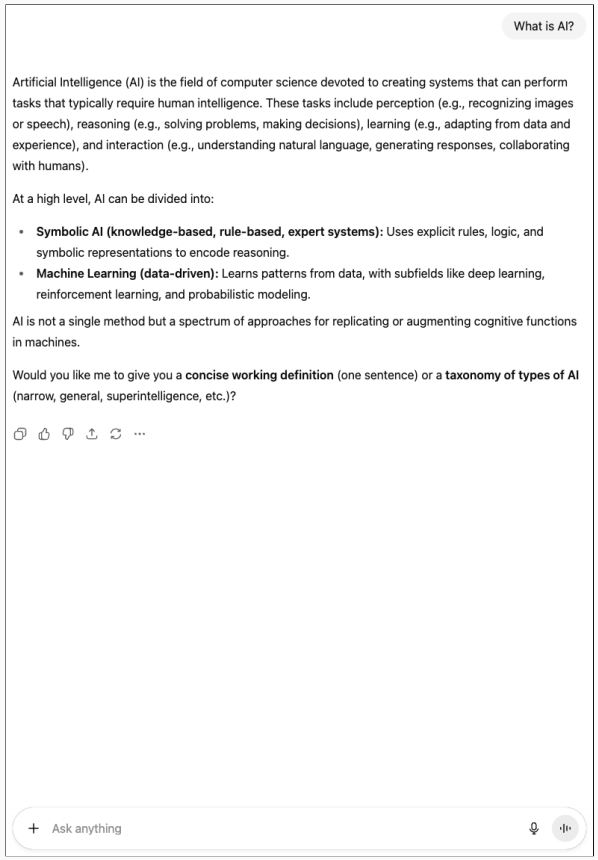
The history of artificial intelligence





The history of artificial intelligence





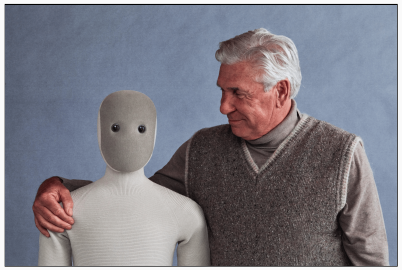
Terminology and concepts

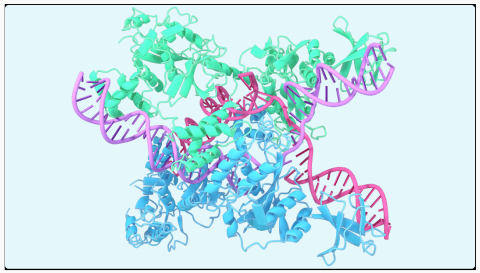
UNIVERSITY
OF OSLO

Terminology and concepts

ChatGPT

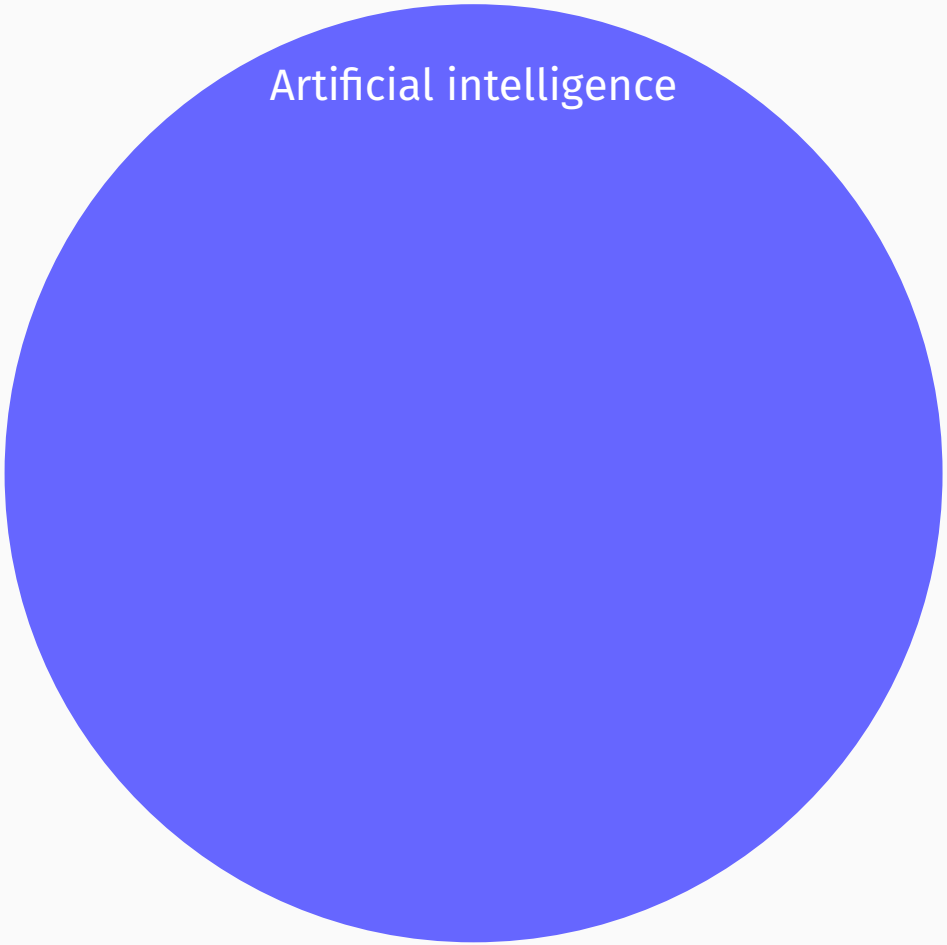
Spot


1X Neo


AlphaFold

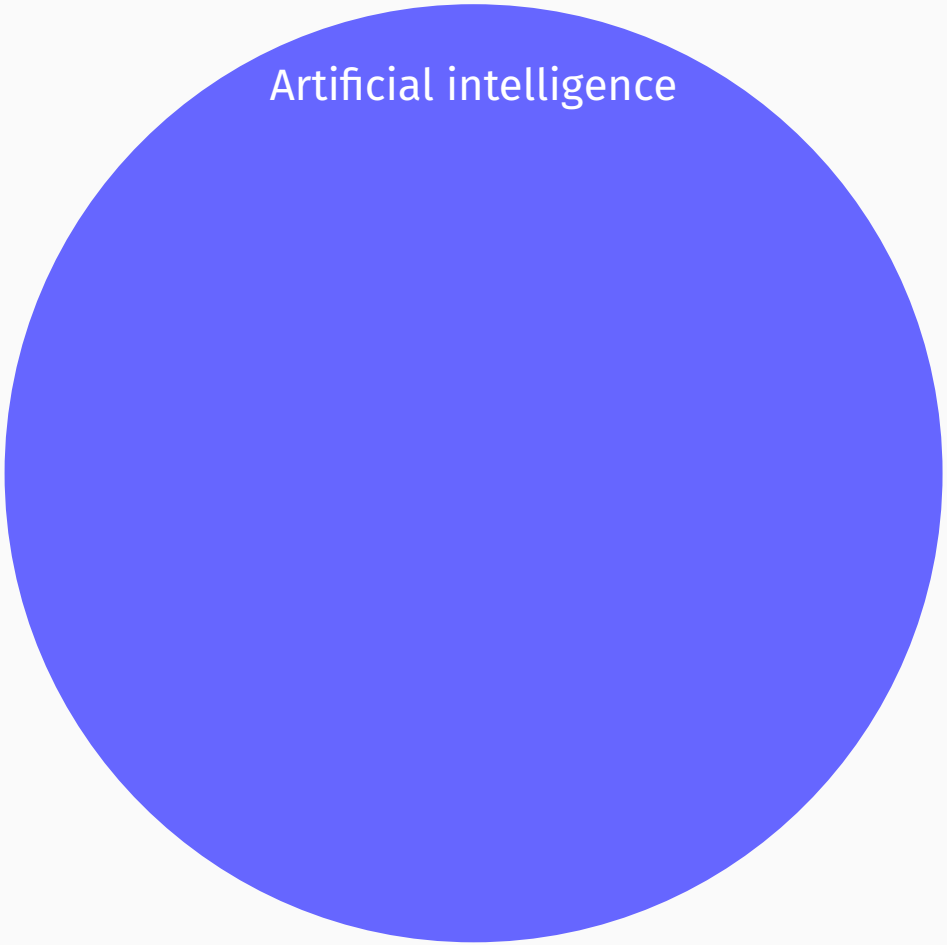

AlphaZero

Terminology and concepts: Taxonomy



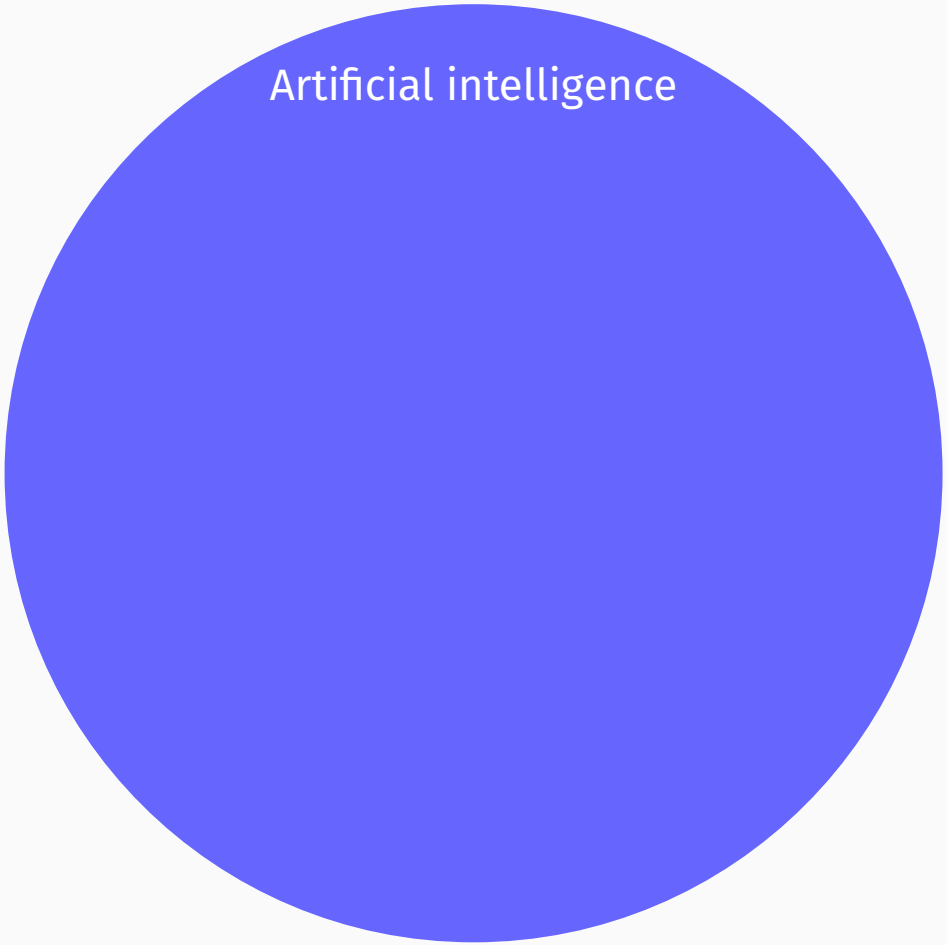
Artificial intelligence:
Machines that solve tasks requiring some kind of (often human-like) intelligence.

Terminology and concepts: Taxonomy



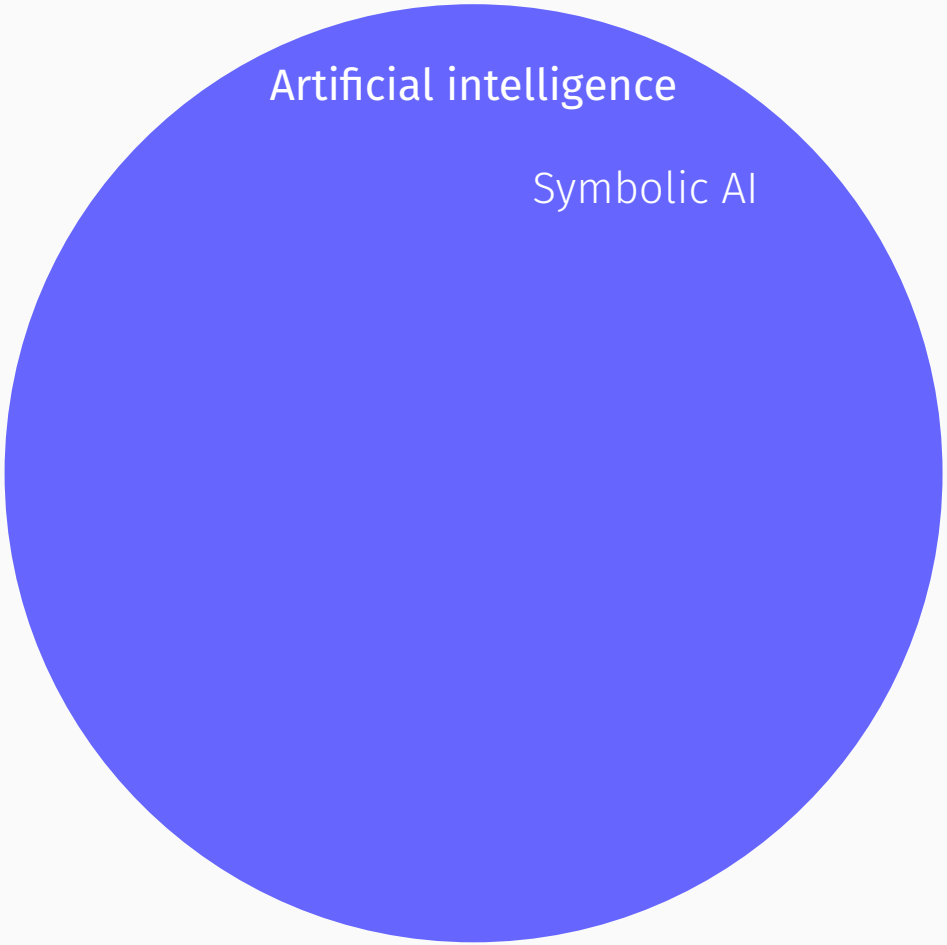
Artificial intelligence:
Machines that solve a wide array of tasks in various environments.

