PSY2301: The psychology of judgement and decision making

Artificial Intelligence and decision-making

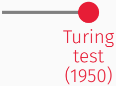
Esten H. Leonardsen

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



Alan Turing

Turing test (1950)

MIND

A QUARTERLY REVIEW

OF

PSYCHOLOGY AND PHILOSOPHY

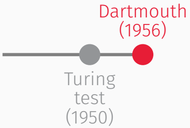
I.—COMPUTING MACHINERY AND INTELLIGENCE

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.

The history of artificial intelligence

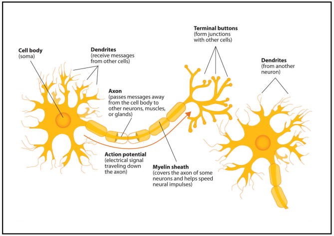


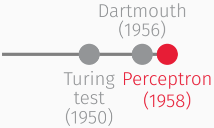
The history of artificial intelligence

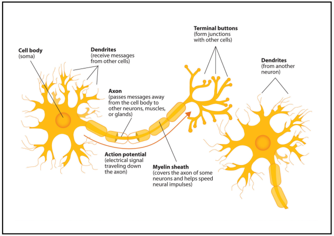
Dartmouth (1956)

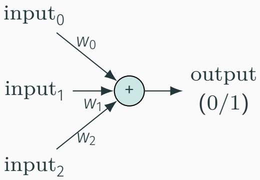
Turing test (1950)

Perceptron (1958)



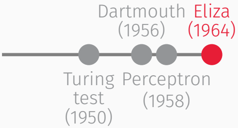


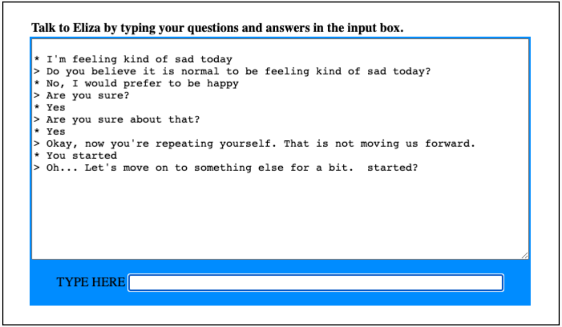




The neuron is the building block of the nervous system

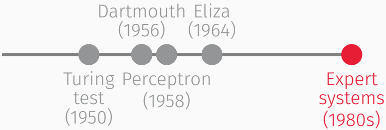
The history of artificial intelligence

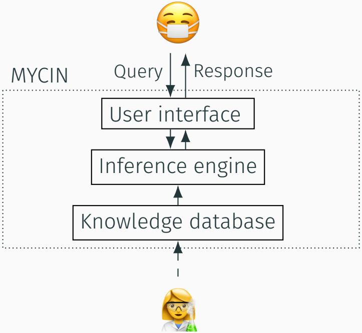




ELIZA: a very basic Rogerian psychotherapist chatbot

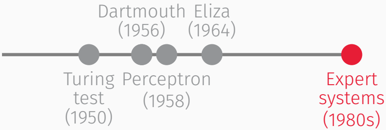
The history of artificial intelligence

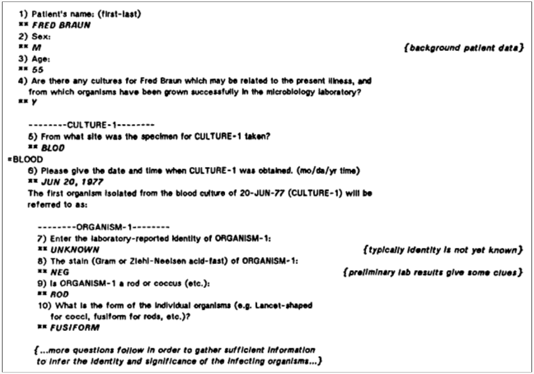




MYCIN: a knowledge-based consultation program for infectious disease diagnosis, William van Melle, *International Journal of Man-Machine Studies*, 1978

The history of artificial intelligence





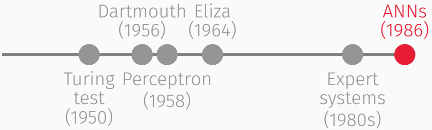
1) Patient's name: (first-last)
** *FRED BRAUN*
2) Sex:
** *M* — *{background patient data}*
3) Age:
** *55*
4) Are there any cultures for Fred Braun which may be related to the present illness, and from which organisms have been grown successfully in the microbiology laboratory?
** *y*

--------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...}

The history of artificial intelligence



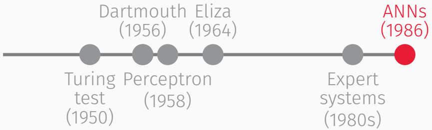
Learning representations by back-propagating errors

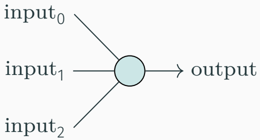
David E. Rumelhart*, Geoffrey E. Hinton† & Ronald J. Williams*

* 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

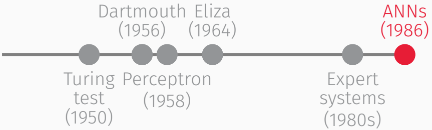
We describe a new learning procedure, back-propagation, for networks of neurone-like units. The procedure repeatedly adjusts the weights of the connections in the network so as to minimize a measure of the difference between the actual output vector of the net and the desired output vector. As a result 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].

