Hoe werkt AI?
&
Hoe kan ik AI verantwoord inzetten?

Meike Nauta

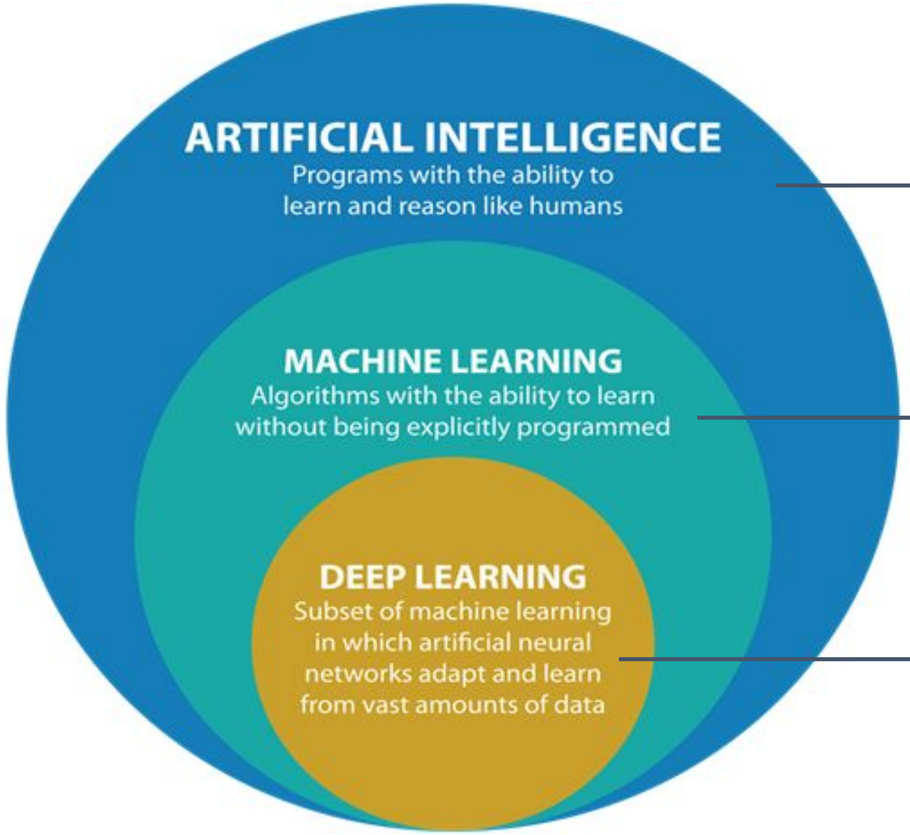
FHI, september 2023

m.nauta@utwente.nl

linkedin.com/in/meikenauta/

AI, Machine Learning, Deep Learning

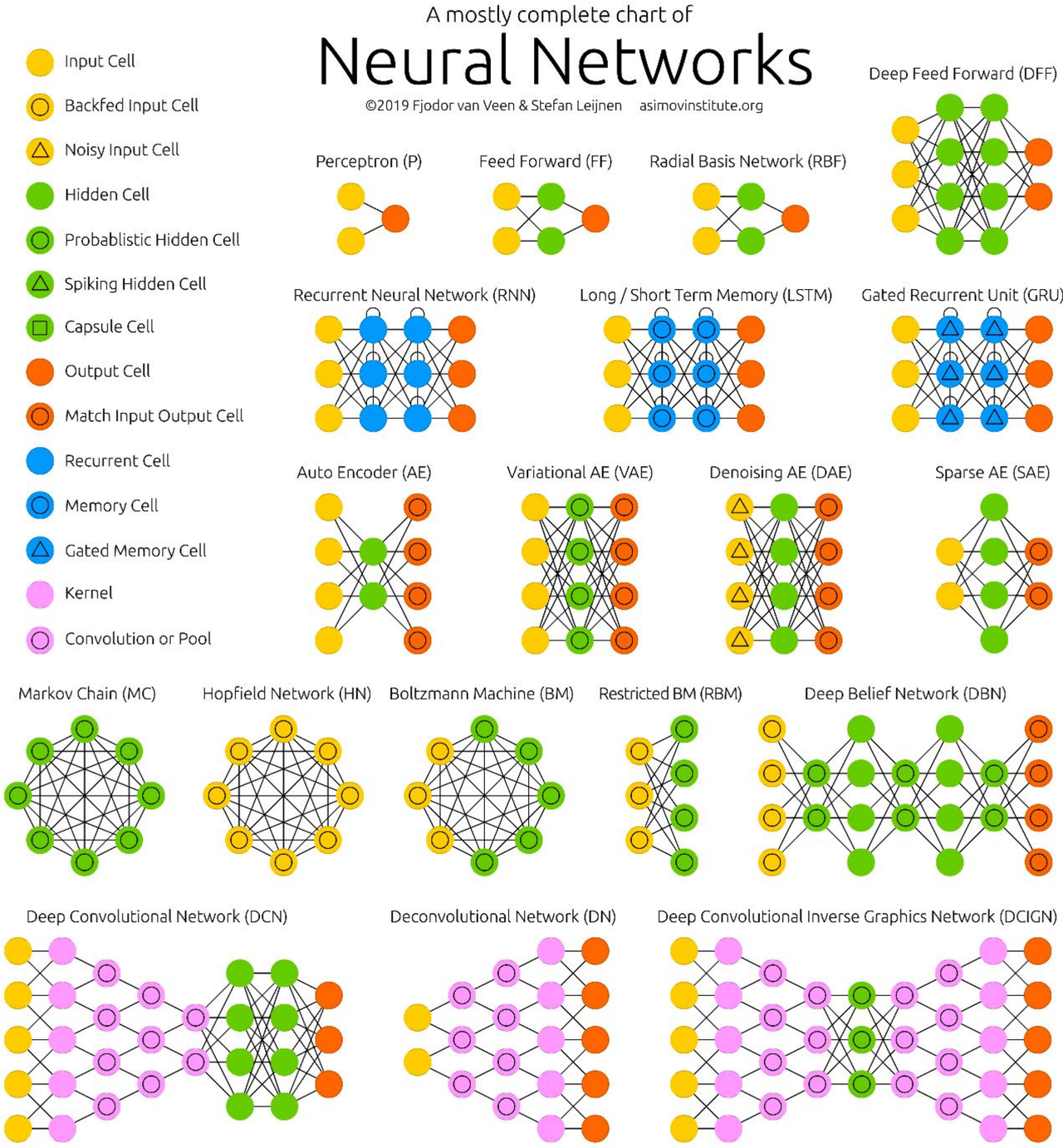