Terminology and concepts: Taxonomy



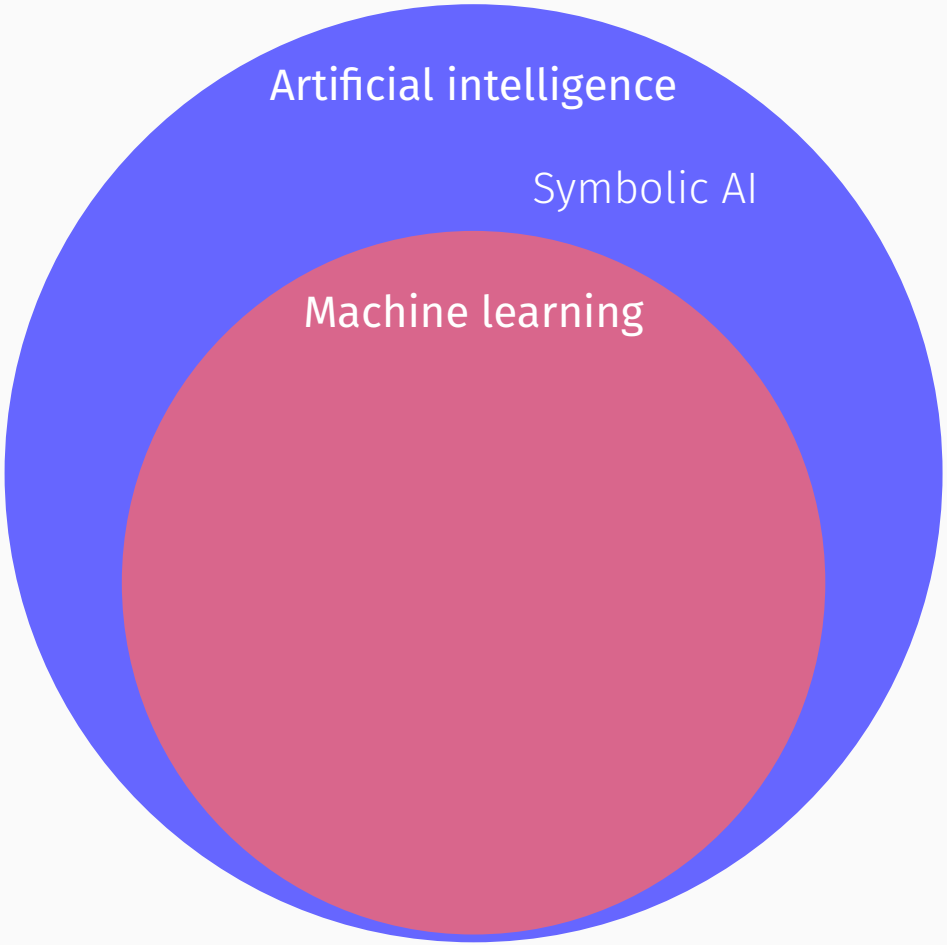
Artificial intelligence:
The field that produces machines that solve a wide array of tasks in various environments.

Terminology and concepts: Taxonomy



Artificial intelligence:
The field that produces machines that solve a wide array of tasks in various environments.

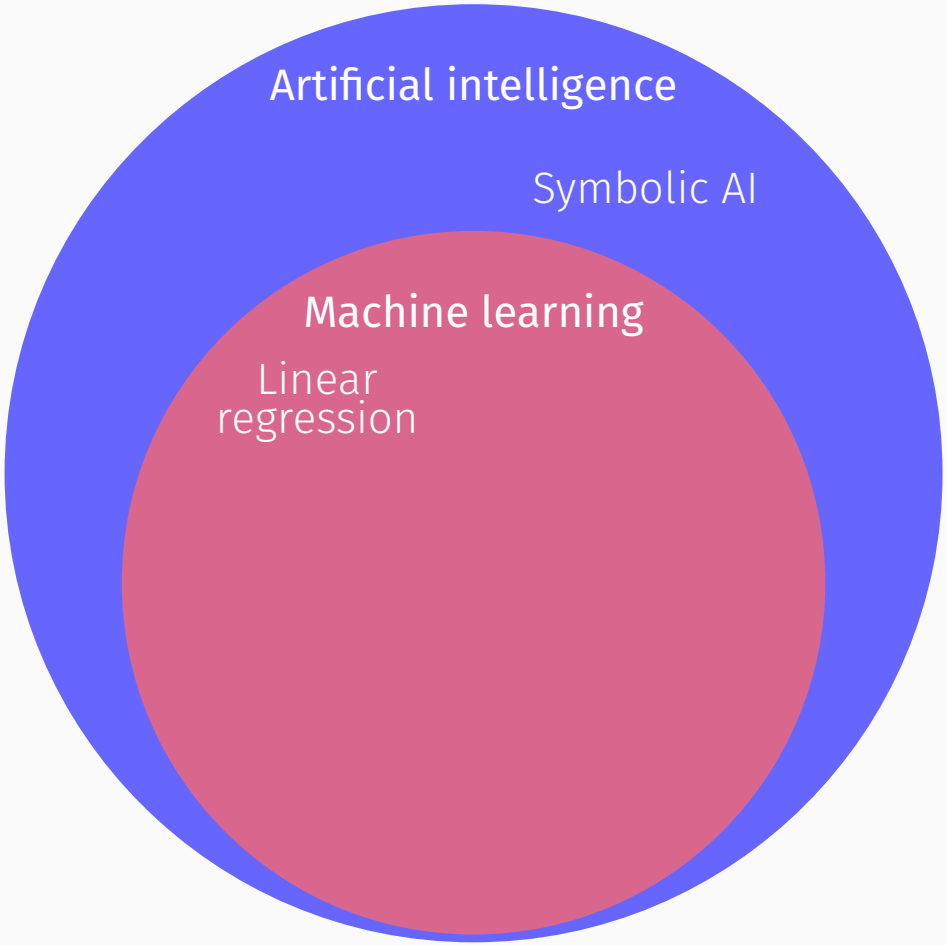
Terminology and concepts: Taxonomy



Artificial intelligence:
The field that produces machines that solve a wide array of tasks in various environments.

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

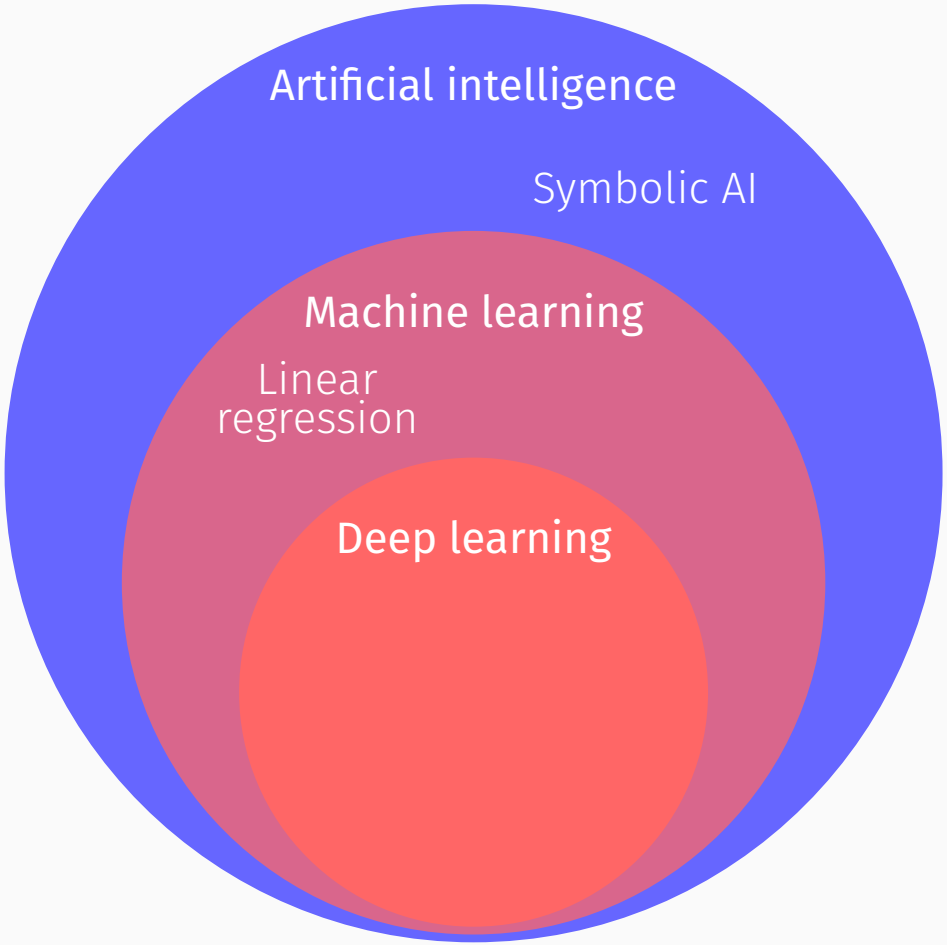
Terminology and concepts: Taxonomy



Artificial intelligence:
The field that produces machines that solve a wide array of tasks in various environments.

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

Terminology and concepts: Taxonomy

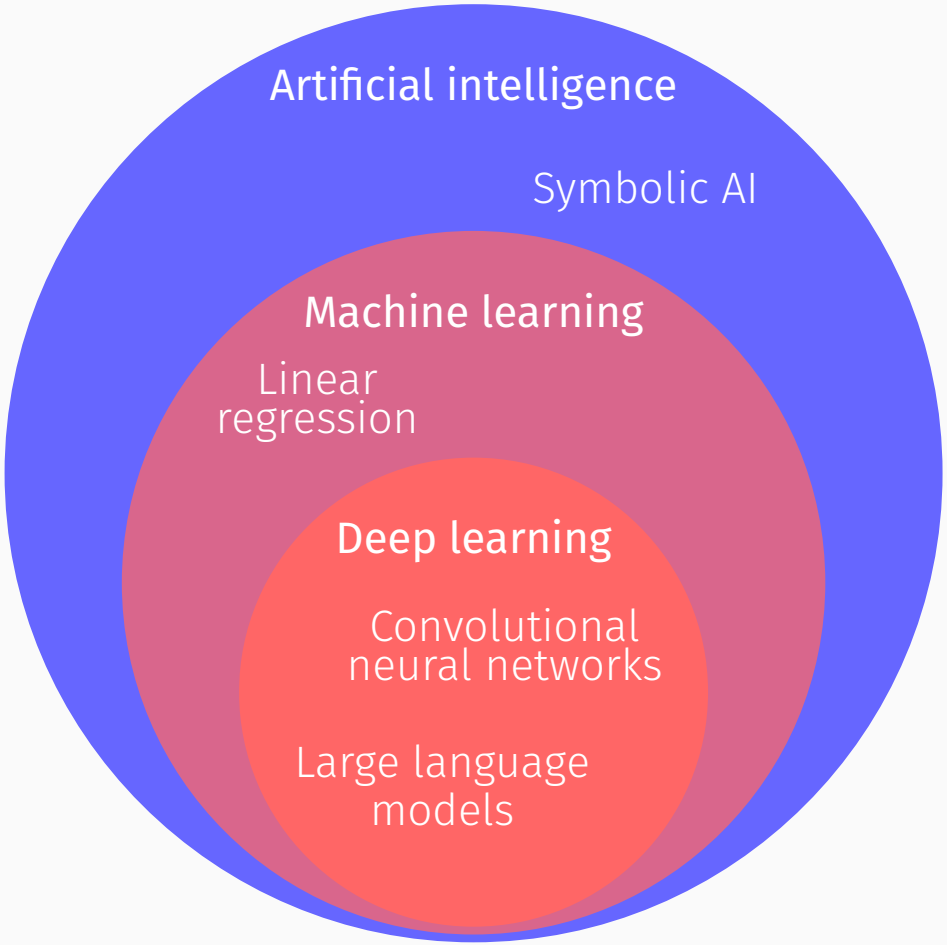


Artificial intelligence:
The field that produces machines that solve a wide array of tasks in various environments.

Machine learning:
Machines that learn to solve tasks by 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:
The field that produces machines that solve a wide array of tasks in various environments.

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

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

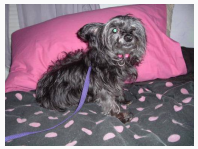
Convolutional neural nets:
Neural networks for image data

Large language models:
Neural networks for natural language (e.g. ChatGPT)

Terminology: Supervision

Supervised learning

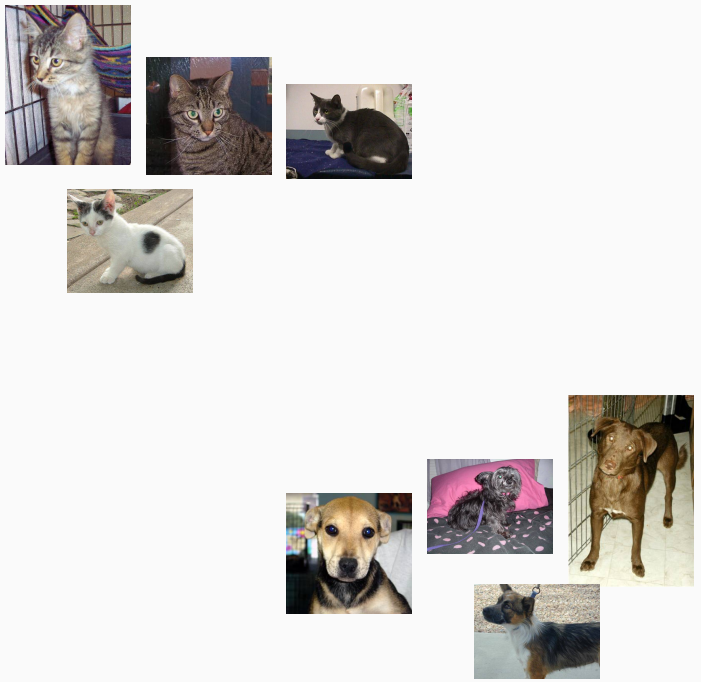
 → Cat

 → Dog

 → Cat

 → Dog

Unsupervised learning



Terminology: Strong and weak AI

Narrow (weak) — General (strong)

Able to solve a broader spectrum of problems in a wider array of domains

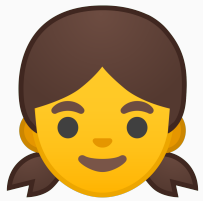
More specific — More general

Terminology: Strong and weak AI

Narrow (weak)

General (strong)

Able to solve a broader spectrum of problems in a wider array of domains



More specific

More general

Terminology: Strong and weak AI

Narrow (weak)

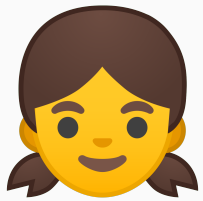
General (strong)

Able to solve a broader spectrum of problems in a wider array of domains

Image diagnostics

Insurance pricing

Document reading



More specific

More general

Terminology: Strong and weak AI

Narrow (weak)

General (strong)

Able to solve a broader spectrum of problems in a wider array of domains

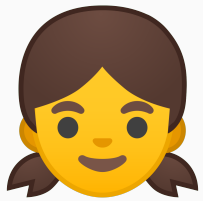
Image diagnostics

Insurance pricing

Document reading

ChatGPT

Tesla



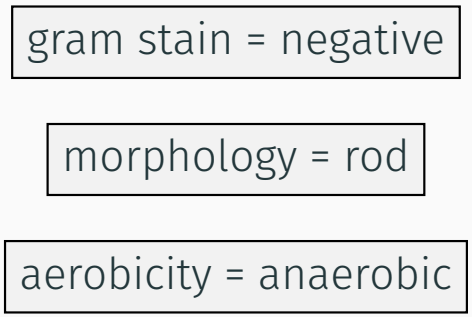
More specific

More general

How does AI make decisions?

