Introduction to AI, fairness and bias

The pictures in these slides were generated using ChatGTP, AI Art and Canva

AI - Introduction

Artificial Intelligence around us

Artificial Intelligence refers to technological applications designed to simulate intelligence, such as the ability to learn, take decisions, and interact with surrounding environments

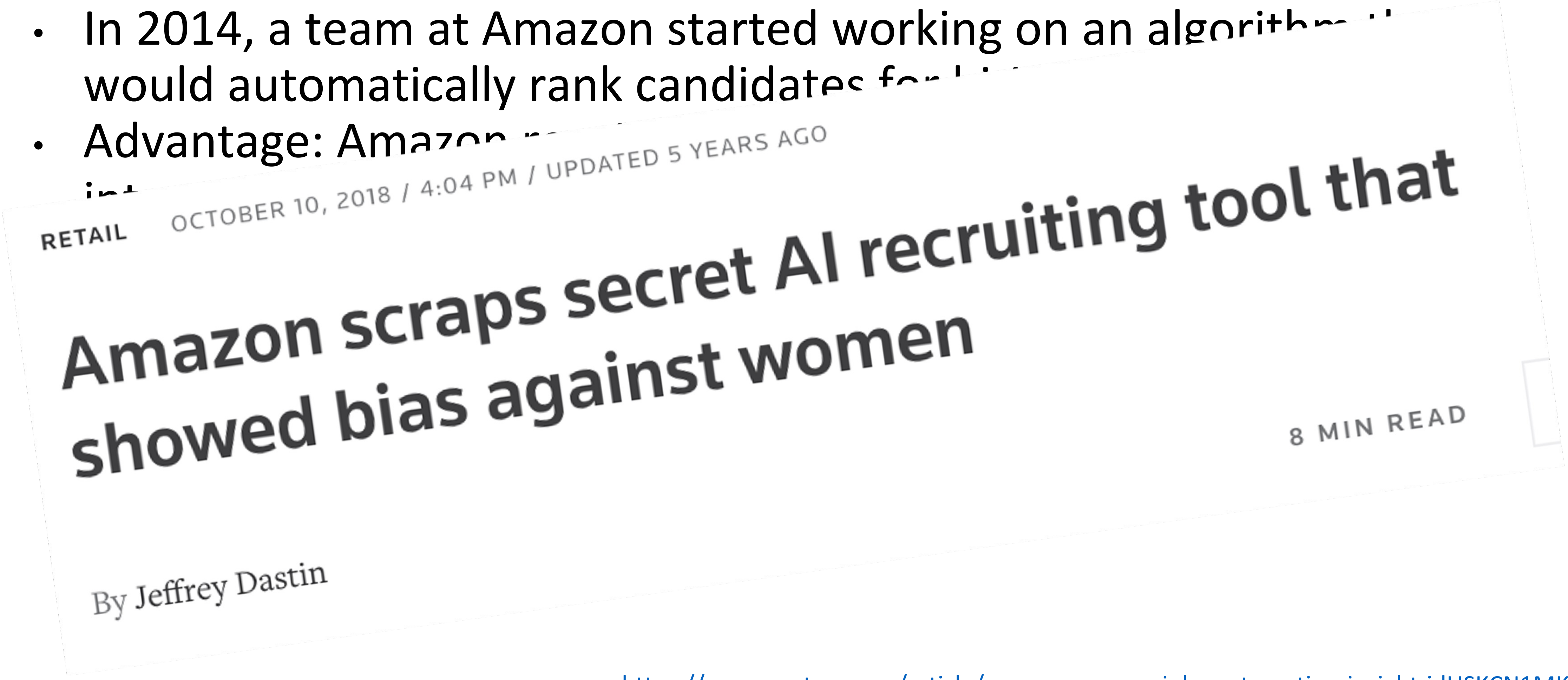
AI applications around us include:

- Recommendation Systems (YouTube, Instagram…)
- Virtual Assistants (Alexa, Siri…)
- Generative algorithms (ChatGPT, Dall-E…)
- Autonomous vehicles
- Playing agents (Chess, Go, various videogames…)

Did you know that the best chess player in the world is a computer program called Stockfish? Its rating (a measure of chess proficiency) is more than 3600. The best ever human player's rating is less than 2900.

A real scenario

- In 2014, a team at Amazon started working on an algorithm would automatically rank candidates for
- Advantage: Amazon int