Definities

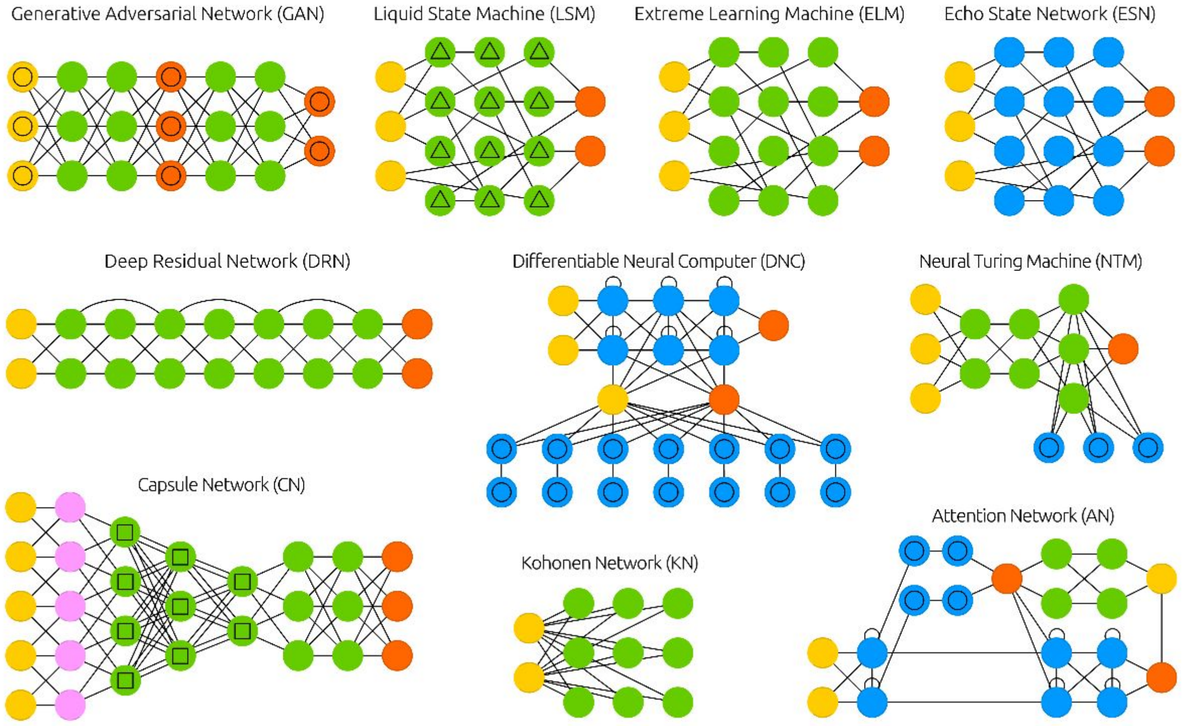


Slecht gedefinieerd, alles gerelateerd aan niet-menselijke intelligentie

Zelflerende algoritmes: leren van data zonder menselijke interventies.
Er bestaan veel verschillende soorten machine learning modellen