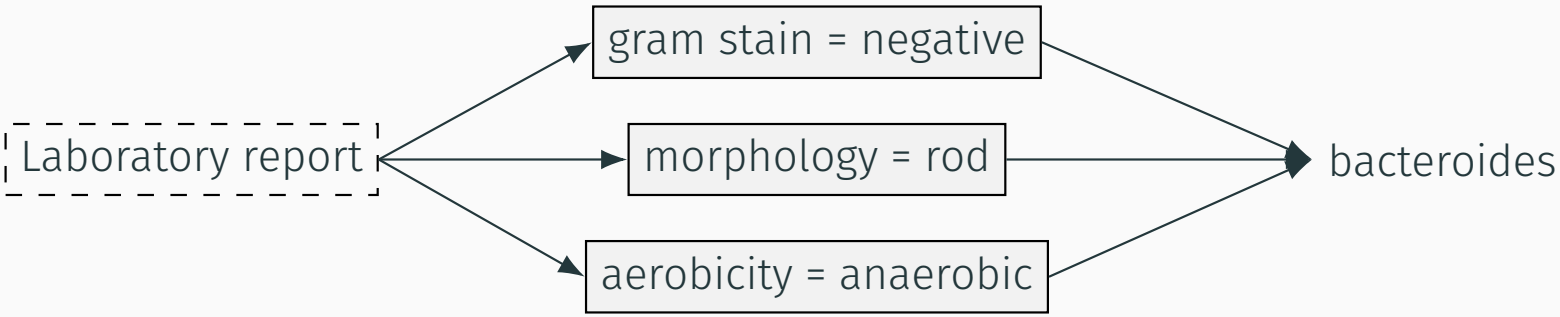
UNIVERSITY
OF OSLO

Decision-making: Expert systems vs. machine learning



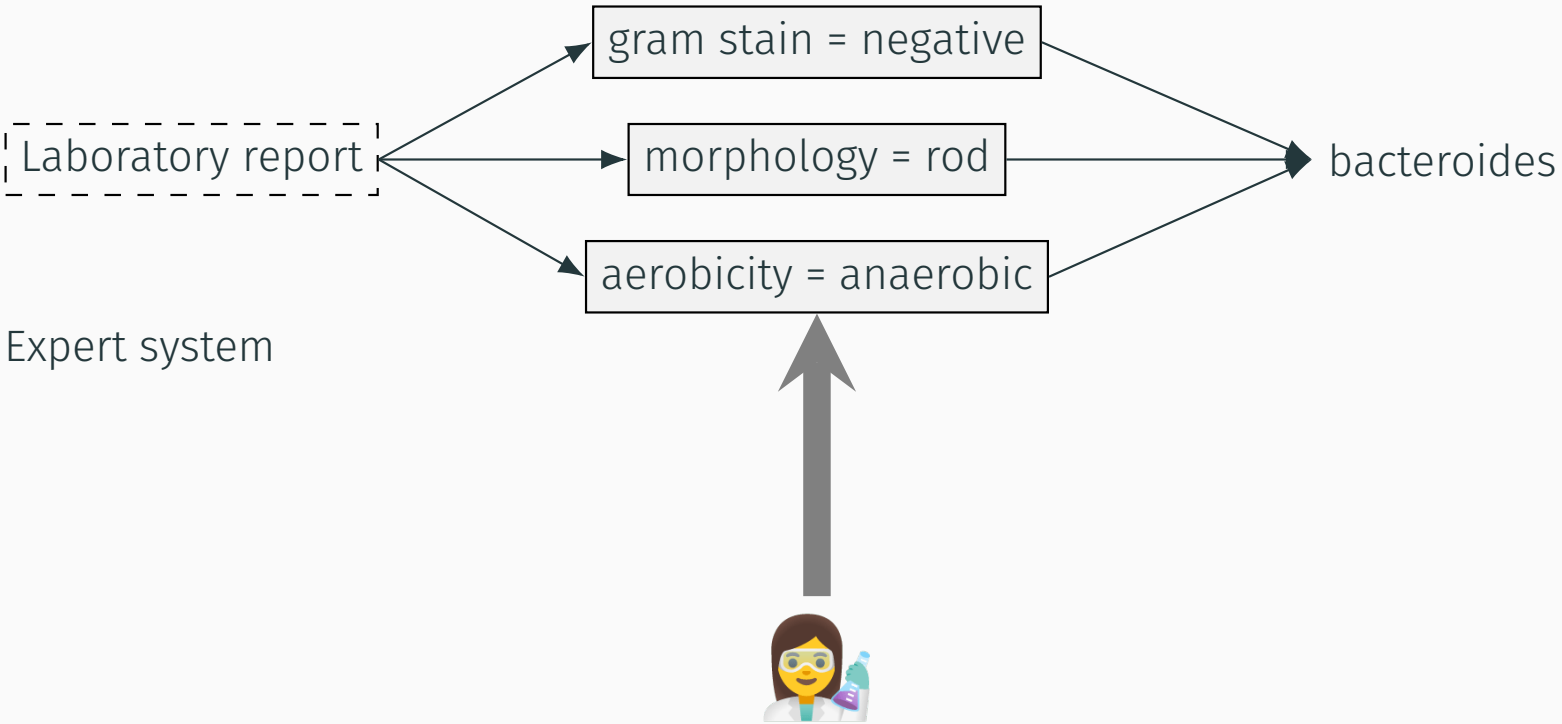
Expert system

Decision-making: Expert systems vs. machine learning

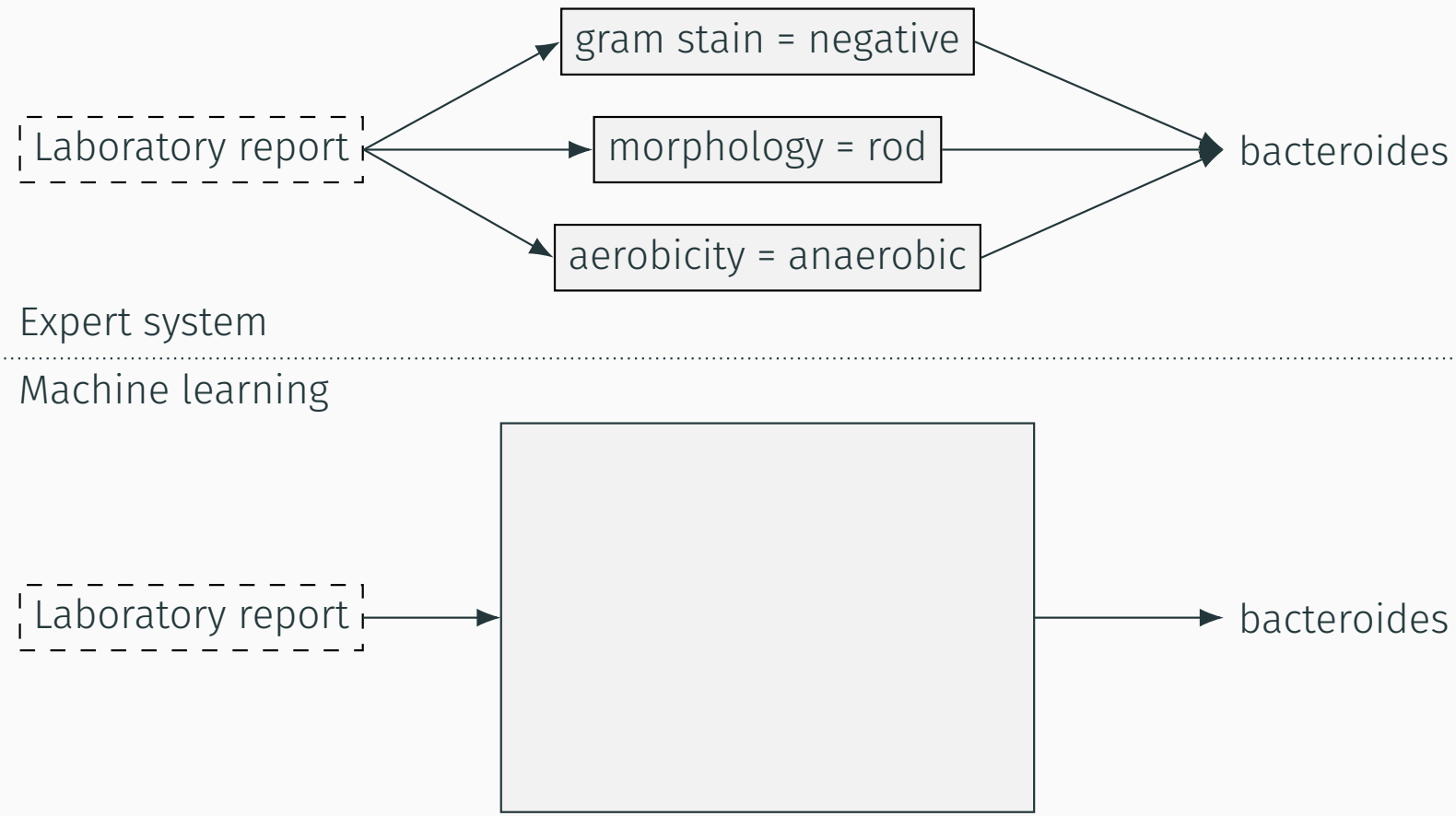


Expert system

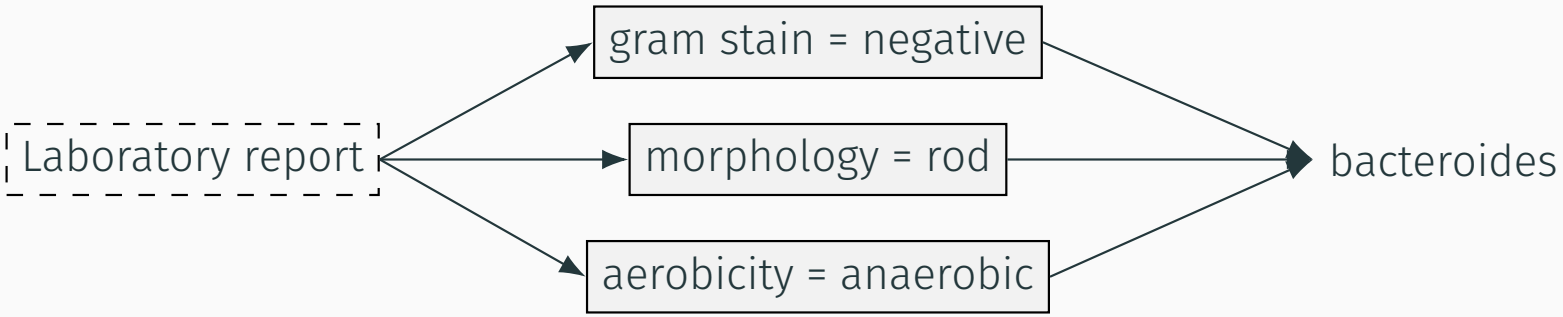
Decision-making: Expert systems vs. machine learning



Decision-making: Expert systems vs. machine learning

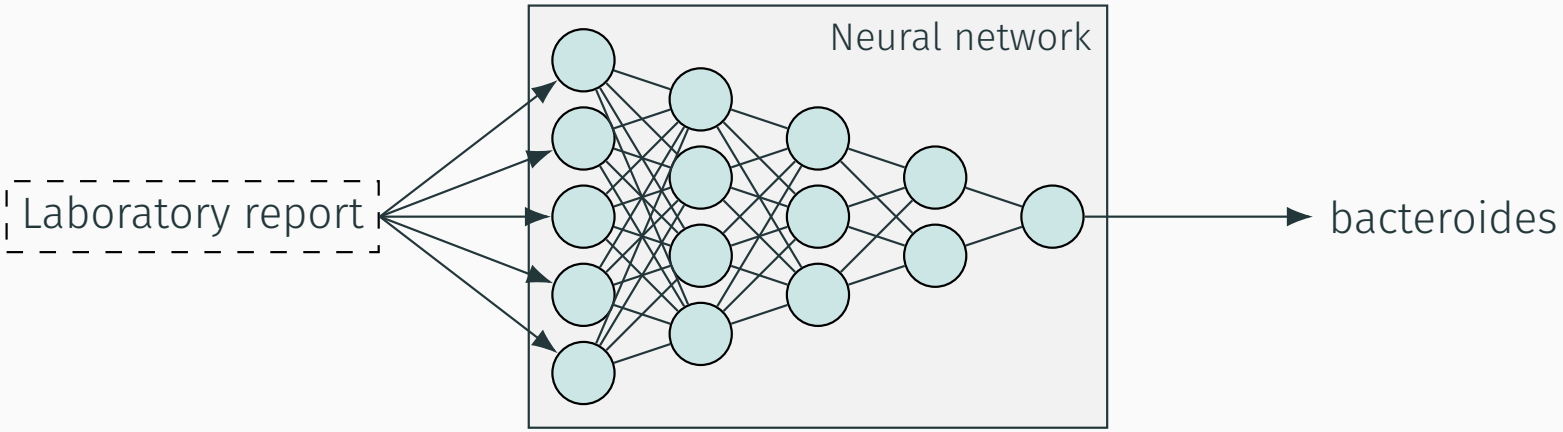


Decision-making: Expert systems vs. machine learning



Expert system

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 is in the image?

Decision-making: Loss functions

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

- **Classification**

  What is in the image?

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

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

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 is in the image?

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

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

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

- **Regression**

  How happy is the person that wrote this sentence on a scale of 1-10?

Decision-making: Loss functions

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

- **Classification**

  What is in the image?

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

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

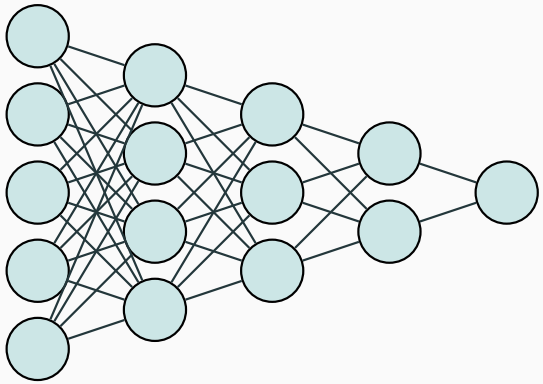
where y_i is the correct label and $\hat{y}_i$ is the predicted probability.

- **Regression**

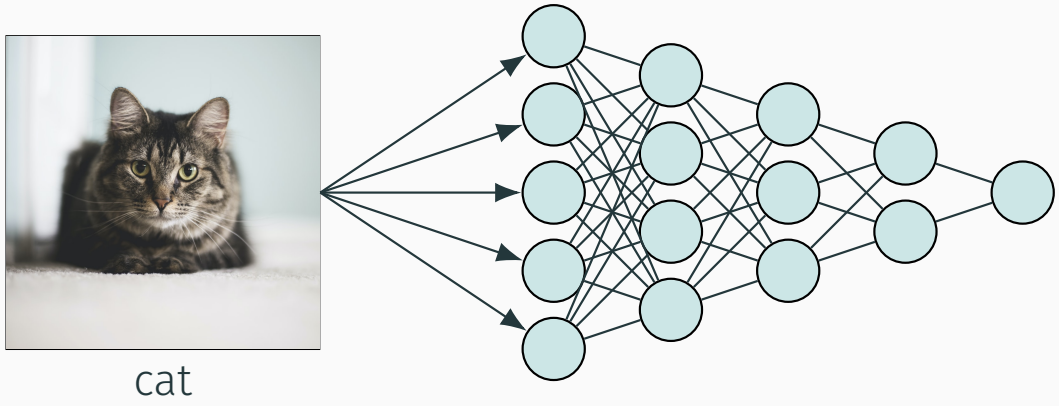
  How happy is the person that wrote this sentence on a scale of 1-10?