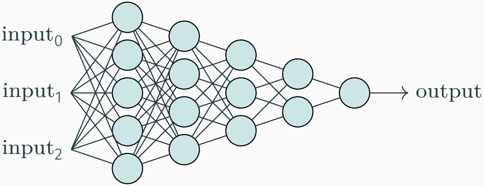
The history of artificial intelligence



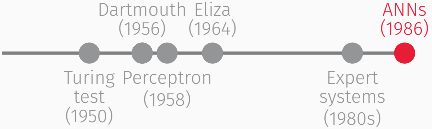


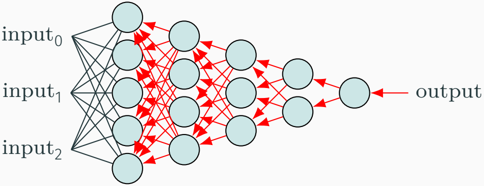
The history of artificial intelligence





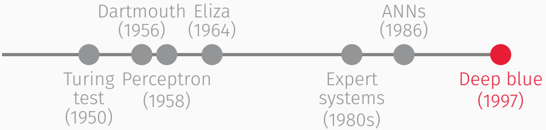
The history of artificial intelligence

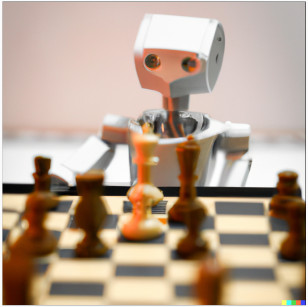




Learning representations by back-propagating errors, Rumelhart, D., Hinton, G. & Williams, R, *Nature*, 1986

The history of artificial intelligence

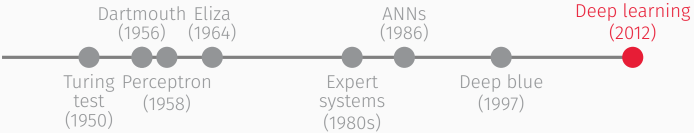




DALL-E: "A robot playing chess"

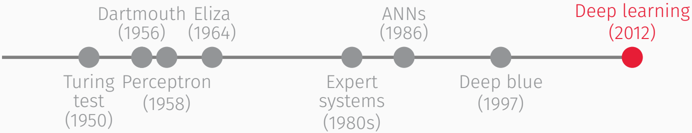
- IBMs Deep Blue became the first computer to beat the reigning human world champion in chess.
- Deep blue won with 3½ points to Garry Kasparovs 2½ after six matches.
- Kasparov famously stated that "Deep Blue was intelligent the way your programmable alarm clock is intelligent."
- Combination of machine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