Machine learning model is een diep kunstmatig neuraal netwerk





ChatGPT

Onderdeel van
'**Transformer**'
neuraal netwerk

Data in, Data uit

Zelflerende Algoritmes

DATA

Table / Database / Excel sheet

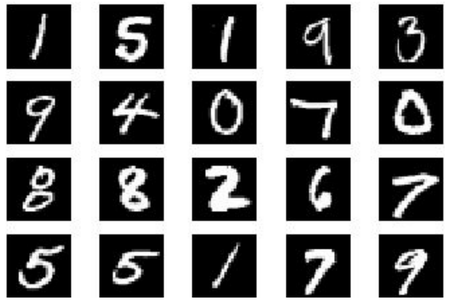
Images

Text / Publications

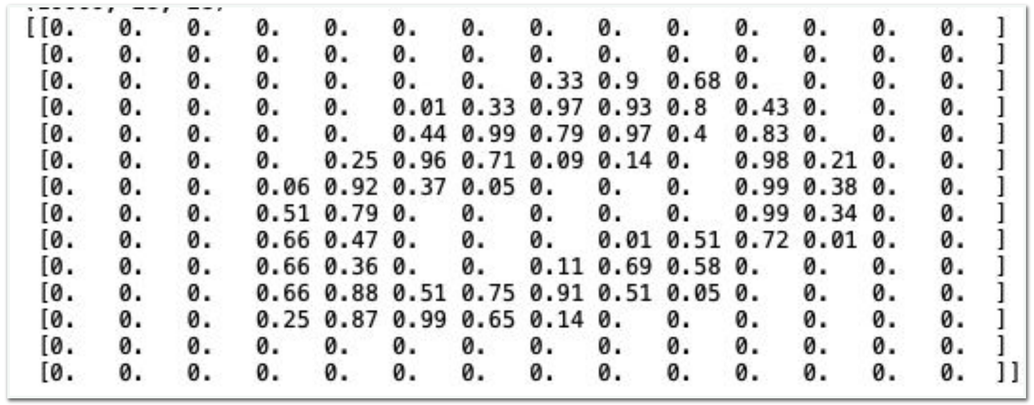
Sensor Data

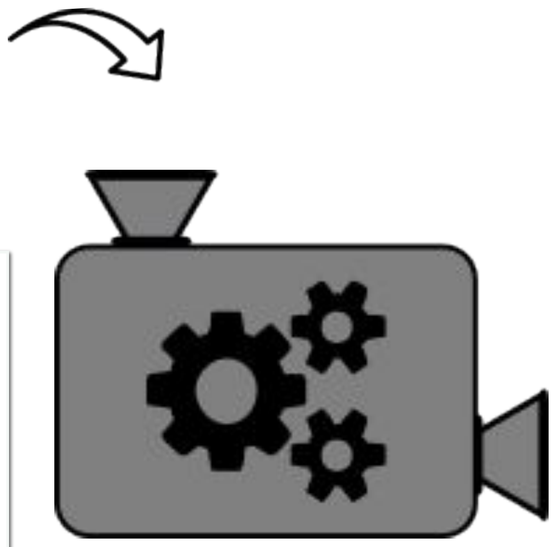
Web

Linked Data



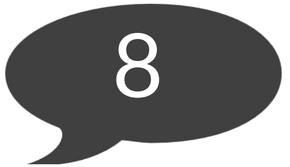
Input x

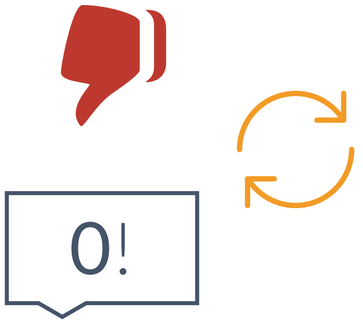




Zelflerend algoritme