→ $$(y-\hat{y})^2$$

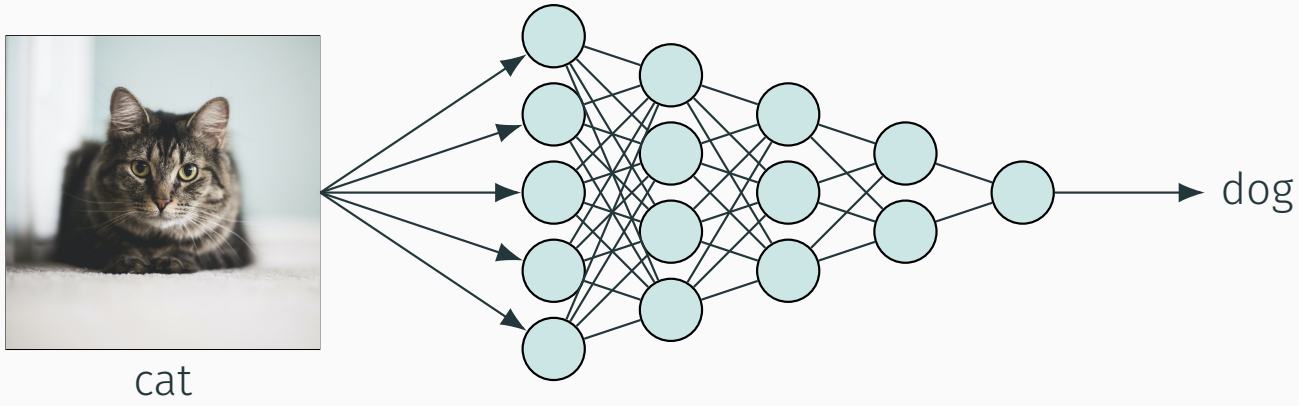
Decision-making: Learning



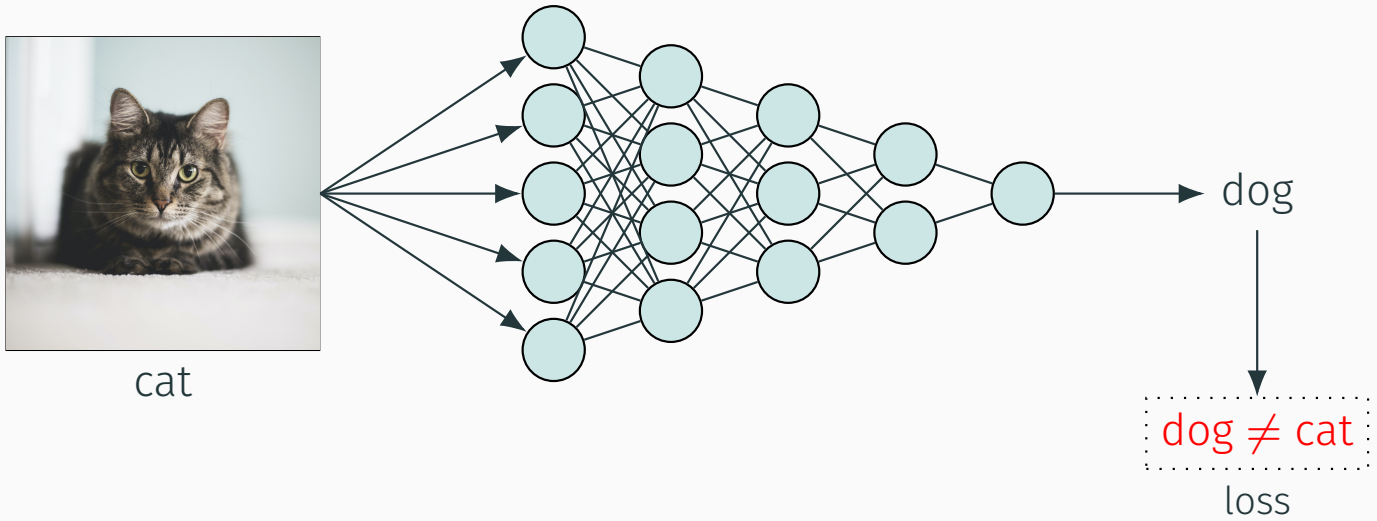
Decision-making: Learning



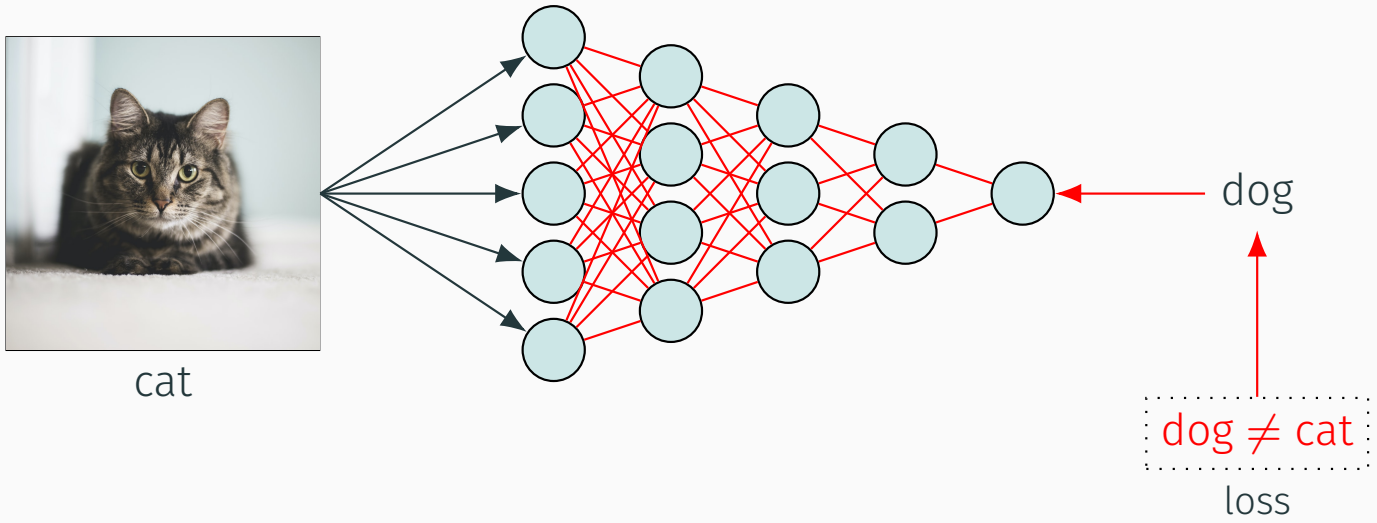
Decision-making: Learning



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.

+ 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 applies patterns from its training data that were sufficient to solve the task there.
- There is no guarantee these patterns are sufficient in new data.
- There are neither any guarantees these patterns are ones we want to use (e.g. bias).

Decision-making: Group work

We are dealing with an automatic system in a bank that automatically decides which of its clients are granted a loan.

- In the center of the system is a machine learning model that predicts the probability of a client defaulting. This model is a fully deterministic mathematical construction that takes some numbers as input (e.g. the clients age, sex, income, size of the loan, etc.) and gives a single number as an output. The model was trained on training data originating from previous customers of the bank.
- Around the neural network is a software system which the user interacts with through a website. After the user has input data, the system gives it to the neural network to make a prediction. If the neural network predicts a probability higher than 20%, the loan is declined. The threshold of 20% was implemented by a programmer, and decided upon by a business analyst.

A client gets his loan declined. Who or what is responsible for the decision?

Decision-making: Group work

We are dealing with an automatic system in a bank that automatically decides which of its clients are granted a loan.

- In the center of the system is a machine learning model that predicts the probability of a client defaulting. This model is a fully deterministic mathematical construction that takes some numbers as input (e.g. the clients age, sex, income, size of the loan, etc.) and gives a single number as an output. The model was trained on training data originating from previous customers of the bank.
- Around the neural network is a software system which the user interacts with through a website. After the user has input data, the system gives it to the neural network to make a prediction. If the neural network predicts a probability higher than 20%, the loan is declined. The threshold of 20% was implemented by a programmer, and decided upon by a business analyst.

A client gets his loan declined. Who or what is responsible for the decision?
The bank, the software system, the neural network, the programmer, the business analyst, previous customers (represented by the training data), the client (represented by his/her characteristics depicted in the input data)?

Decision-making: Generalization

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

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.” - Reasoning, Judging, Deciding: The Science of Thinking, 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.” - Reasoning, Judging, Deciding: The Science of Thinking, 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 (often 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." - Reasoning, Judging, Deciding: The Science of Thinking, 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 (often called the test set).

Decision-making: Generalization

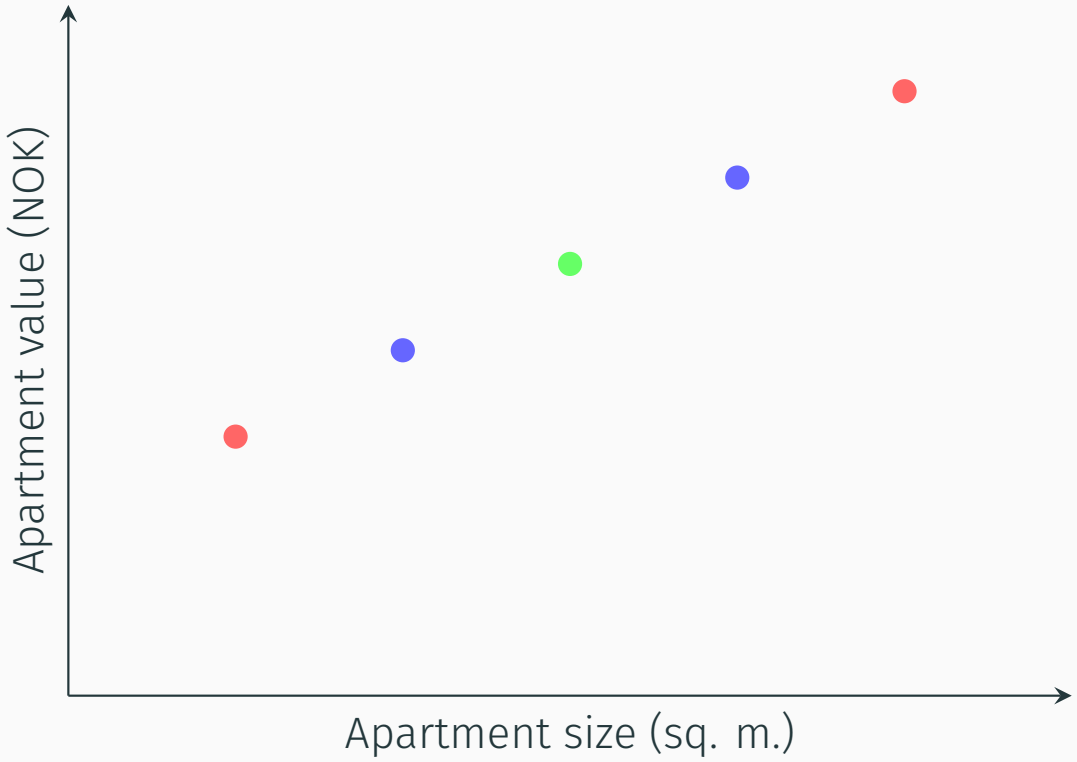
There is no guarantee the patterns the model has learned are sufficient in new data.

- ”AI that is based on datasets cannot go beyond what is in the data.” - Reasoning, Judging, Deciding: The Science of Thinking, 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 (often 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.” - Reasoning, Judging, Deciding: The Science of Thinking, 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 (often called the test set).



Decision-making: Biases

There is no guarantee the patterns the models have learned are ones we want to use

- A 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 from their training data.

Decision-making: Biases

There is no guarantee the patterns the models have learned are ones we want to use

- A 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 from their 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.

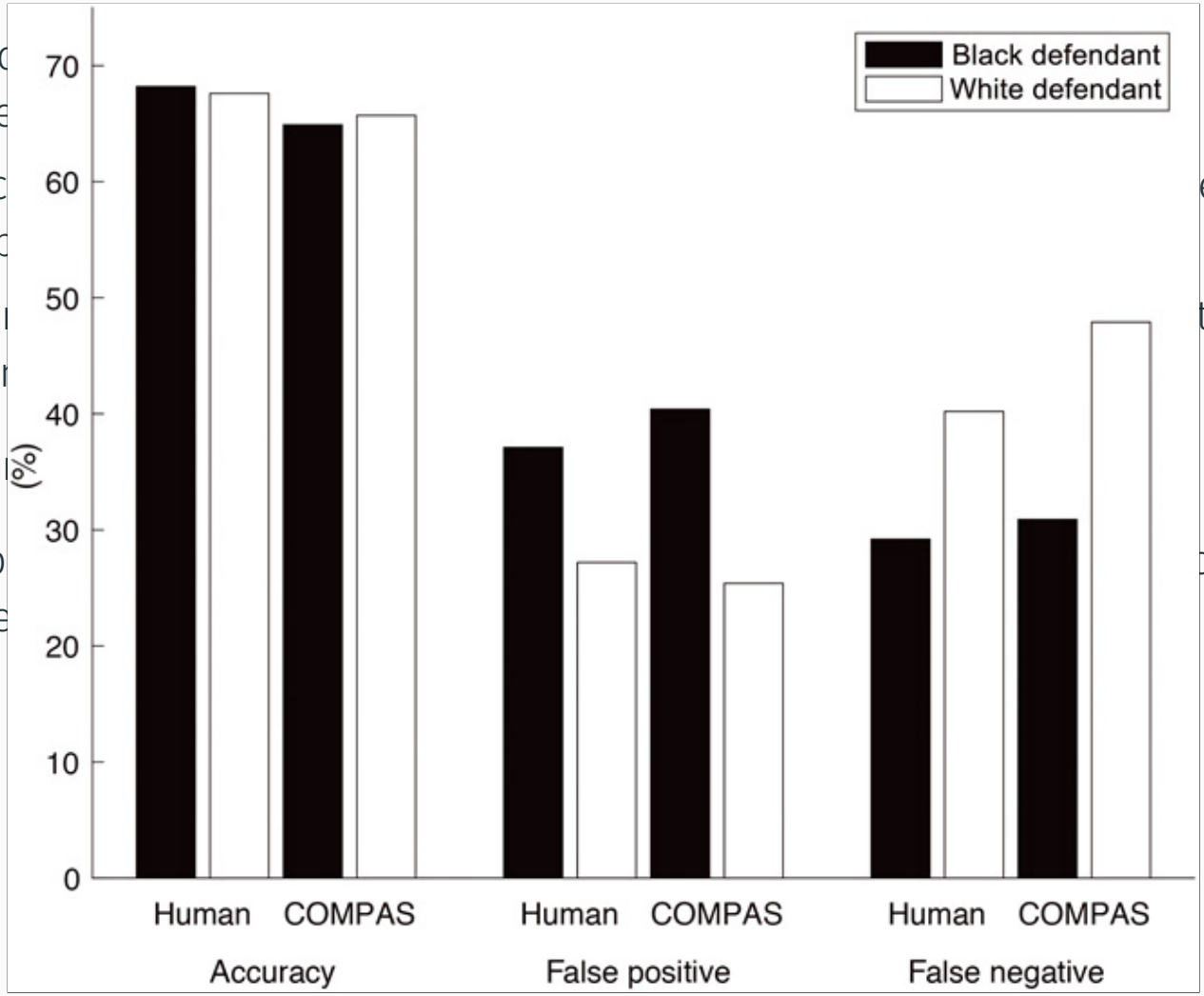
Decision-making: Biases

There is no guarantee the patterns the models have learned are ones we want to use

- A mod (age, gende
- This c ese variab
- Thus t their traini

Bias in cri

- Comp oftware, and non-e



The accuracy, fairness, and limits of predicting recidivism, Dressel J. and Farid H., *Science Advances*, 2018

Decision-making: Biases

There is no guarantee the patterns the models have learned are ones we want to use

- A 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 from their 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 ethnicity was not included in the data.