The history of artificial intelligence



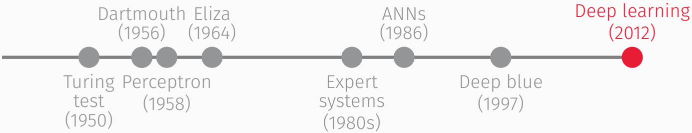
Cat

The history of artificial intelligence

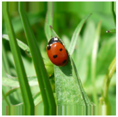


Cat

The history of artificial intelligence




Sunflower


Ladybug

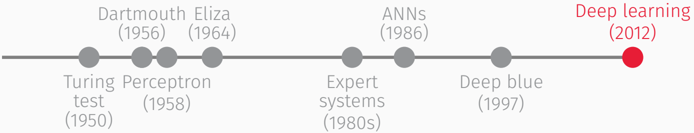

Cat

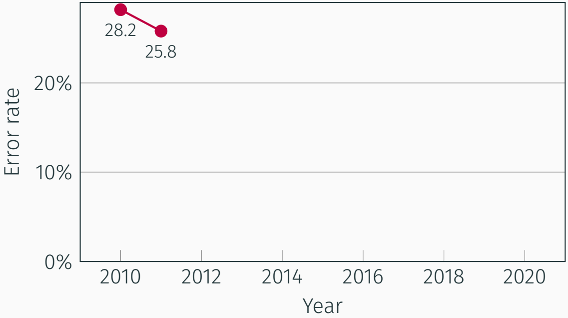

Airplane


Shark

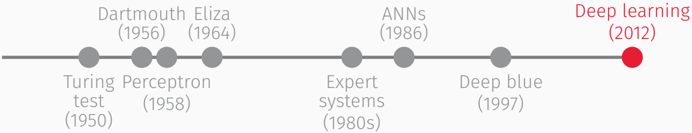
ImageNet: ∼14m images, ∼22k categories

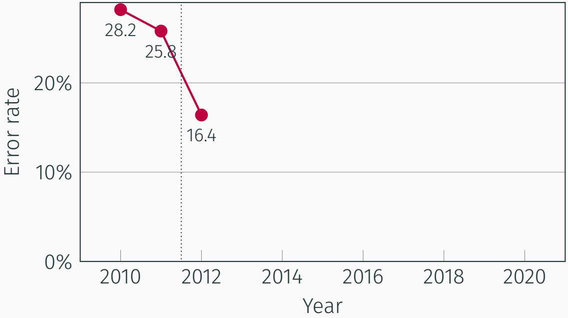
The history of artificial intelligence



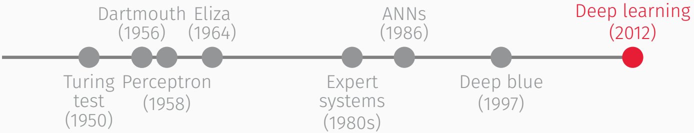


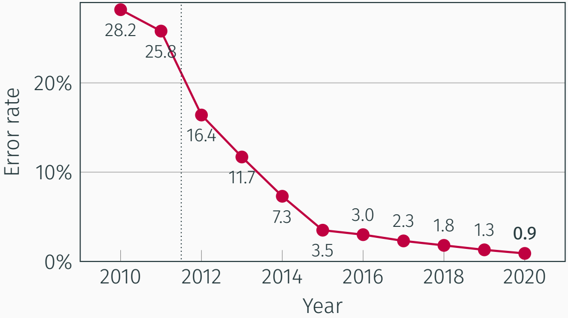
The history of artificial intelligence



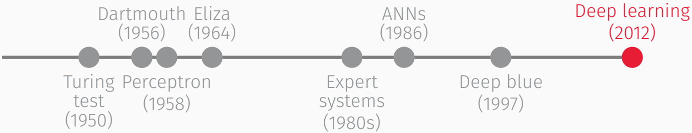


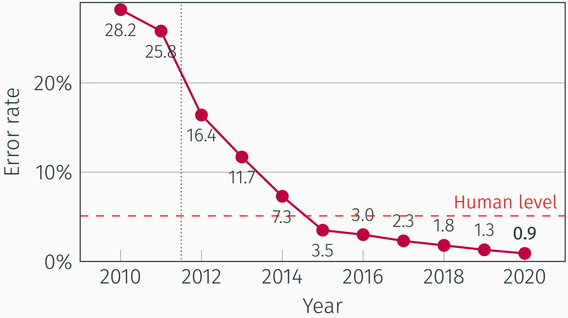
The history of artificial intelligence



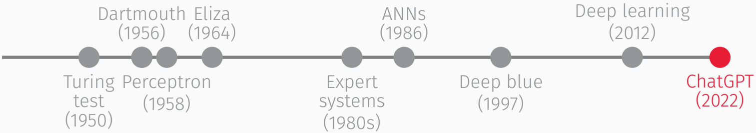


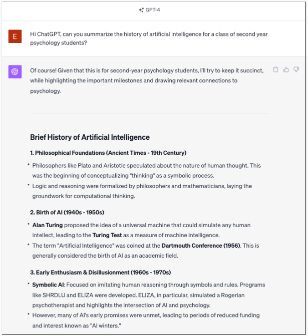
The history of artificial intelligence





The history of artificial intelligence





GPT-4

E

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" as a symbolic process.
- Logic and reasoning were formalized by philosophers and mathematicians, laying the groundwork for computational thinking.

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

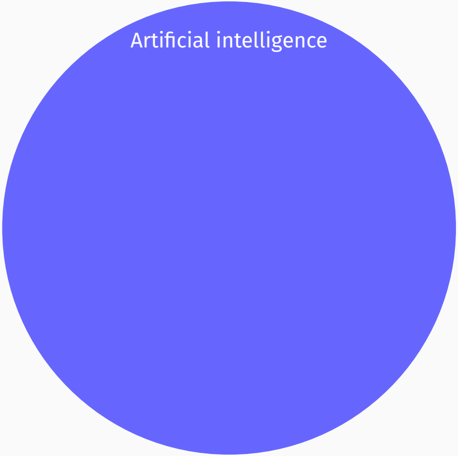
- **Alan Turing** proposed the idea of a universal machine that could simulate any human intellect, leading to the **Turing Test** as a measure of machine intelligence.
- The term "Artificial Intelligence" was coined at the **Dartmouth Conference (1956)**. This is generally considered the birth of AI 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 intersection 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."

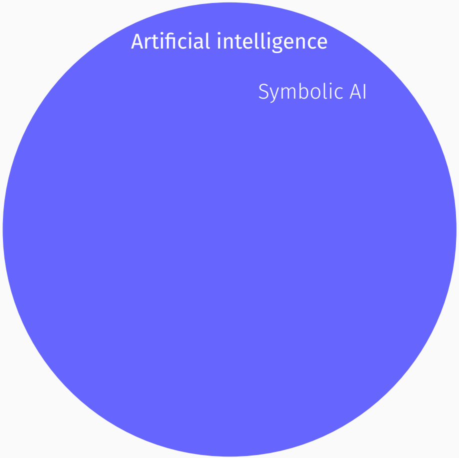
Terminology and concepts

Terminology and concepts: Taxonomy



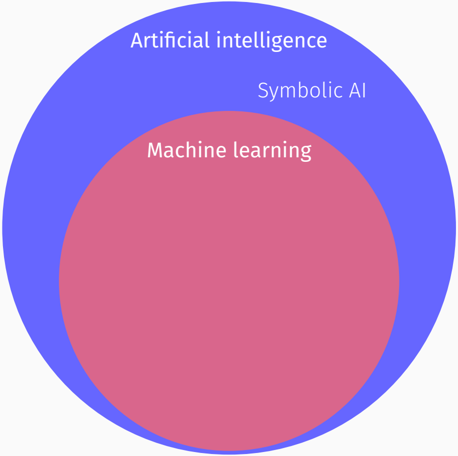
Artificial intelligence:
Machines that solve tasks requiring intelligence