Output y (voorspelling)





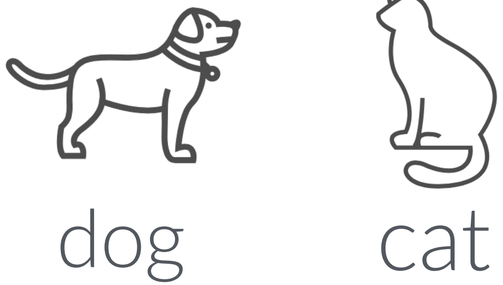
- Classificatie
- Detectie
- Segmentatie
- Generatie

Types of (Machine) Learning

SUPERVISED LEARNING

Teacher gives direct, concrete feedback

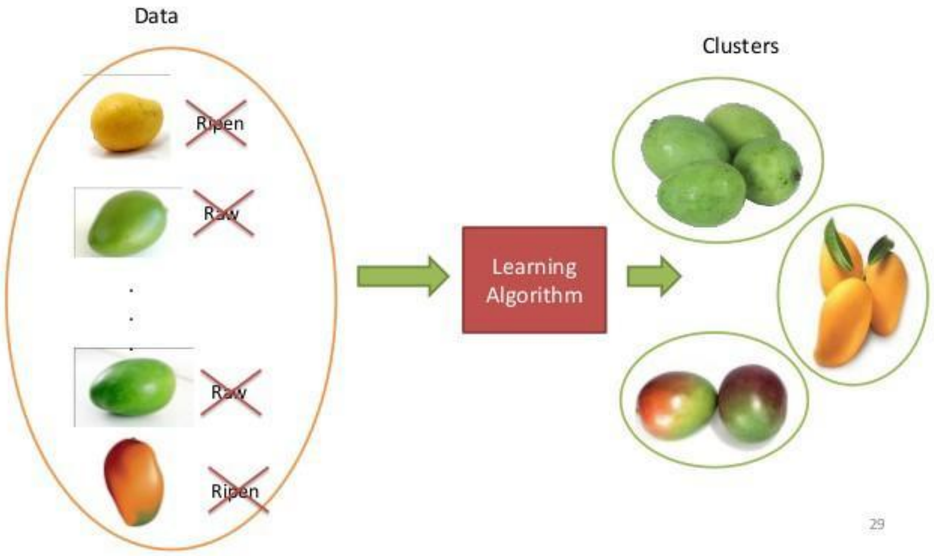
Often used, needs labelled data. Useful for classification, regression, segmentation, detection, …



UNSUPERVISED LEARNING

No teacher, no feedback

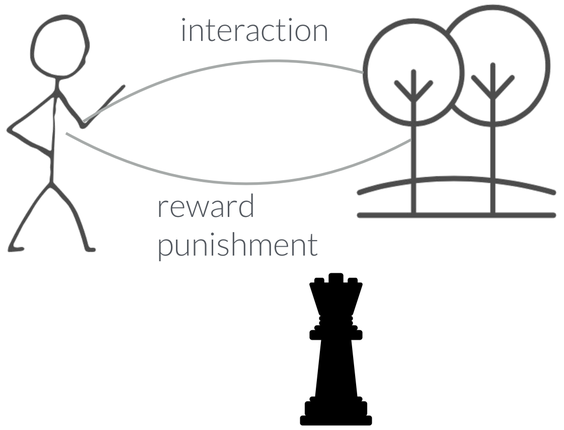
Useful for clustering and data exploration