Decision-making: Biases

There is no guarantee the patterns the models have learned are ones we want to use

- A 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 from their 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 ethnicity was not included 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

There is no guarantee the patterns the models have learned are ones we want to use

- A 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 from their training data.

Bias in hiring (Bertrand & Mullainathan, 2004)

- Evaluation of bias in human decision-making in help-wanted advertisements in the United States.

$\rightarrow$ "Applicants" were given very African American or European-sounding names.

Are Emily and Greg More Employable than Lakisha and Jamal? ..., Bertrand M. & Mullainathan S., *American economic review*, 2004

Decision-making: Biases

There is no guarantee the patterns the models have learned are ones we want to use

- A 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 from their training data.

Bias in hiring (Bertrand & Mullainathan, 2004)

- Evaluation of bias in human decision-making in help-wanted advertisements in the United States.

$\rightarrow$ ”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?

Decision-making: Theory of mind

Does AI consider humans as thinking and feeling beings?

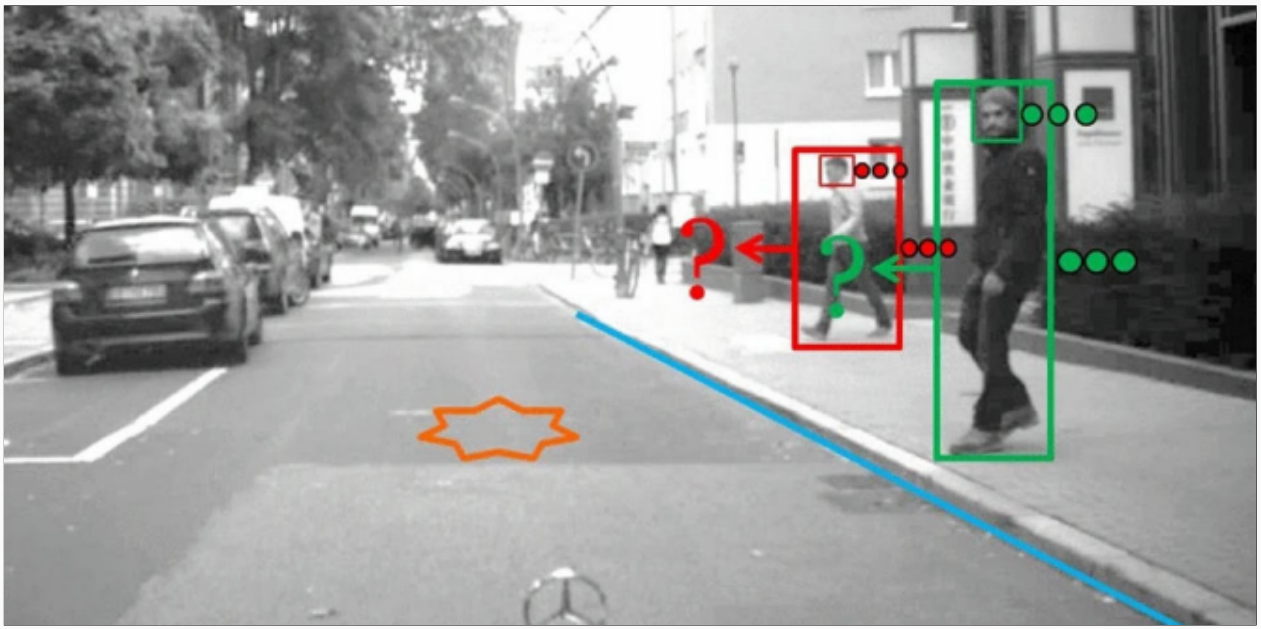
- "... This is an instance of AI programs lacking true Theory of Mind capability." - Reasoning, Judging, Deciding: The Science of Thinking, Ch. 15
- Theory of mind: The ability to "track others' unobservable mental states, such as their knowledge, intentions, beliefs, and desires." (Kosinski 2023)

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." - Reasoning, Judging, Deciding: The Science of Thinking, 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)



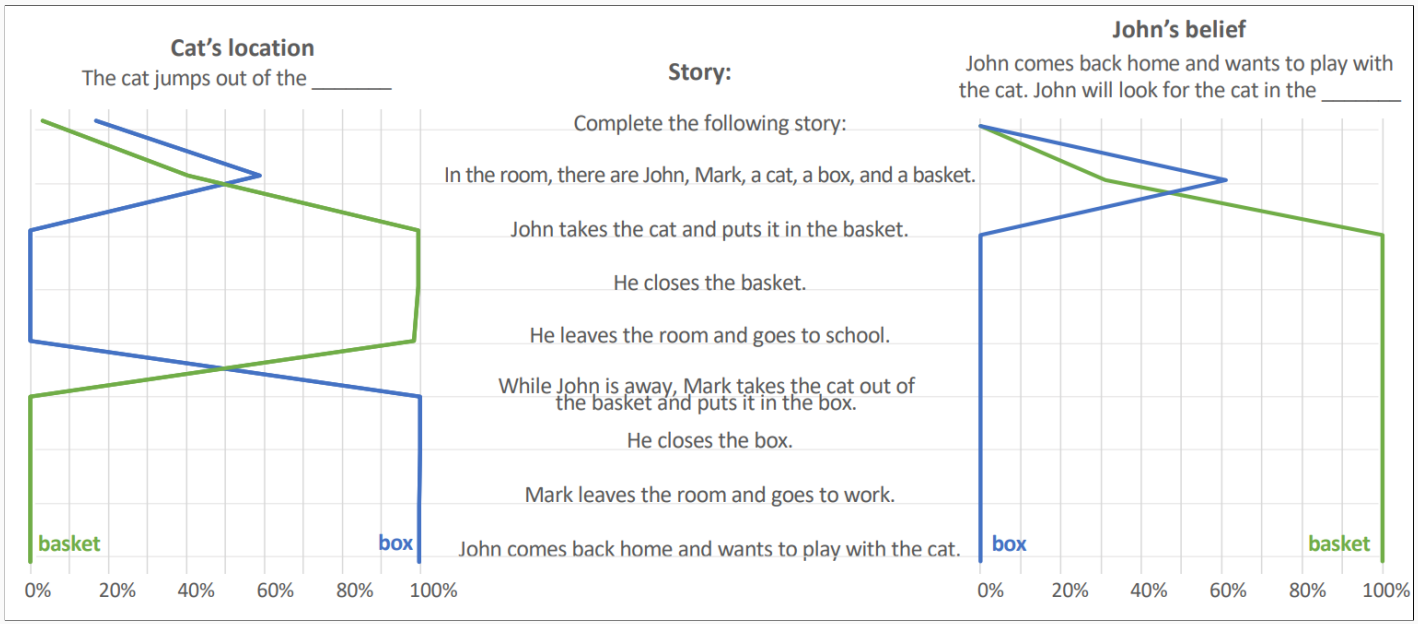
A Survey on Motion Prediction of Pedestrians and Vehicles for Autonomous Driving, Gulzar, M. et al, *IEEE Access 9*, 2021.

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." - Reasoning, Judging, Deciding: The Science of Thinking, 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)



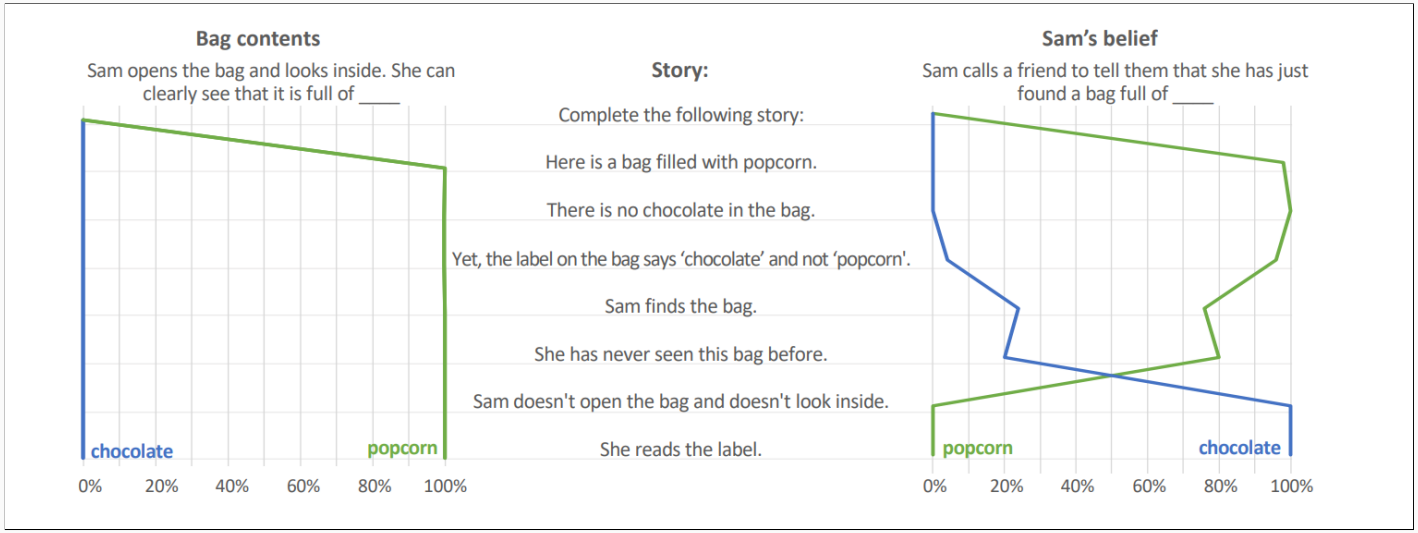
Theory of Mind Might Have Spontaneously Emerged in Large Language Models, Kosinski, M., *PNAS*, 2024.

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." - Reasoning, Judging, Deciding: The Science of Thinking, 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)



Theory of Mind Might Have Spontaneously Emerged in Large Language Models, Kosinski, M., *PNAS*, 2024.

Decision-making: Creativity

Can AI create anything that is truly new?

Decision-making: Creativity

Can AI create anything that is truly new?

- ”AI does not truly create” - Reasoning, Judging, Deciding: The Science of Thinking, Ch. 15
- ”AI lacks true imagination” - Reasoning, Judging, Deciding: The Science of Thinking, Ch. 15

Decision-making: Creativity

Can AI create anything that is truly new?

- ”AI does not truly create” - Reasoning, Judging, Deciding: The Science of Thinking, Ch. 15
- ”AI lacks true imagination” - Reasoning, Judging, Deciding: The Science of Thinking, Ch. 15

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

Decision-making: Wisdom

Are AIs wise?

- "... machines don't have common sense" - Korteling et al., 2021
- "... 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." - Reasoning, Judging, Deciding: The Science of Thinking, Ch. 15 (adopted from Birren and Svensson, attributed to Baltes and Smith)

Decision-making: Wisdom

Are AIs wise?

- "... machines don't have common sense" - Korteling et al., 2021
- "... 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." - Reasoning, Judging, Deciding: The Science of Thinking, 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 (which can lead to surprising errors).
- Mostly unimodal (e.g. relies only on text) and non-causal, little opportunity to interact with the world.
- **Little introspection towards its own limits or uncertainties.**

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 decisions made by AI systems?

- 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 that is truly new.
- Lacks wisdom, a fundamental understanding of the world, and common sense.
- **Reliable and objective (in one sense of the word)**

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 decisions made by AI systems?

- 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 that is truly new.
- Lacks wisdom, a fundamental understanding of the world, and common sense.
- **Reliable and objective (in one sense of the word)**

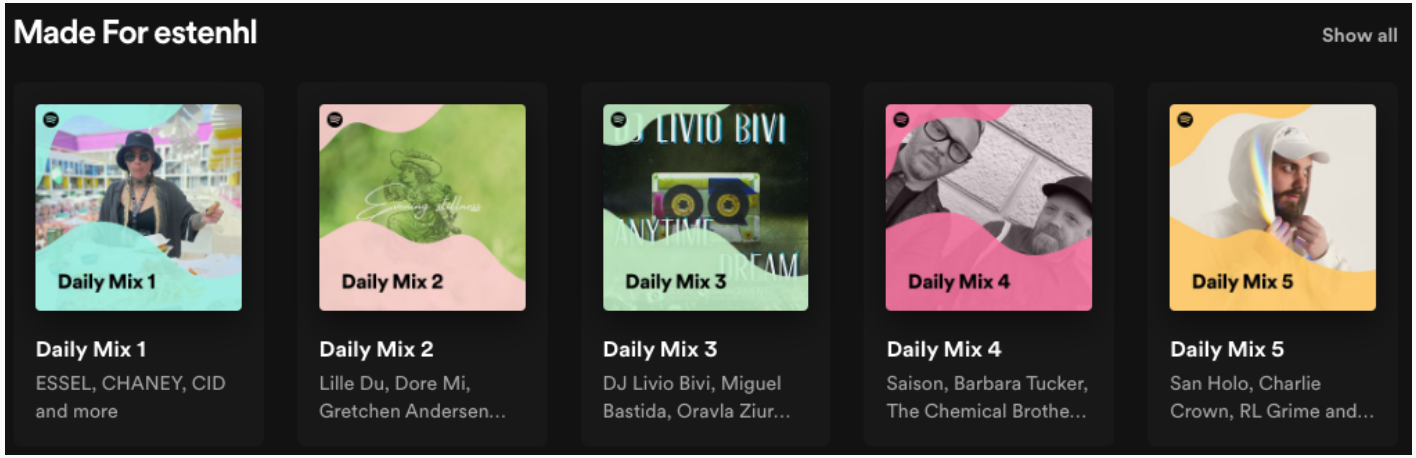
→ What is wisdom, creativity?

AI for decision-support

UNIVERSITY
OF OSLO

Decision-support: Content personalization

Helping users decide what to listen to



- Recommends content to users based on their history.
- Has been around for a long time.
- Extremely intricate trade-offs between exploitation, showing users what they like, and exploration, showing users new content.
- **Based around recommendation, not clear cut decisions.**
- Can potentially lead to feedback loops?

Decision-support: Fracture detection

Helping doctors detect fractures in X-rays

- Bærum sykehus is the first norwegian hospital to implement an AI powered decision-support system into the clinic.
- Helps alleviate a 12.5% year-on-year increase in the prevalence of fractures.
- 60% to 70% of all X-rays are normal, but still need to be reviewed by a radiologist.