Terminology and concepts: Taxonomy



Artificial intelligence:
Machines that solve tasks requiring intelligence

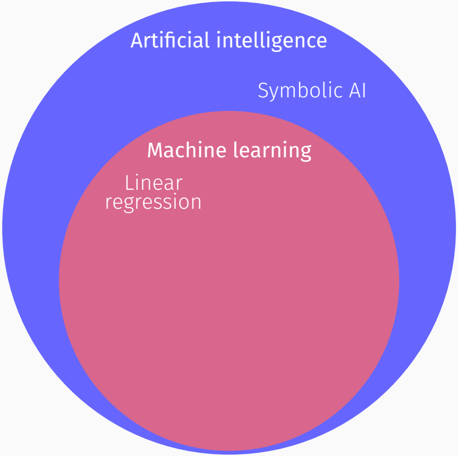
Terminology and concepts: Taxonomy



Artificial intelligence:
Machines that solve tasks requiring intelligence

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

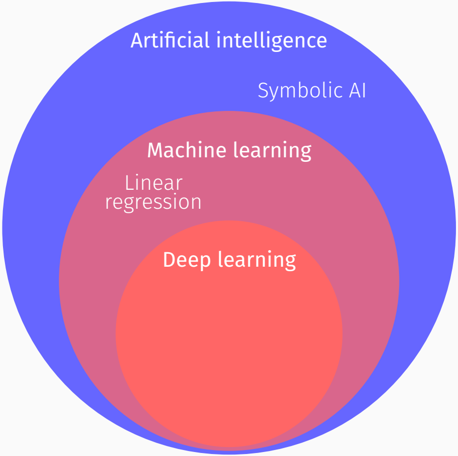
Terminology and concepts: Taxonomy



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

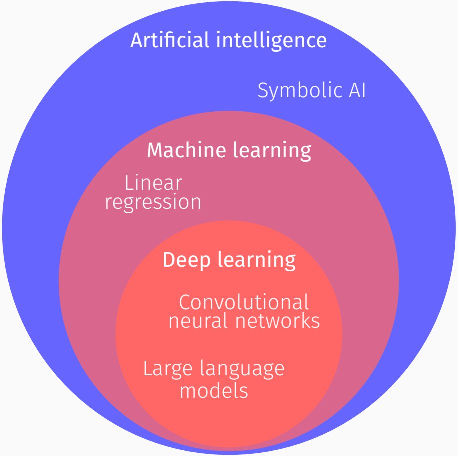


Artificial intelligence:
Machines that solve tasks requiring intelligence

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

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

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

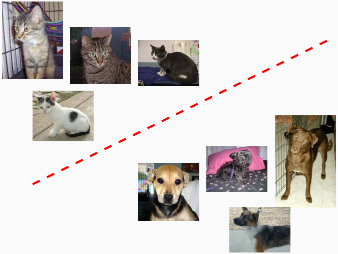
Convolutional neural nets:
Neural networks for image data

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

Terminology: Supervision

Supervised learning

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

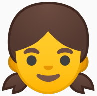
More general

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

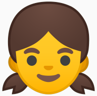
Narrow (weak) General (strong)

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

Image diagnostics

Insurance pricing

Document reading



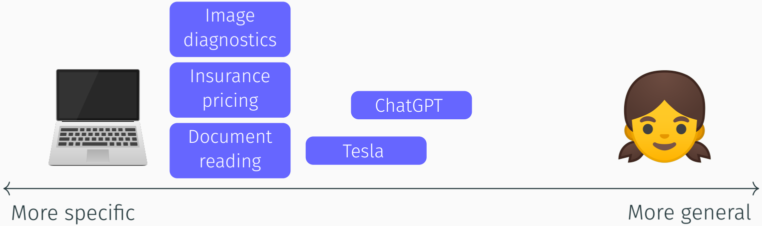
More specific More general

Terminology: Strong and weak AI

Narrow (weak)

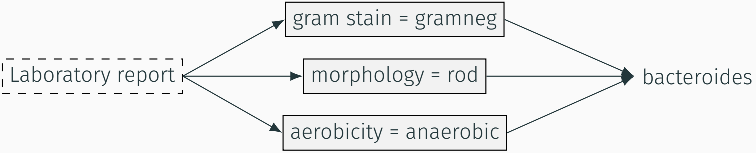
General (strong)

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



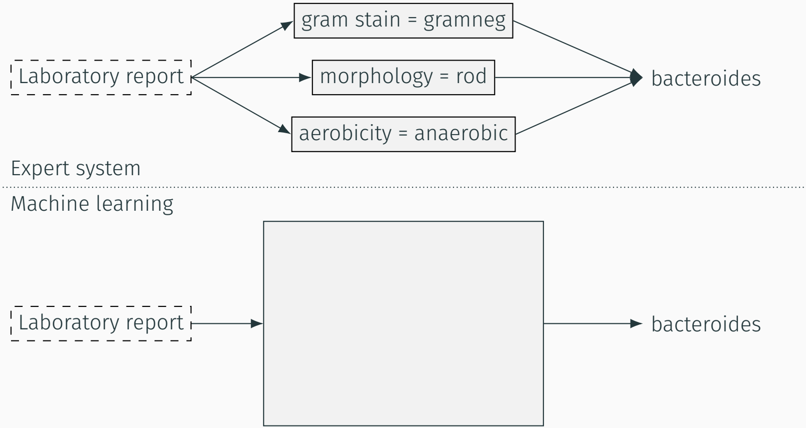
How does AI make decisions?

Decision making: Expert systems vs. machine learning

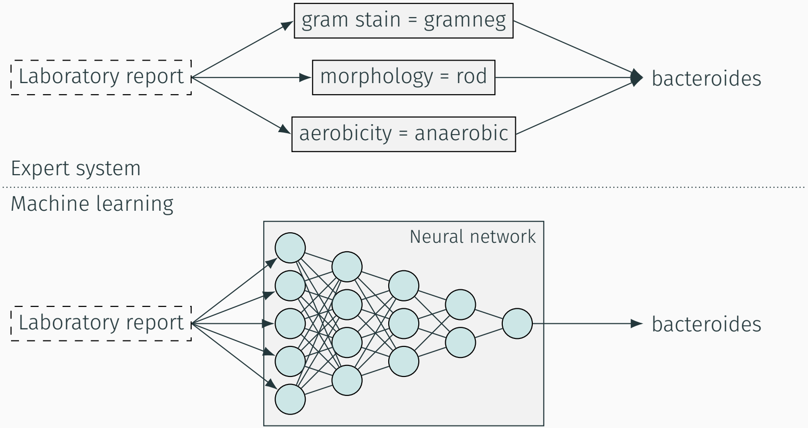


Expert system

Decision making: Expert systems vs. machine learning



Decision making: Expert systems vs. machine learning



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

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?
  $\rightarrow$ 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?