REINFORCEMENT/ EVOLUTIONARY LEARNING

No teacher, delayed/unspecific feedback

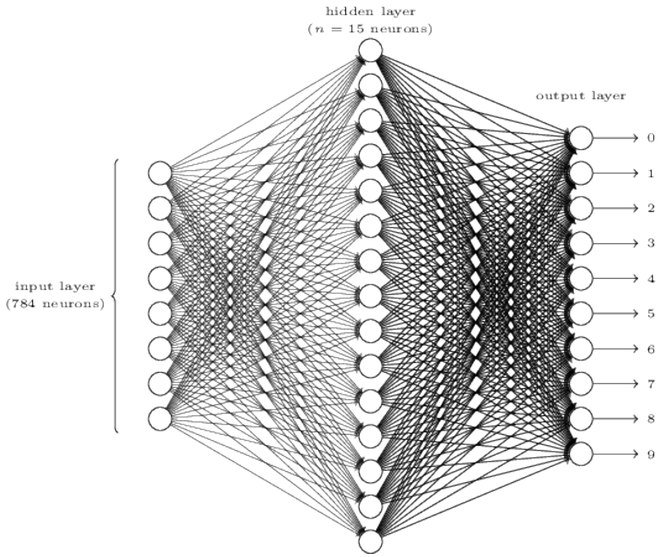
Often used for optimization and generation



Explainable AI (XAI) to the Rescue

Deep Neural Networks become black boxes

Complex neural networks have over 1 BILLION weights!
Not interpretable anymore



Explainable AI

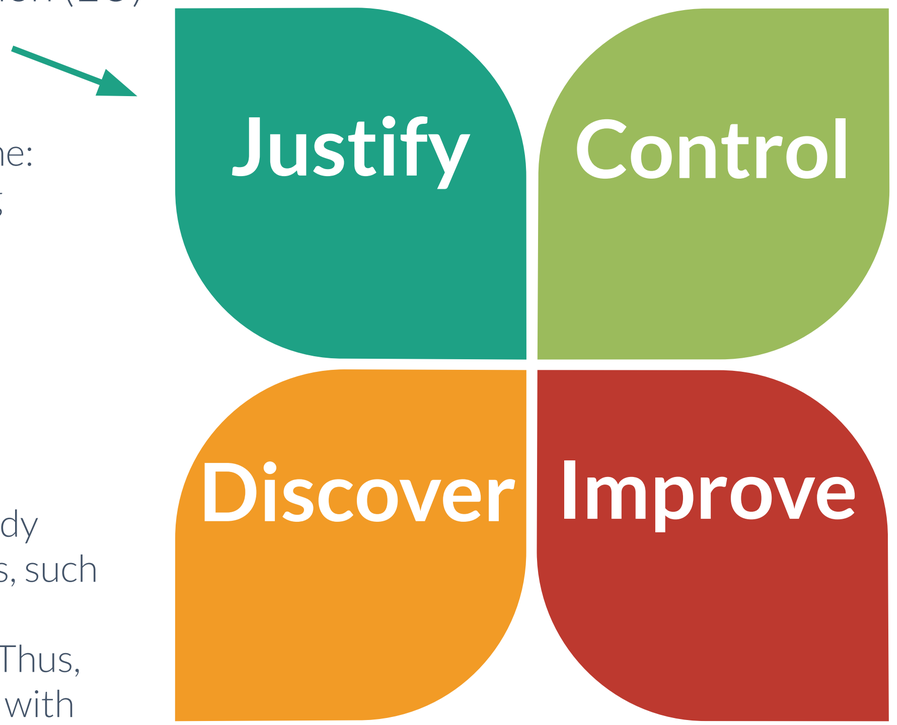
An **explanation** is a presentation of the reasoning of a machine learning model in **human-understandable terms**.

Source: Nauta et al. "From anecdotal evidence to quantitative evaluation methods: A systematic review on evaluating explainable AI." ACM Computing Surveys 55.13s (2023): 1-42.

Need for Explainable AI

Explainable Artificial Intelligence (XAI)

GDPR (Algemene verordening gegevensbescherming): Elke betrokkene dient dan ook het recht te hebben, te weten en te worden meegedeeld […] **welke logica er ten grondslag ligt aan een eventuele automatische verwerking van de persoonsgegevens”**, Regulation (EU) 2016/679.



Justify

Justifications for a certain outcome: getting insight into the underlying reasoning of a decision.