RØNTGEN: Radiograf Jonas Vatle plasserer røntgenmaskinen over hånden til Davejit Bhuller (41). Formålet er å finne ut om han har brukket hånden. Foto: Janne Møller-Hansen / VG

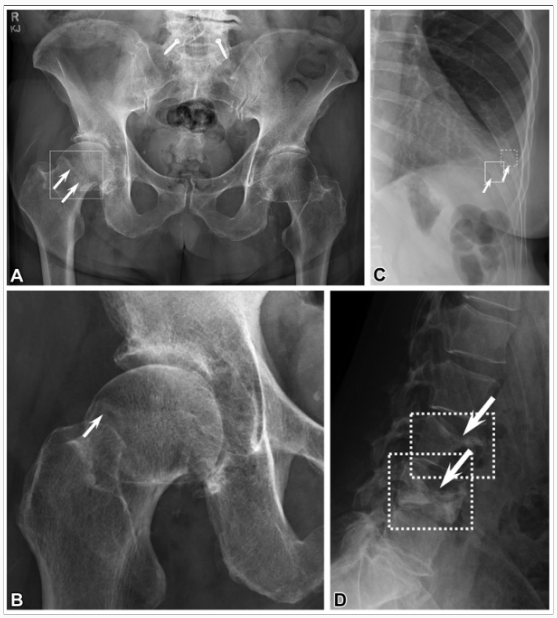
Fikk hånden analysert av kunstig intelligens: – Resultatet kom så raskt

Decision-support: Fracture detection

Helping doctors detect fractures in X-rays

- Bærum sykehus is the first norwegian hospital to implement an AI powered decision-support system into the clinic.
- Helps alleviate a 12.5% year-on-year increase in the prevalence of fractures.
- 60% to 70% of all X-rays are normal, but still need to be reviewed by a radiologist.

Assessing the efficacy of AI in fracture detection (Guermazi et al., 2022)



Improving Radiographic Fracture Recognition Performance and Efficiency Using Artificial Intelligence, Guermazi, A. et al, *Radiology*, 2022.

Decision-support: Fracture detection

Helping doctors detect fractures in X-rays

- Bærum sykehus is the first norwegian hospital to implement an AI powered decision-support system into the clinic.
- Helps alleviate a 12.5% year-on-year increase in the prevalence of fractures.
- 60% to 70% of all X-rays are normal, but still need to be reviewed by a radiologist.

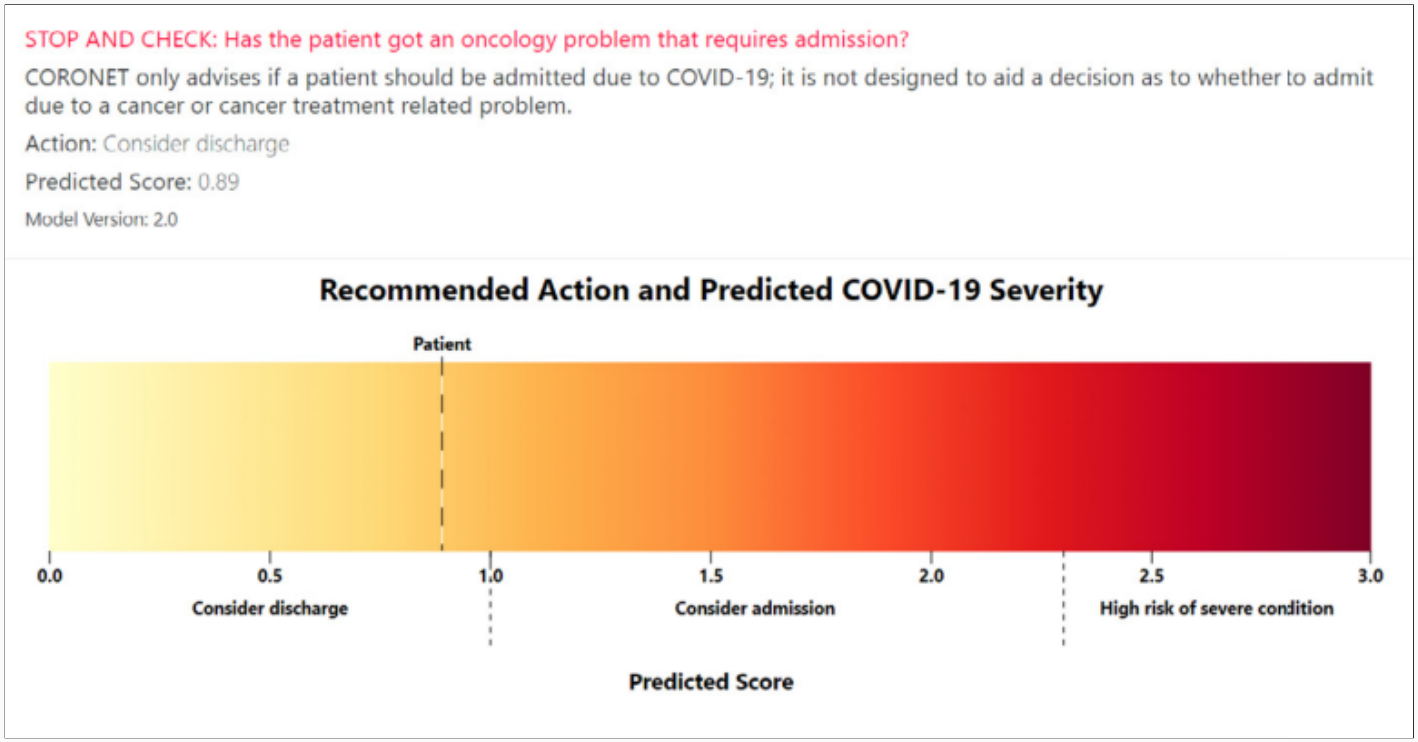
Assessing the efficacy of AI in fracture detection (Guermazi et al., 2022)

- AI assistant on its own achieved an area under the receiver operating characteristic curve of 0.97.
- Radiologist in conjunction with the AI assistant achieved a 10.4% increase in sensitivity (64.8% to 75.2%), and an increase in specificity (90.6% vs 95.6%).
- **Assistance from the AI reduced average reading time with 6.3 seconds**.

Decision-support: COVID-19 severity

Helping doctors decide the severity of COVID-19 cases (Wysocki et al., 2023)

- 23 healthcare professionals tasked to assess the severity of COVID-19 in ten patients using the COVID-19 Risk in ONcology Evaluation Tool (CORONET) tool.

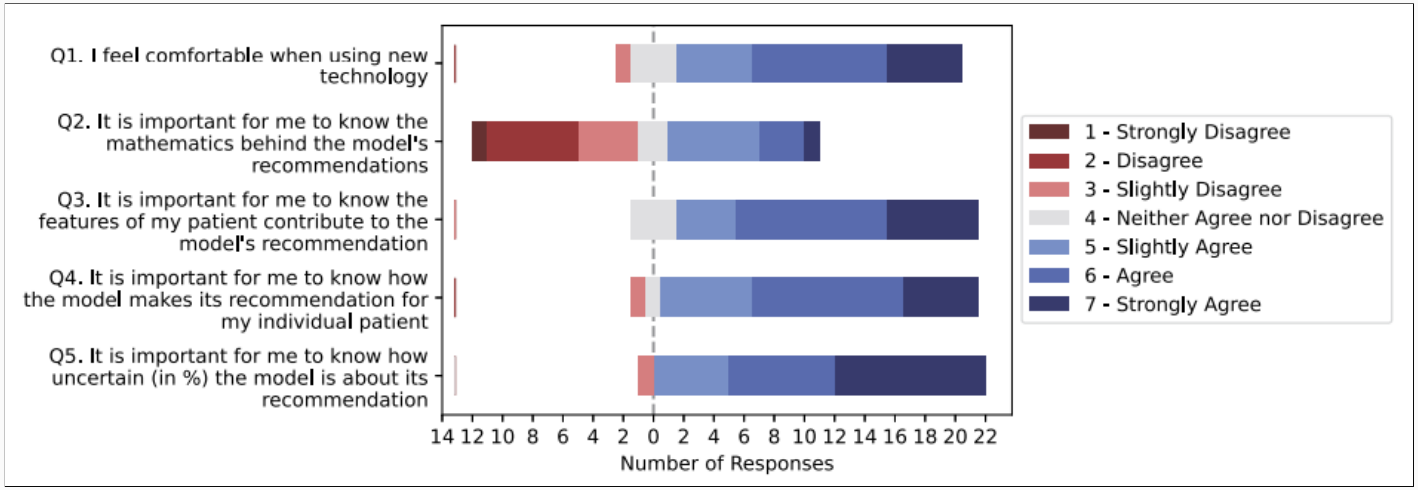


Assessing the communication gap between AI models and healthcare professionals ..., Wysocki, O. et al, *Artificial Intelligence*, 2023.

Decision-support: COVID-19 severity

Helping doctors decide the severity of COVID-19 cases (Wysocki et al., 2023)

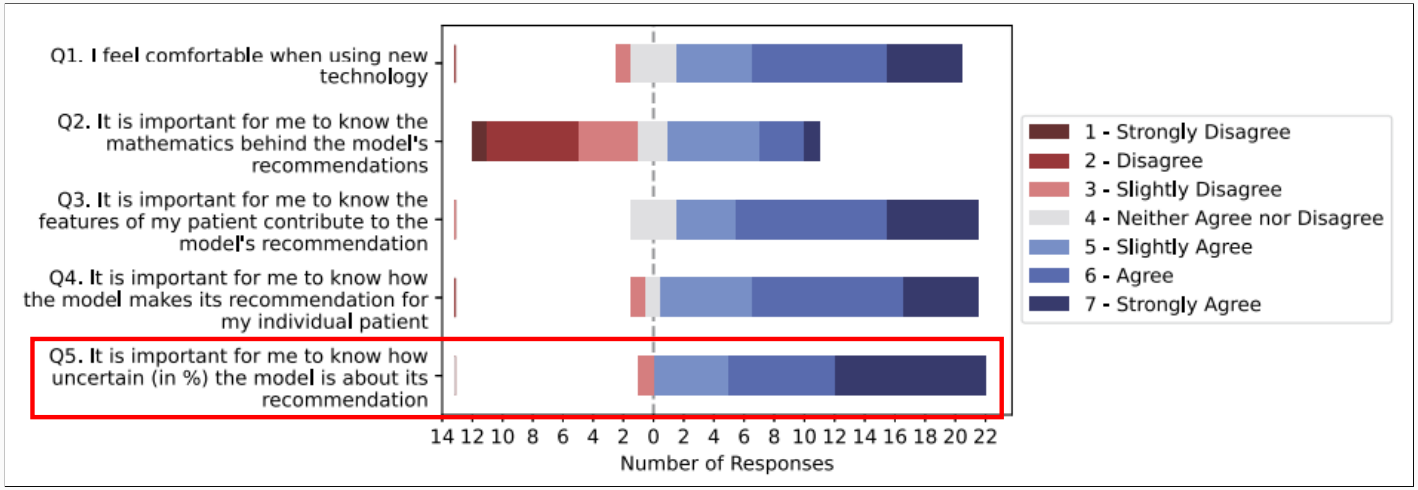
- 23 healthcare professionals tasked to assess the severity of COVID-19 in ten patients using the COVID-19 Risk in ONcology Evaluation Tool (CORONET) tool.
- Questioned about their experience using the tool.



Decision-support: COVID-19 severity

Helping doctors decide the severity of COVID-19 cases (Wysocki et al., 2023)

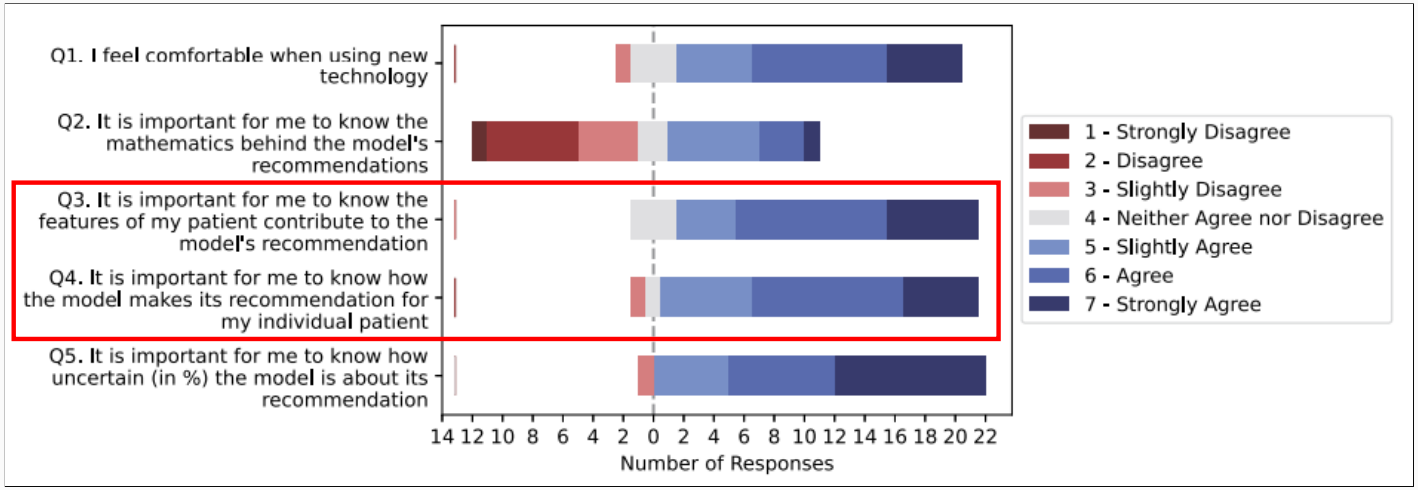
- 23 healthcare professionals tasked to assess the severity of COVID-19 in ten patients using the COVID-19 Risk in ONcology Evaluation Tool (CORONET) tool.
- Questioned about their experience using the tool.



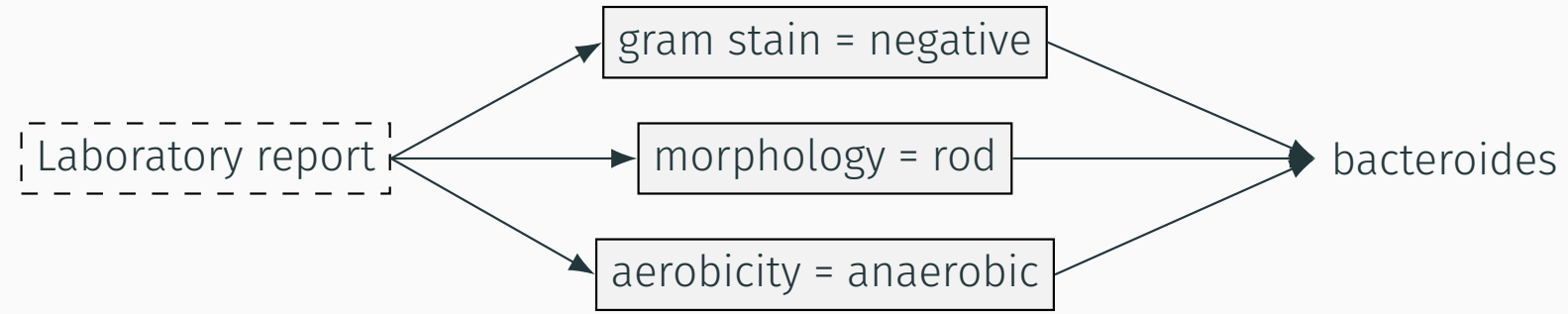
Decision-support: COVID-19 severity

Helping doctors decide the severity of COVID-19 cases (Wysocki et al., 2023)

- 23 healthcare professionals tasked to assess the severity of COVID-19 in ten patients using the COVID-19 Risk in ONcology Evaluation Tool (CORONET) tool.
- Questioned about their experience using the tool.