  $\rightarrow$ What is the probability that input is a cat/dog/giraffe?

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

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

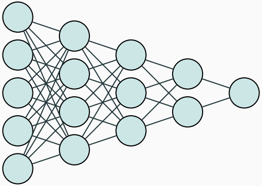
Decision making: Loss functions

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

- Classification
  What category does the input belong to?
  $\rightarrow$ What is the probability that input is a cat/dog/giraffe?
  $\rightarrow$ $-\frac{1}{N}\sum_{i=0}^{N}\left[y_i \log \hat{y}_i + (1-y_i)\log(1-\hat{y}_i)\right]$

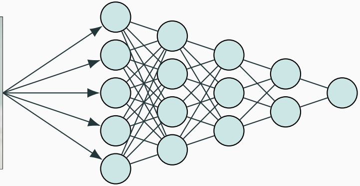
  where y_i is the correct label and $\hat{y}_i$ is the predicted probability.
- Regression
  What is the correct (continuous) output for the input?
  $\rightarrow$ $(y-\hat{y})^2$.

Decision making: Learning

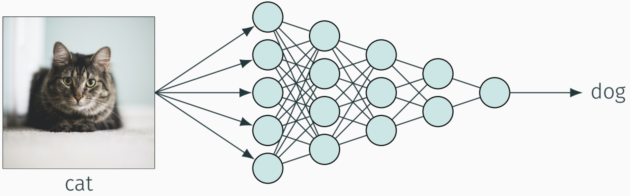


Decision making: Learning

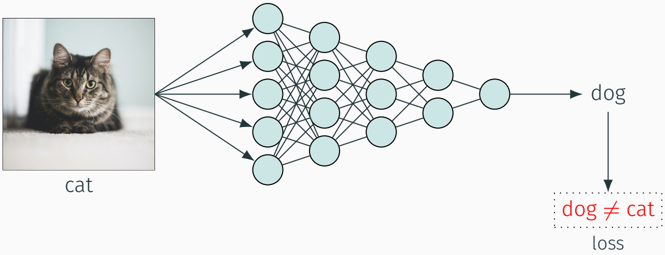

cat



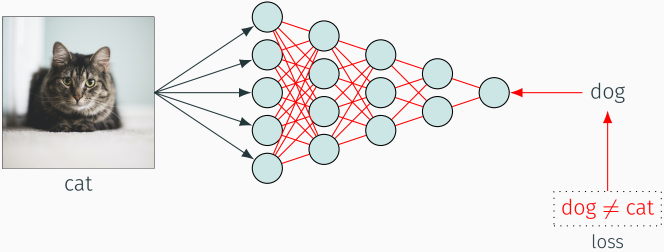
Decision making: Learning



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: Group work

We are dealing with an automatic systems in a bank that decide which clients are allowed 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 formula that takes some numbers in a give a number out. The model was trained on training data from the bank.
- Around the neural network is a software system which the user interacts with. After the user has input data, the system gives it to the neural network. If the neural network predicts a probability higher than 20%, the loan is declined. The threshold of 20% was implemented by a programmer, informed on the basis of a business analyst.

A client gets his loan declined. Who or what made the decision?

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 (called the test set).

Decision making: Generalization

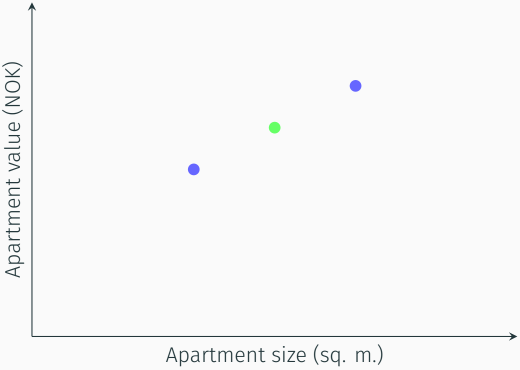
There is no guarantee the patterns the model has learned are sufficient in new data.

- ”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 (called the test set).

Decision making: Generalization

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

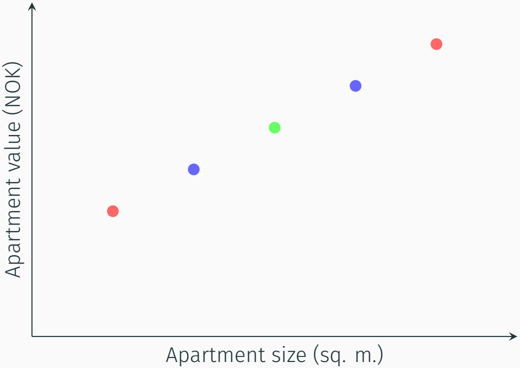
- ”AI that is based on datasets cannot go beyond what is in the data.” - Reasoning, Judging, Deciding: The Science of Thinking, Ch. 15
- While machine learning models are trained on a specific dataset (commonly referred to as the training set), they are almost always evaluated on a different dataset (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 (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 occuring 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 occuring 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.