Control

Understanding the system’s behavior to identify previously unknown flaws.
Is the model right for the right reasons (no bias/discrimination)?

Discover

Machine learning algorithms already outperform humans on many tasks, such as playing the game of Go [1] and identifying cervical precancer [2]. Thus, explainable models can provide us with new knowledge.

Improve

Using the explanations of the systems, users can make the system smarter.
Human-in-the-loop development: ongoing iteration and improvement between human and machine.

[1] D. Silver et al., Mastering the game of go with deep neural networks and tree search," nature, vol. 529, no. 7587, p. 484, 2016.
[2] L. Hu et al., An observational study of deep learning and automated evaluation of cervical imagesfor cancer screening," JNCI: Journal of the National Cancer Institute, 2019.

Adapted from : Adadi, & Berrada (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). IEEE access, 6, 52138-52160.

XAI reveals Shortcut Learning

Wat heeft het AI model geleerd?

Fractuurdetectie

No fracture

Fracture

Verborgen shortcut in de data:
dit type röntgenfoto wordt alleen gemaakt van immobiele patiënten op de spoedeisende hulp! Het AI-model zal dus bijna altijd goed gokken dat er een breuk op zo'n foto aanwezig is.

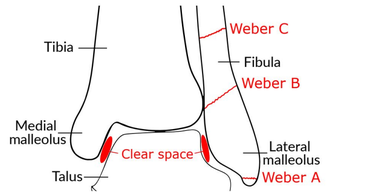
AI leert altijd de makkelijkste route om van A naar B te komen!

Source: Nauta et al., "Interpreting and Correcting Medical Image Classification with PIP-Net", XI-ML workshop @ ECAI, 2023

Interpretable Image Classification

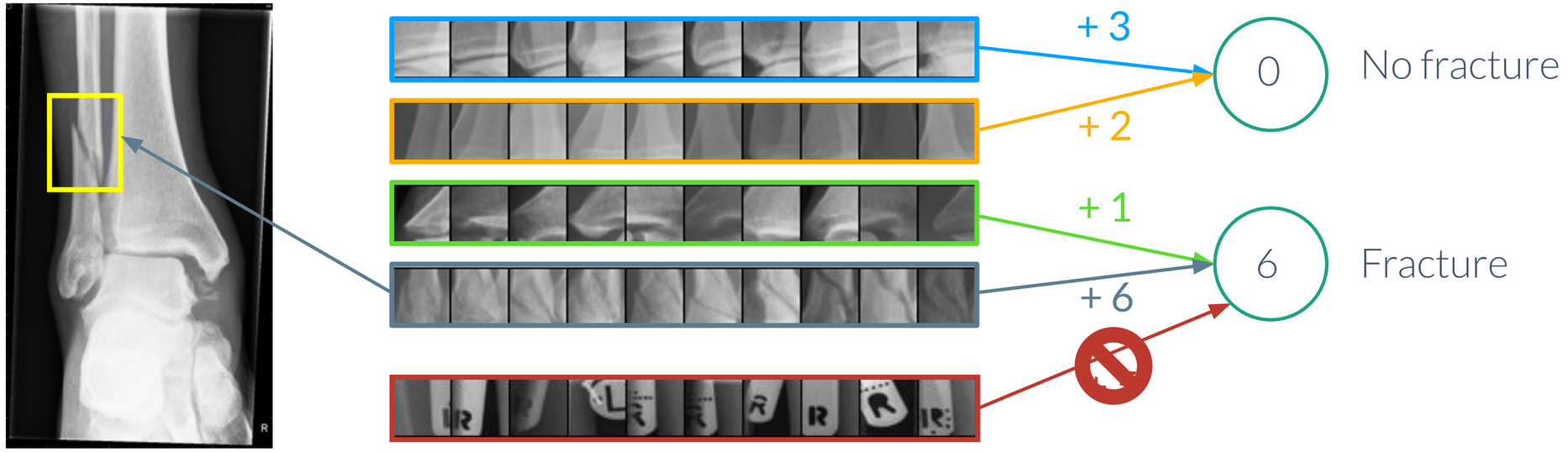
With part-prototype learning

Fracture detection



(a) Classification standard for ankle fractures

Few Learned Part-Prototypes



Justify

"Why did the model give this prediction?"