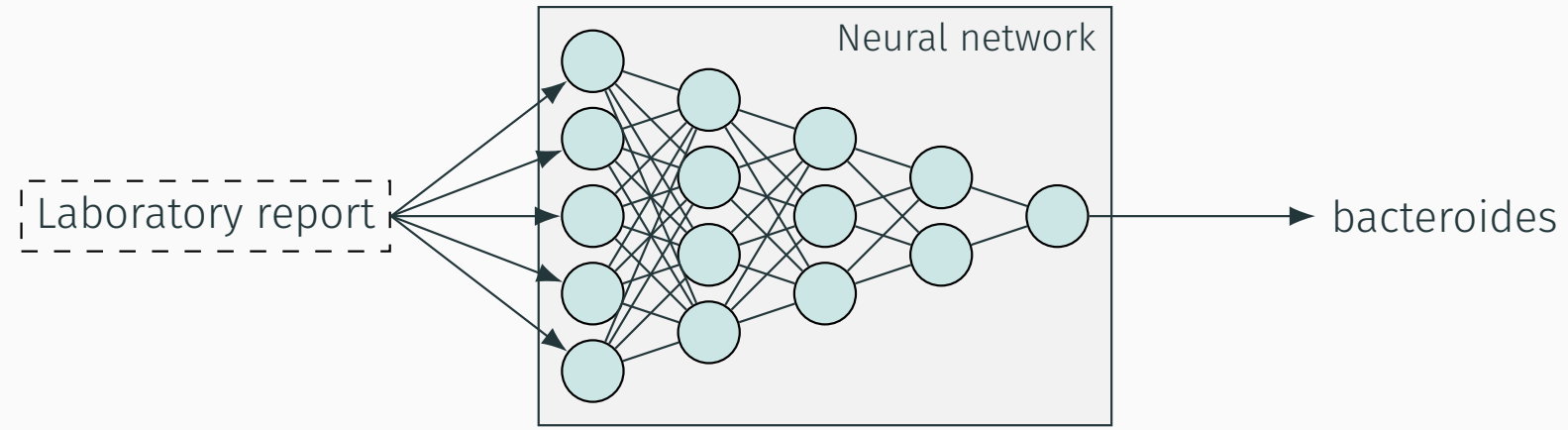


Decision-support: Interpretability

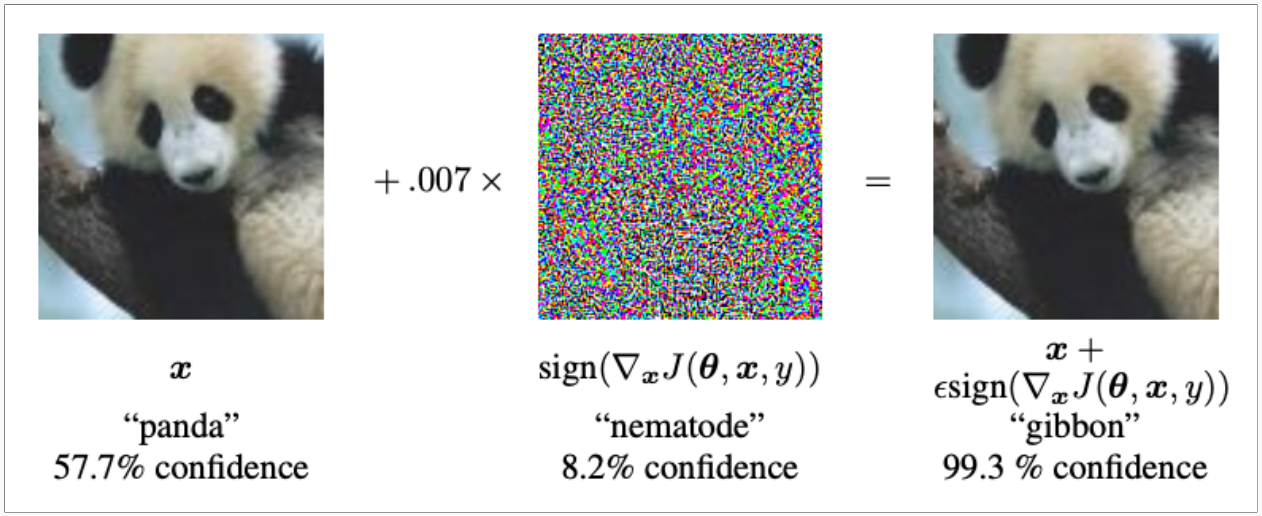


Expert system

Machine learning

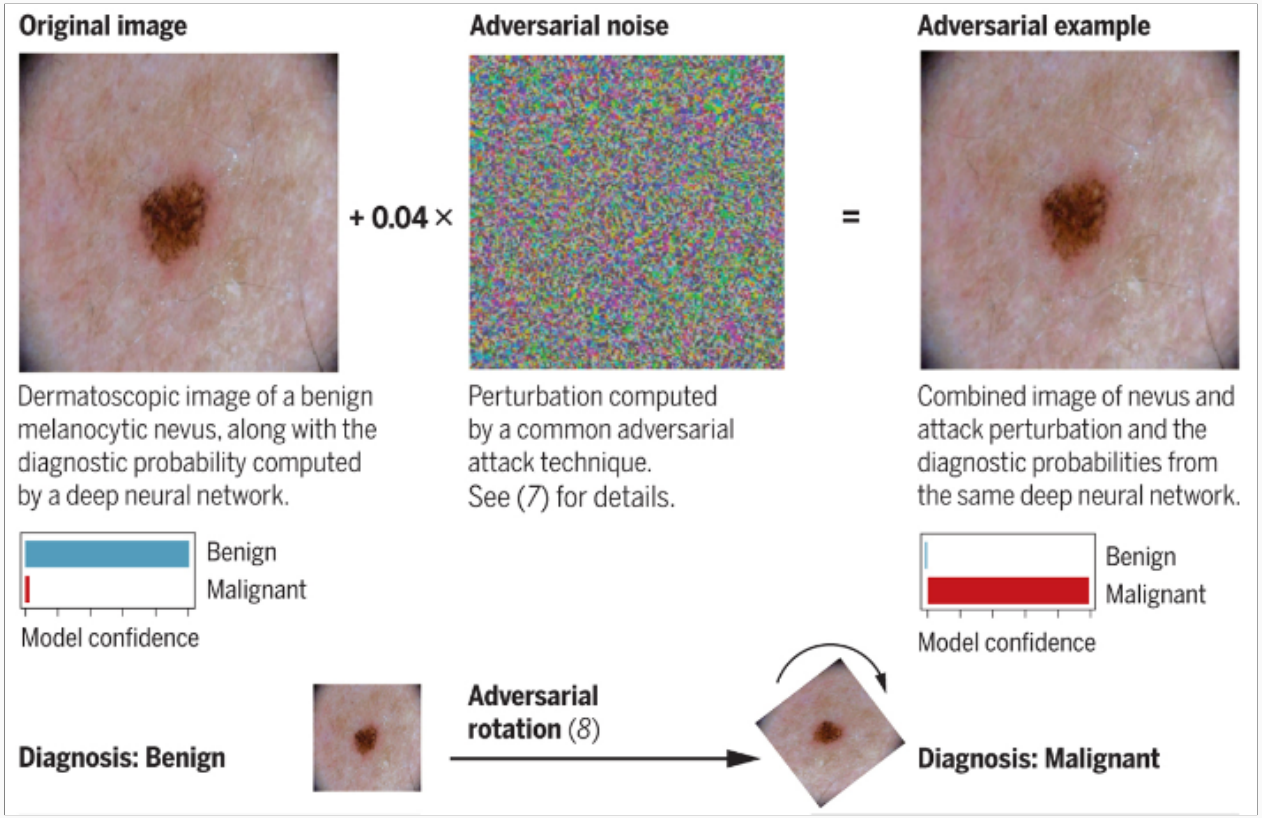


Decision-support: Interpretability



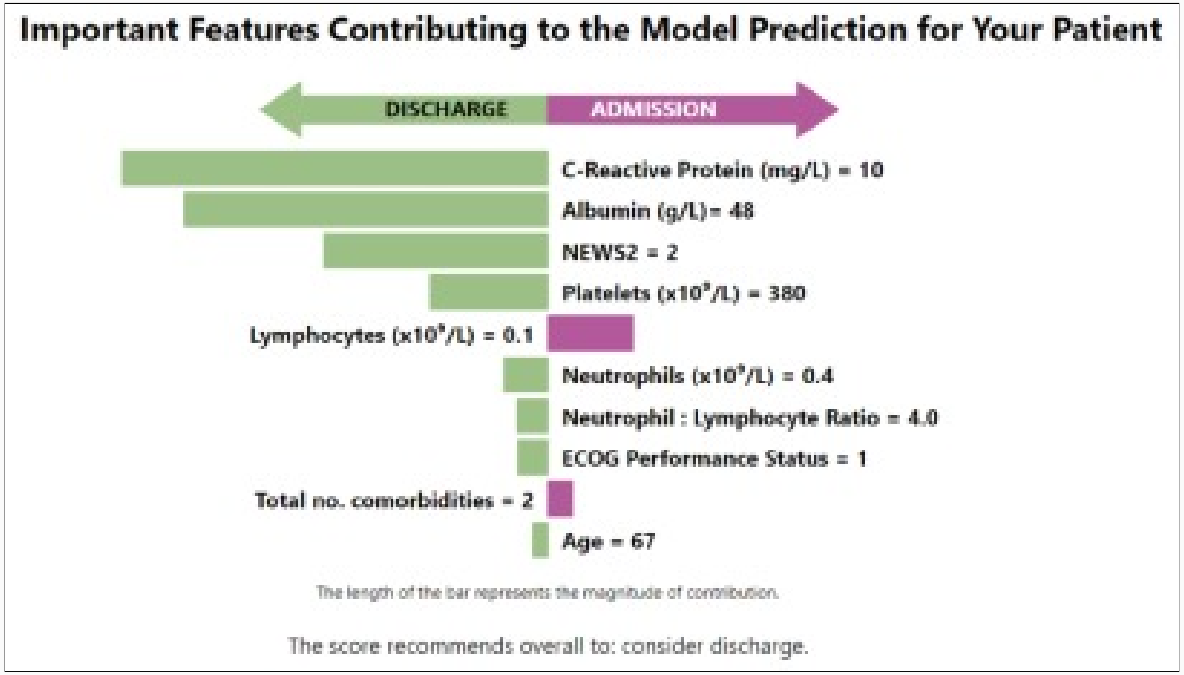
Explaining and Harnessing Adversarial Examples, Goodfellow, I. J. et al., *preprint at arXiv*, 2014

Decision-support: Interpretability

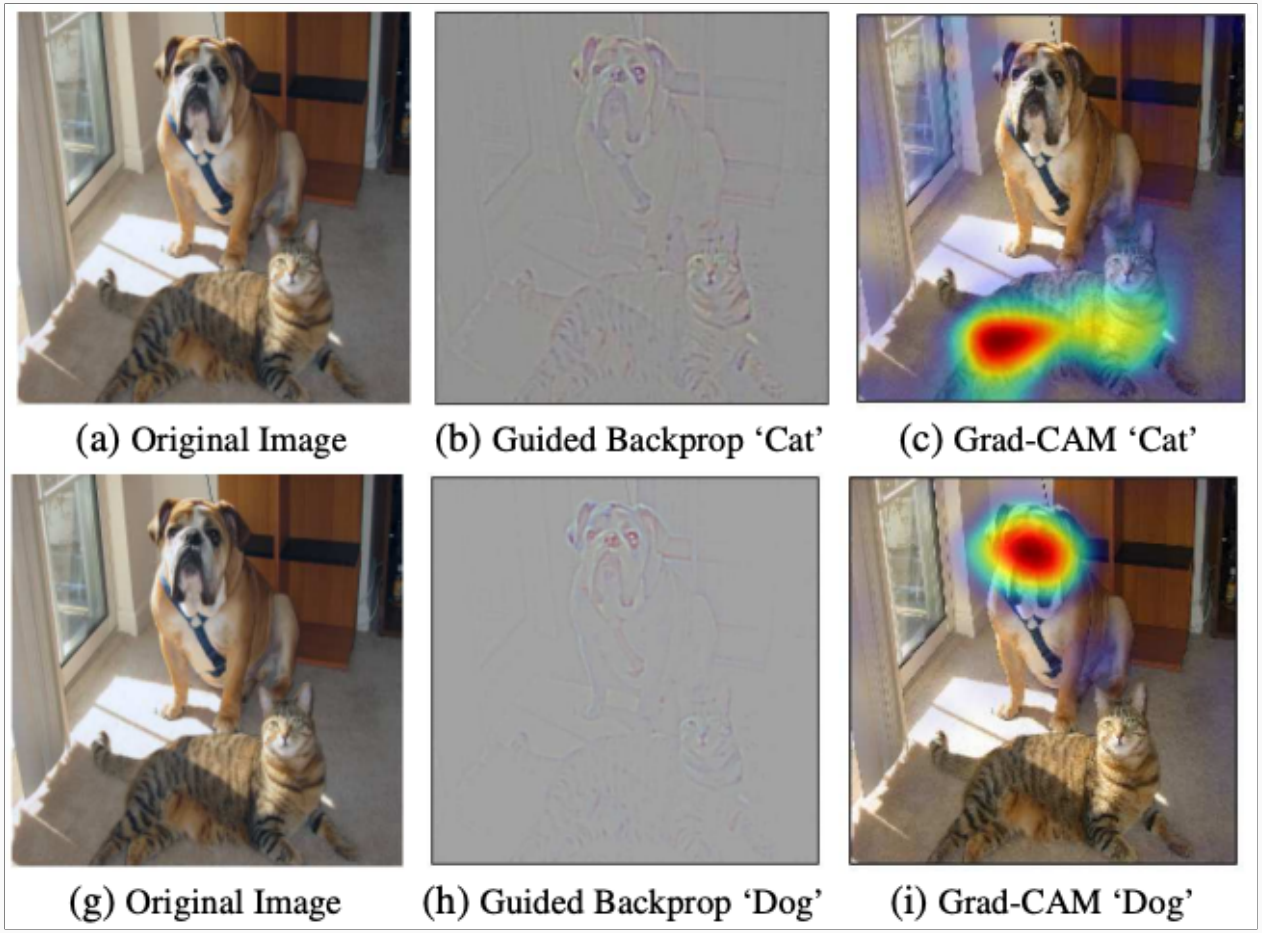


Adversarial attacks on medical machine learning, Finlayson, S. G. et al., *Science*, 2019