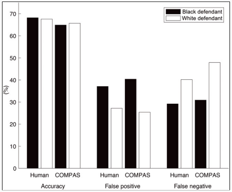
Decision making: Biases

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 occuring 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 occuring 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.

The accuracy, fairness, and limits of predicting recidivism, Julia Dressel & Hany Farid, *Science Advances*, 2018.

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 occuring 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.”

The accuracy, fairness, and limits of predicting recidivism, Julia Dressel & Hany Farid, *Science Advances*, 2018.

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 occuring 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 neighbourshoods considered higher class received more callbacks.
- Employers listing themselves as an ”Equal Opportunity Employer” were as biased as others.

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

Decision making: Theory of mind

Does AI consider humans as thinking and feeling beings?

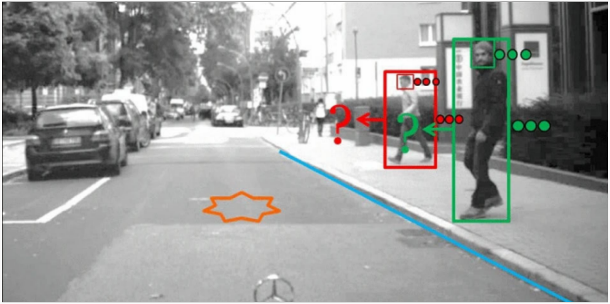
- ”… This is an instance of AI programs lacking true Theory of Mind capability.” - Reasoning, Judging, Deciding: The Science of Thinking, Ch. 15
- Theory of mind: The ability to ”track others’ unobservable mental states, such as their knowledge, intentions, beliefs, and desires.” (Kosinski 2023)

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)



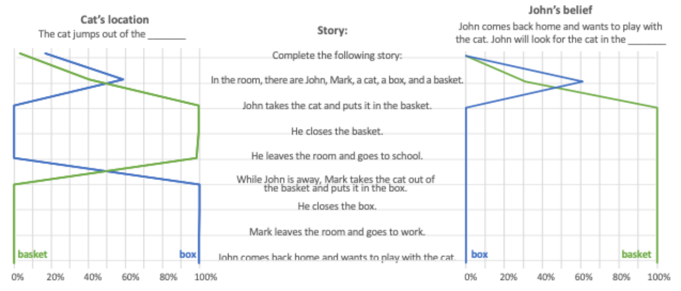
A Survey on Motion Prediction of Pedestrians and Vehicles for Autonomous Driving, Gulzar, Mahir, Yar Muhammad, and Naveed Muhammad, *IEEE Access* 9, 2021.

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)



Decision making: Creativity

Can AI create anything 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 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: Creativity

Can AI create anything 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

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 OR NOT x_2 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 satis fiability 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 straightfor ward 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.

Sparks of artificial general intelligence: Early experiments with gpt-4, Bubeck, Sébastien, et al., *arxiv*, 2023.

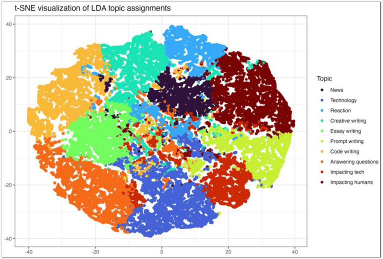
Decision making: Creativity

Can AI create anything 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

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"



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.” - easoning, 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.
- Unimodal (e.g. relies only on text), 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.

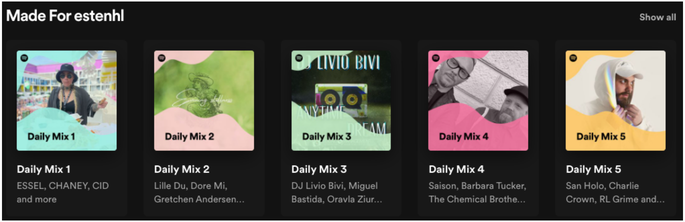
What can we expect from the decision made by AI?

- 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.

Decision support

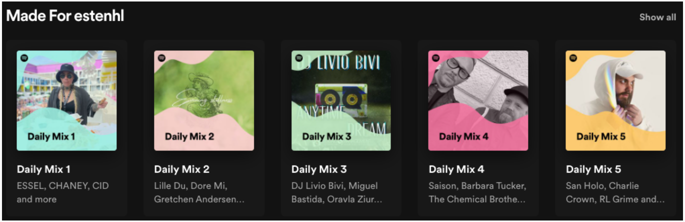
Decision support: Content personalization

Helping users decide what to listen to



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?

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

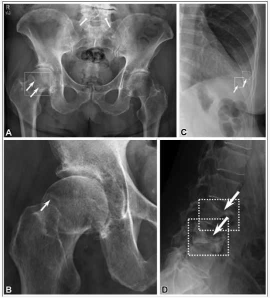
Fikk hånden analysert av kunstig intelligens: – Resultatet kom så raskt

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)



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.

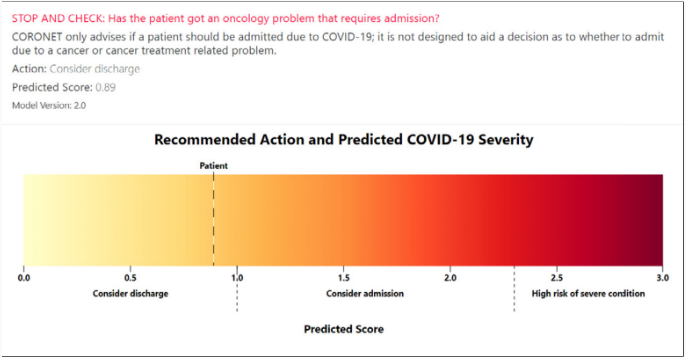
Assessing the efficacy of AI in fracture detection (Guermazi et al., 2022)

- AI assistant on its own achieved an area under the receiver operator 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**.