Validate

"Is the model right for the right reasons?"

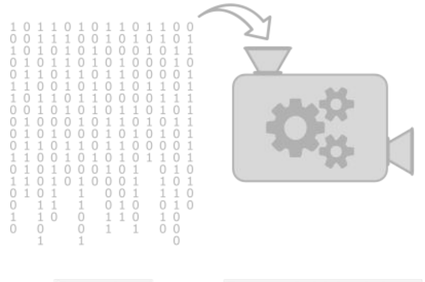
Discover

"Do the explanations reveal new information?"

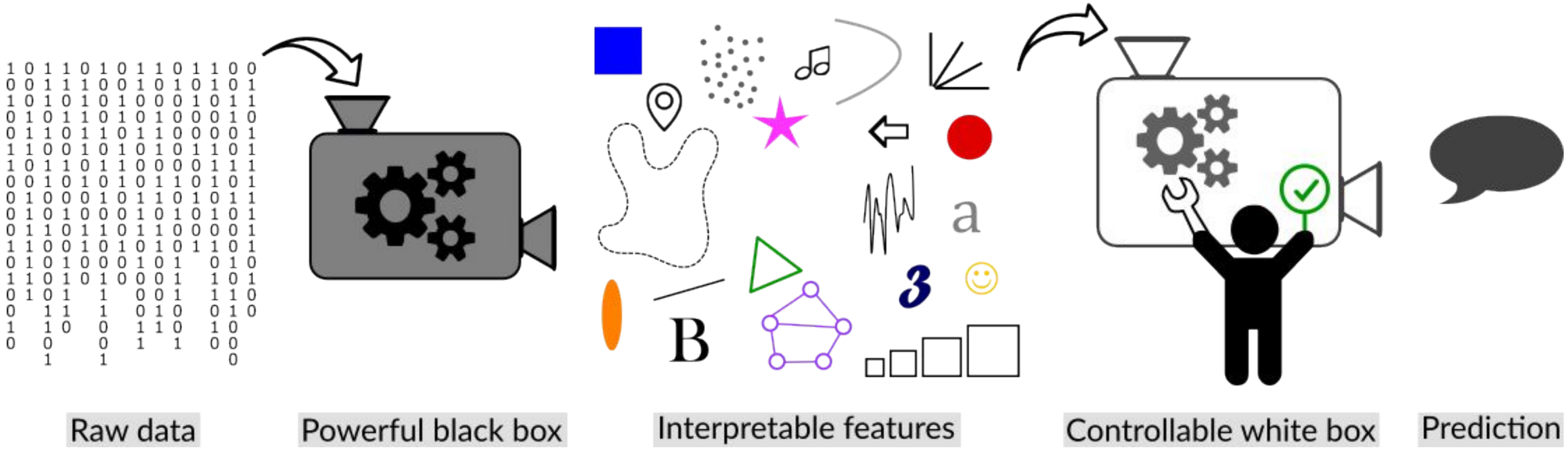
Source: Nauta et al., "Interpreting and Correcting Medical Image Classification with PIP-Net", XI-ML workshop @ ECAI, 2023

Future of Responsible AI

With in-model interpretability



Raw data Powerful black box Prediction



Raw data | Powerful black box | Interpretable features | Controllable white box | Prediction

→ let the user check and **adapt** the prediction model

Source: Conclusion of PhD thesis Meike Nauta, "Explainable AI and Interpretable Computer Vision: From Oversight to Insight" (2023)

Wat en waarvan heeft AI geleerd?

de Correspondent

Update • 2 dagen geleden • Leestijd 4 minuten

Veel bedrijven die met kunstmatige intelligentie bezig zijn, noemen zichzelf ‘open’: ze zijn transparant over wat ze doen en hun software is voor iedereen gratis toegankelijk. In werkelijkheid is dat vooral een marketingstrategie. ‘Alsof je een vliegreis duurzaam noemt omdat de maaltijd aan boord vegetarisch is.’

De wassen neus van ‘open’ kunstmatige intelligentie

https://opening-up-chatgpt.github.io/

There is a growing amount of instruction-tuned text generators billing themselves as 'open source'. How open are they really? paper repo

Dingemanse, M. 2023. "Opening up ChatGPT: Tracking Openness, Transparency, and Accountability in Instruction-Tuned Text Generators." In CUI '23: Proceedings of the 5th International Conference on Conversational User Interfaces. July 19-21, Eindhoven.