Decision-support: Interpretability

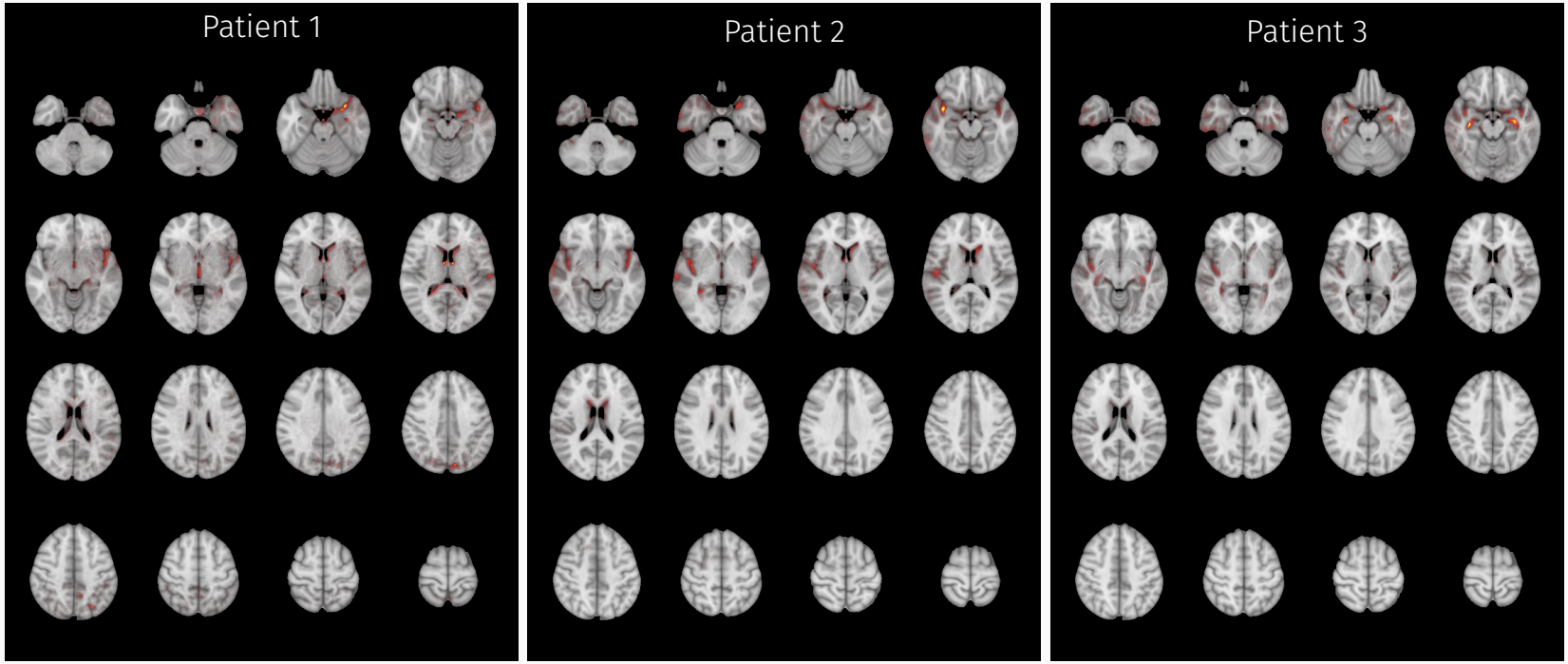


Decision-support: Interpretability



Grad-cam: Visual explanations from deep networks via gradient-based localization, Selvaraju, R. R. et al., *Proceedings of the IEEE ICCV*, 2017

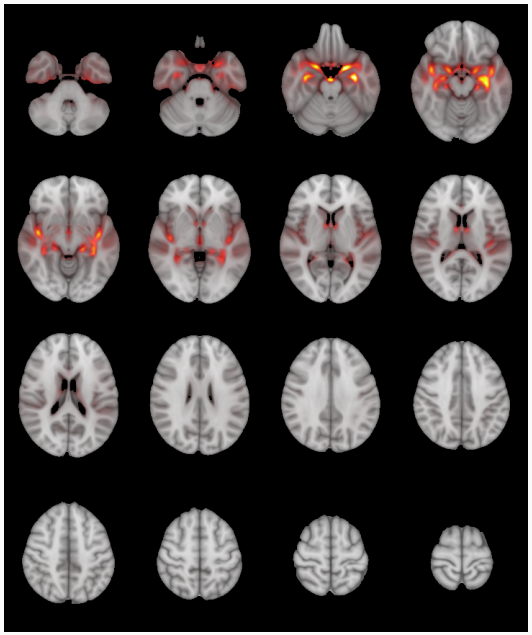
Decision-support: Interpretability



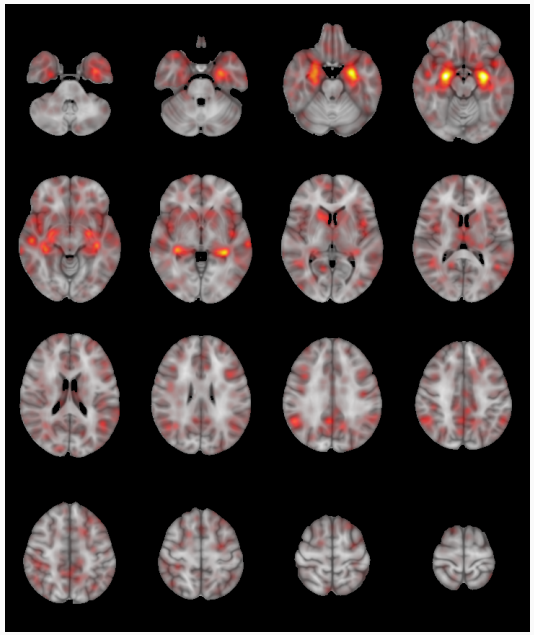
Constructing personalized characterizations of structural brain aberrations in patients with dementia using explainable artificial intelligence, Leonardsen, E. H. et al., *Digital Medicine*, 2024

Decision-support: Interpretability

AI



Human



Decision-support: Summary

AI already implemented in many domains for **decision-support**, also those considered high stakes.

- Can help improve predictive performance, and reduce time needed from domain experts.
- Lack of understanding of what underlies the decisions made by AI systems is a problem.
- Explainability is a hot topic in research, but still in its infancy.

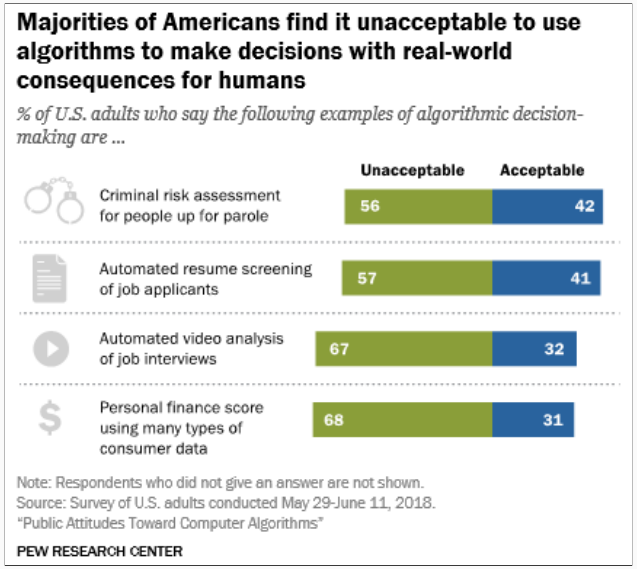
How are decisions made by AIs perceived?

UNIVERSITY
OF OSLO

Perception of AI: Skepticism

In some studies, people show low acceptance for AI making high stake decisions

- A majority of US Americans consider it unacceptable to use algorithmic decision-making in situations with real life consequences (Smith, 2018).
- 58% of Americans feel that computer programs will always reflect some level of human bias (Smith 2018).
- Concerns that they (algorithms) may violate privacy, are unfair, and **lack nuance** (Smith 2018).



Public Attitudes Toward Computer Algorithms, Smith A., *Pew Research Center Newsletter*, 2018

Perception of AI: Skepticism

In some studies, people show low acceptance for AI making high stake decisions

- A majority of US Americans consider it unacceptable to use algorithmic decision-making in situations with real life consequences (Smith, 2018).
- 58% of Americans feel that computer programs will always reflect some level of human bias (Smith 2018).
- Concerns that they (algorithms) may violate privacy, are unfair, and **lack nuance** (Smith 2018).
- Participants found it less permissible for AI to make decisions about life and death driving situations, parole (Bigman & Gray, 2018).
- Participants found it less permissible for AI to make decisions about potentially life-saving, but risky, medical procedures, than a doctor (Bigman & Gray, 2018).
- AI is seen as having less agency and experience, and thus are **less able to make moral decisions**, even when they coincide with humans (Bigman & Gray, 2018).

People are Averse to Machines Making Moral Decisions, Bigman, Y. E. & Gray, K., *Cognition*, 2018

Perception of AI: Positivism

In other studies, people show high acceptance for AI making high stake decisions (Araujo et al., 2020)

A study among a representative sample in the Netherlands asked participants to rate usefulness, fairness, and risk of AI (vs. human) decision-making in the media, health sector, and justice system.

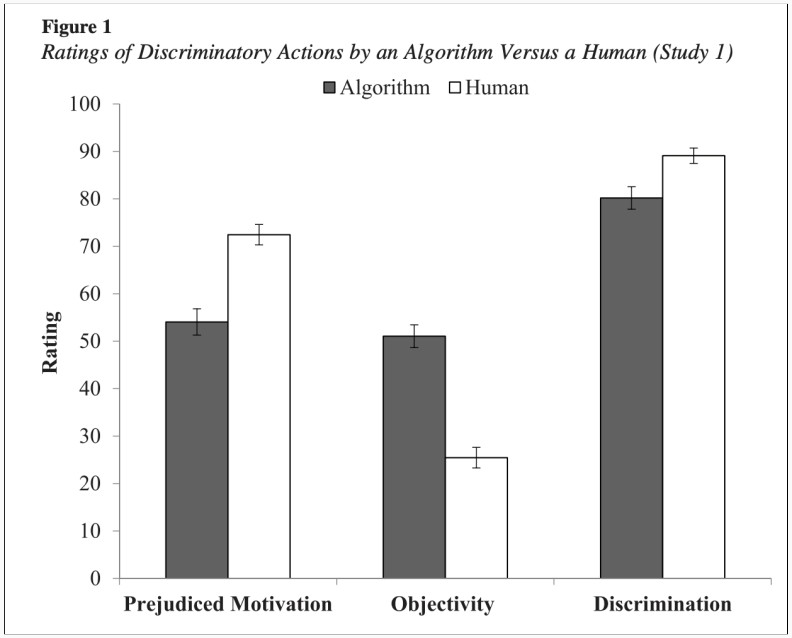
- For high-stake decisions, participants perceived decisions by AI (vs. human) to be slightly more useful, fairer and less risky in high-stake domains (typically health and justice), with no differences concering low-stakes decisions.
- Perceived usefulness increased with knowledge on AI, programming, and algorithms (self-reported).
- "... people are by and large concerned about risks and have mixed opinions about fairness and usefulness of automated decision-making at a societal level, with general attitudes influenced by individual characteristics."

In AI we trust? Perceptions about automated decision-making by artificial intelligence, Araujo, T. et al., *AI & society*, 2020

Perception of AI: Man vs machine

More moral outrage when humans discriminate than AI (Bigman et al., 2023)
Participants were asked to assess degree of discrimination, objectivity, prejudice and moral outrage after reading about a discriminatory hiring process. The discrimination was performed either by an AI or a human (HR specialist).

- When discrimination was performed by the AI, participants perceived the process as more objective, less discriminatory, and less prejudiced.



Algorithmic discrimination causes less moral outrage than human discrimination, Bigman, Y. E. et al., *Journal of Experimental Psychology*, 2023

Perception of AI: Man vs machine

More moral outrage when humans discriminate than AI (Bigman et al., 2023)
Participants were asked to assess degree of discrimination, objectivity, prejudice and moral outrage after reading about a discriminatory hiring process. The discrimination was performed either by an AI or a human (HR specialist).

- When discrimination was performed by the AI, participants perceived the process as more objective, less discriminatory, and less prejudiced.
- More moral outrage when the discrimination was performed by a human.
- Less permissible that CVs are screened by an algorithm.
- Liability of the company was smaller when the biased screening procedure was performed by an AI.

Perception of AI: Trust in AI

What predicts trust in AI (Kaplan et al., 2023)

A meta-anaylsis was performed across 65 studies that empirically investigated what leads people to trust, defined as ”the reliance by an agent that actions prejudicial to their well-being will not be undertaken by influential others” in AI.

- In humans (interacting with the AI), competency, understanding and expertise were the most important factors for facilitating trust.
- In the AI itself, reliability was the most important factor, succeeded by performance.
- Attributes such as personality, anthropomorphism, behaviour and reputation were also significant predictors of trust.
- The context of the relationship between the human and the AI was also important, with the length of the relationship the most important predictor.

Trust in artificial intelligence: Meta-analytic findings, Kaplan, Alexandra D., et al., *Human Factors*, 2023

Perception of AI: Perception of humans?

Humans overrate their ability to understand eachother (Bonezzi et al., 2022)
Participants were tasked with evaluating how well they understood the decision process of an agent (human or an AI) performing one of three tasks: (1) evaluating risk for recidivism, (2) examining video interviews, (3) examining a Magnetic Resonance Image to diagnose a disease.

- When only the decision of the agent was made available, without explanation, respondents reported a higher degree of understanding the humans.
- This difference was reduced when an explanation was provided alongside the decision.
- People project their own decision-making processes onto others.
- People overestimate their ability to understand the decision-making processes of other humans.
- Could also have a negative effect, e.g. by projecting ones own biases onto others.

The human black-box ..., Bonezzi, A., *Journal of Experimental Psychology: General*, 2023

Perception of AI: Perception of humans?

Humans overrate their ability to understand eachother (Bonezzi et al., 2022)
Participants were tasked with evaluating how well they understood the decision process of an agent (human or an AI) performing one of three tasks: (1) evaluating risk for recidivism, (2) examining video interviews, (3) examining a Magnetic Resonance Image to diagnose a disease.

- When only the decision of the agent was made available, without explanation, respondents reported a higher degree of understanding the humans.
- This difference was reduced when an explanation was provided alongside the decision.
- People project their own decision-making processes onto others.
- People overestimate their ability to understand the decision-making processes of other humans.
- Could also have a negative effect, e.g. by projecting ones own biases onto others.
- **Are we unfair when asking AIs to explain themselves? Is the only thing that matters predictive proficiency?**

Perception of AI: Summary

There is generally a large amount of variability in how people perceive AI (and automated or algorithmic systems in general), and whether they trust their decisions.

- There is a tendency towards not trusting AI to make high-stake decisions, although this varies depending on the exact task at hand, the person doing the trusting, the algorithm being trusted, and the general context.
- Although we trust AIs less, we are also less inclined to blame them (or their owners/creators) when they make mistakes, at least morally.
- Reliability and performance, both metrics of efficacy, are the most important factors for trust in AI.
- Human-like attributes in the AI increase trust.

Practice questions

Repetition

- Explain the differences between narrow and general intelligence.
- Explain how AI may lead to biased decisions, although their algorithms are objective mathematical constructs.
- Discuss the similarities and dissimilarities between human and artificial intelligence, in terms of their capacities and limitations.
- Describe why it is hard to interpret the decisions of modern AI, and what is being done to counteract this.
- How do people perceive decisions made by AI systems? Refer to two examples.

Reflection

- What kind of decisions would you be comfortable with AI making on your behalf?
- What would it take to change your view?