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

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

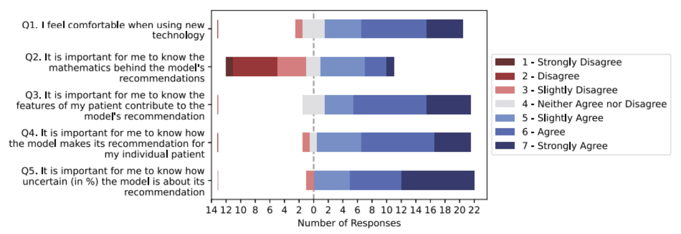
- 23 healthcare professionals tasked to assess the severity of COVID-19 in ten patients using the COVID-19 Risk in ONcology Evaluation Tool (CORONET) tool.



Decision support: COVID-19 severity

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

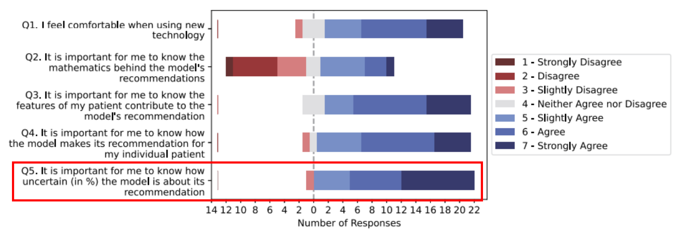
- 23 healthcare professionals tasked to assess the severity of COVID-19 in ten patients using the COVID-19 Risk in ONcology Evaluation Tool (CORONET) tool.
- Asked about their experience using the tool.



Decision support: COVID-19 severity

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

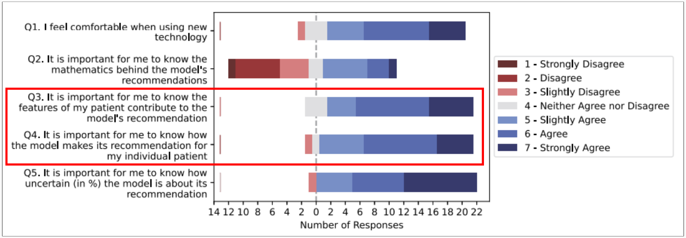
- 23 healthcare professionals tasked to assess the severity of COVID-19 in ten patients using the COVID-19 Risk in ONcology Evaluation Tool (CORONET) tool.
- Questioned about their experience using the tool.



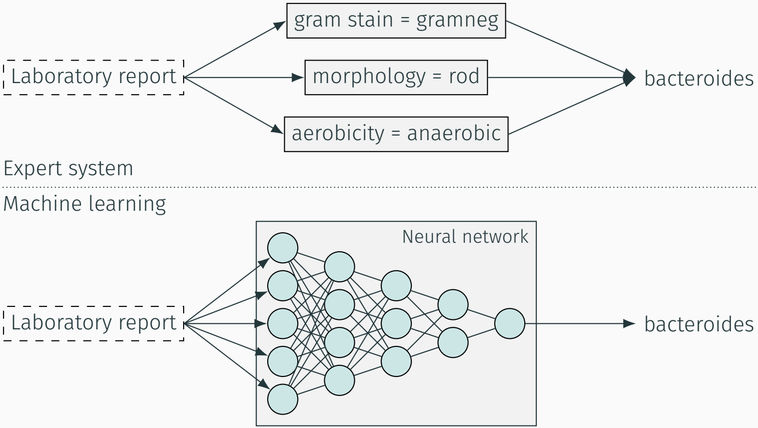
Decision support: COVID-19 severity

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

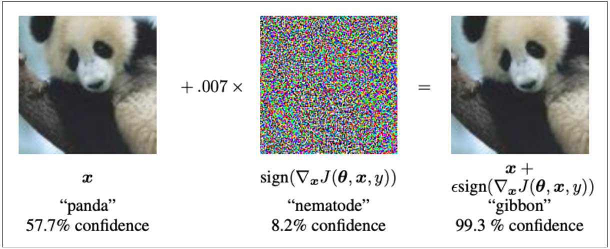
- 23 healthcare professionals tasked to assess the severity of COVID-19 in ten patients using the COVID-19 Risk in ONcology Evaluation Tool (CORONET) tool.
- Questioned about their experience using the tool.