Project	Availability						Documentation						Access	
(maker, bases, URL)	Open code	LLM data	LLM weights	RLHF data	RLHF weights	License	Code	Architecture	Preprint	Paper	Modelcard	Datasheet	Package	API
BLOOMZ	✓	✓	✓	✓	~	~	✓	✓	✓	✗	✓	✓	✗	✓
bigscience-workshop	LLM base: BLOOMZ, mT0			RL base: xP3										§
Pythia-Chat-Base-7…	✓	✓	✓	✓	✗	✓	✓	✓	~	✗	~	~	✓	✗
togethercomputer	LLM base: EleutherAI pythia			RL base: OIG										§
Open Assistant	✓	✓	✓	✓	✗	✓	✓	✓	~	✗	✗	✗	✓	✓
LAION-AI	LLM base: Pythia 12B			RL base: OpenAssistant Conversations										§
dolly	✓	✓	✓	✓	✗	✓	✓	✓	~	✗	✗	✗	✓	✗
databricks	LLM base: EleutherAI pythia			RL base: databricks-dolly-15k										§
RedPajama-INCITE…	~	✓	✓	✓	✓	~	~	~	✗	✗	✓	✓	✗	~
TogetherComputer	LLM base: RedPajama-INCITE-7B-Base			RL base: various (GPT-JT recipe)										§
trlx	✓	✓	✓	~	✗	✓	✓	~	✗	✗	✗	✗	~	✓
carperai	LLM base: various (pythia, flan, OPT)			RL base: various										§
MPT-7B Instruct	✓	~	✓	~	✗	✓	✓	~	✗	✗	✓	✗	✓	✗
MosaicML	LLM base: MosaicML			RL base: dolly, anthropic										§
Stanford Alpaca	✓	✗	~	~	~	✗	~	✓	✗	✗	✗	✗	✗	✗
Stanford University CRFM	LLM base: LLaMA			RL base: Self-Instruct (synthetic)										§
Koala 13B	✓	~	~	~	✗	~	~	~	✗	✗	✗	✗	✗	✗
BAIR	LLM base: LLaMA 13B			RL base: HC3, ShareGPT, alpaca (synt…										§
LLaMA2 Chat	✗	✗	~	✗	~	✗	✗	~	~	✗	~	✗	✗	~
Facebook Research	LLM base: LLaMA2			RL base: Meta, StackExchange, Anthro…										§
ChatGPT	✗	✗	✗	✗	✗	✗	✗	✗	~	✗	~	✗	✗	✗
OpenAI	LLM base: GPT 3.5			RL base: Instruct-GPT										§

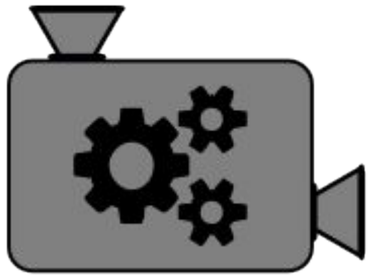
How to use this table. Every cell records a three-level openness judgement (✓ open, ~ partial or ✗ closed) with a direct link to the available evidence; on hover, the cell will display the notes we have on file for that judgement. At the end of a row, the § is a direct link to source data. The table is sorted by cumulative openness, where ✓ is 1, ~ is 0.5 and ✗ is 0 points.

De Mogelijkheden zijn Eindeloos!

AI voor innovatie & vooruitgang

- Object- of defectherkenning in foto's
- Structuur geven in ongestructureerde documenten:
  automatisch classificeren of doorzoeken van documenten, e-mails, …
- Datagedreven decision making
- Predictive maintenance
- Procesoptimalisatie voor efficiëntere routes, opslag of voorraadbeheer
- Genereren van potentieel nieuwe producten: medicijnen, onderdelen, ingrediënten, …
- …

Wees bewust van risico's op shortcut learning, biases, privacy en andere beperkingen



Powerful black box

Controllable white box

Inzichtelijk, aanpasbaar, open

Hoe werkt AI?
&
Hoe kan ik AI verantwoord inzetten?

Meike Nauta

FHI, september 2023

m.nauta@utwente.nl

linkedin.com/in/meikenauta/