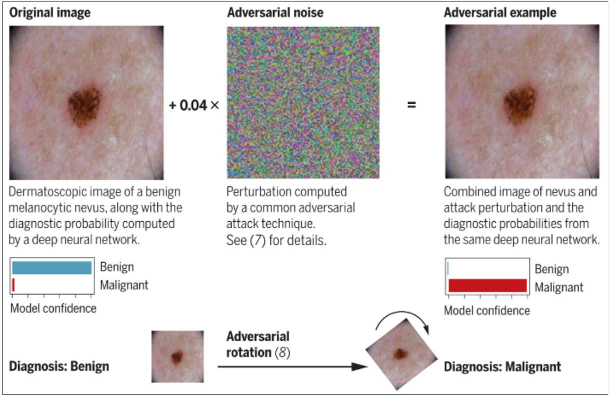
Decision support: Black boxes



Decision support: Black boxes

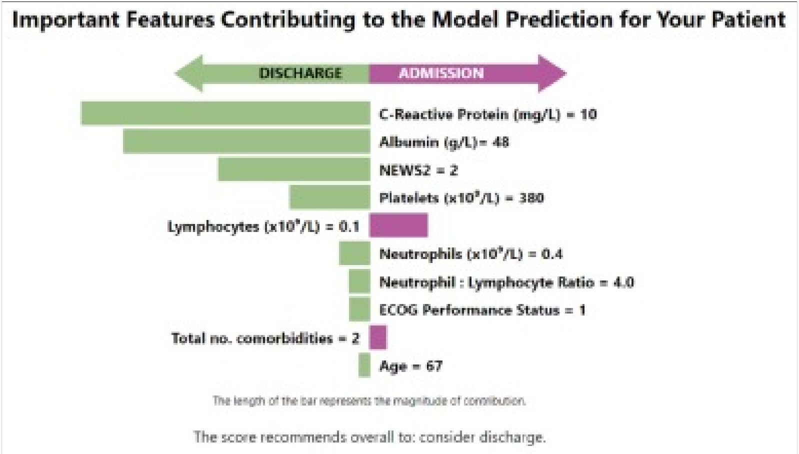


Explaining and Harnessing Adversarial Examples, Goodfellow, Ian J., Jonathon Shlens, and Christian Szegedy, *arXiv*, 2014.

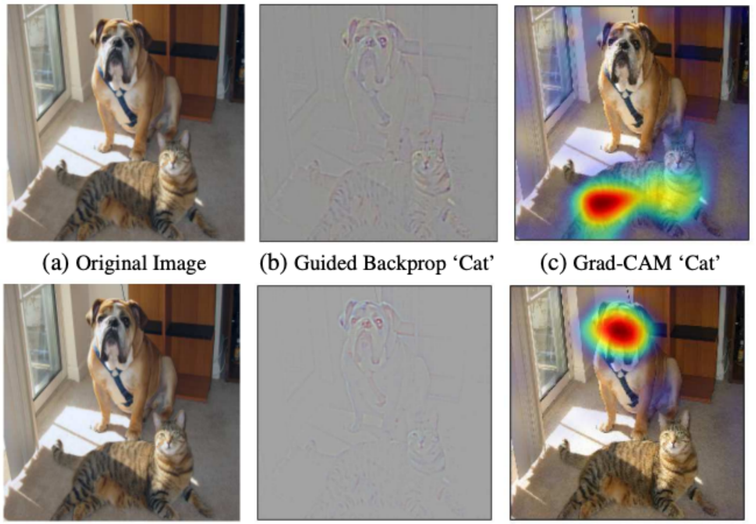


Adversarial attacks on medical machine learning, Finlayson, Samuel G., et al, *Science*, 2019.

Decision support: Explainability



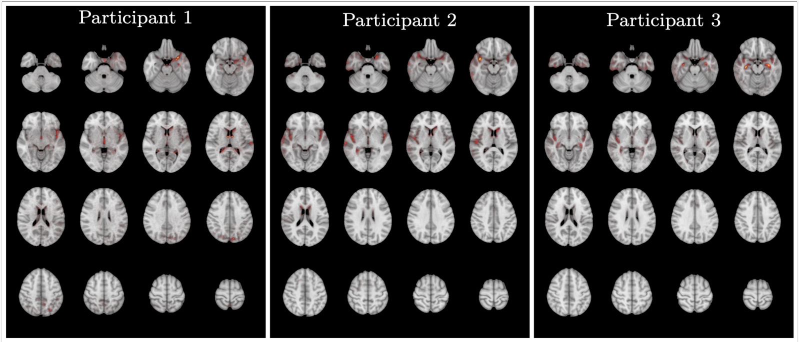
Decision support: Explainability



(a) Original Image (b) Guided Backprop 'Cat' (c) Grad-CAM 'Cat'

(g) Original Image (h) Guided Backprop 'Dog' (i) Grad-CAM 'Dog'

Decision support: Explainability



Decision support: Summary

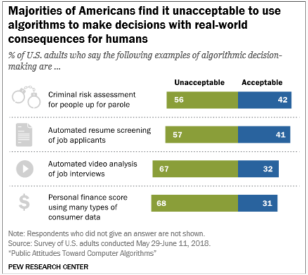
- AI already implemented in many domains for decision support, also those considered high stakes.
- Can help improve predictive performance, and reduce time.
- Lack of understanding of how the AI makes its decisions is a problem.
- Explainability is a hot topic in research, but still in its infancy.

How are decisions made by AIs perceived?

Perception of AI: Skepticism

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

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



Perception of AI: Skepticism

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

- 58% of Americans feel that computer programs will always reflect some level of human bias (Smith 2018).
- A majority of US Americans consider it unacceptable to use algorithmic decision making with real life consequences (Smith, 2018).
- Concerns that they (algorithms) may violate privacy, are unfair, and lack nuance (Smith 2018).
- AI is seen as having less agency, and thus are less able to make moral decisions (Bigman & Gray, 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).
- AI perceived to have less agency and experience, mediating the lower permissibility (Bigman & Gray, 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 more useful, fairer and less risky in health and justice contexts (no difference for low-stake decisions)
- Perceived usefulness and fairness increased with knowledge on AI, programming, and algorithms (self-reported).

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.

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 in AI. Trust defined as ”the reliance by an agent that actions prejudicial to their well-being will not be undertaken by influential others.”

- 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.
- Also attributes such as personality, anthropomorphism, behaviour and reputation were 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.

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.
- 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?

The human black-box ..., Kaplan, Bonezzi, Andrea, Massimiliano Ostinelli, and Johann Melzner, *ournal of Experimental Psychology: General*, 2022.

Perception of AI: Summary

- 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.

Perception of AI

The robot in the museum.

Decision making: Group work

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

Practice questions

- 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 AI decision making? Refer to two examples.