RETAIL OCTOBER 10, 2018 / 4:04 PM / UPDATED 5 YEARS AGO

Amazon scraps secret AI recruiting tool that showed bias against women

8 MIN READ

By Jeffrey Dastin

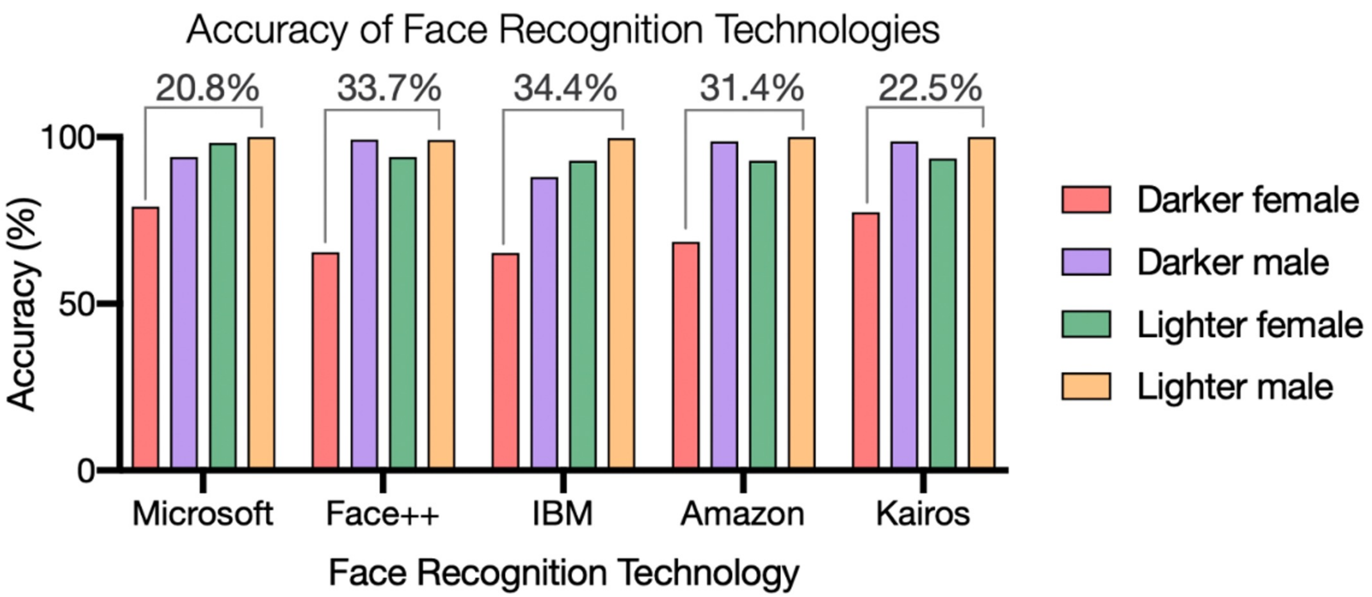
source: https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G/

AI under scrutiny

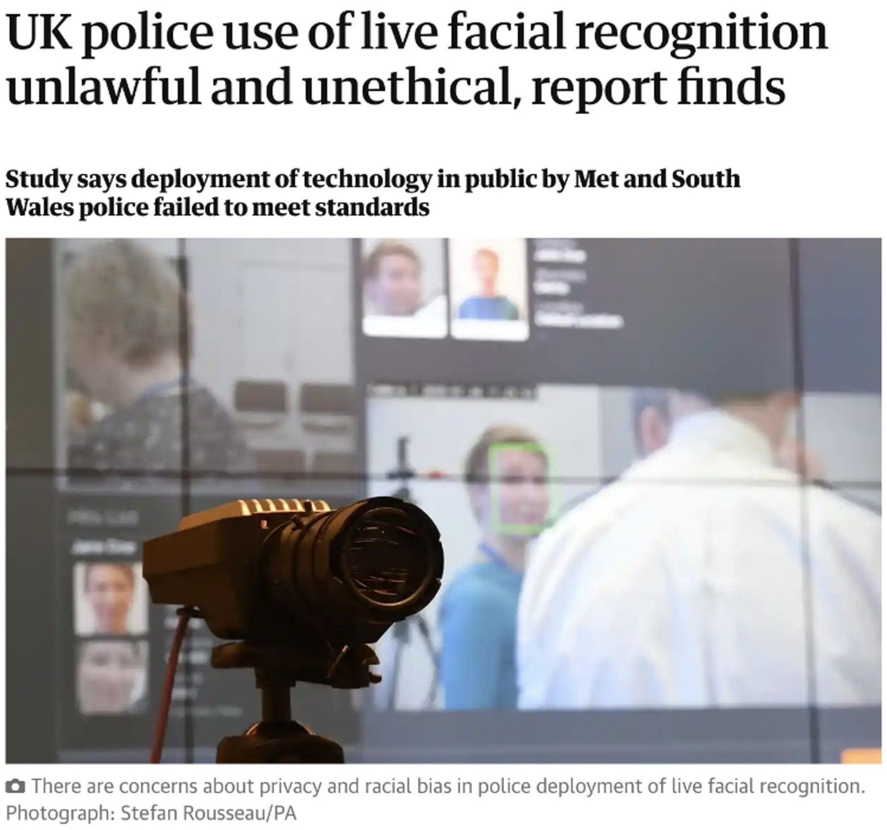
Today, there is a call for increased attention toward responsible use of AI applications, including a focus on aspects such as:

- **Fairness =** the idea that outcomes of an AI application must be equitable, i.e. no group should be discriminated against/receive a different treatment
- **Accountability =** clearly identify people responsible for the outcomes and derived consequences
- **Transparency =** the ability to explain outcomes and decisions, as well as transparency in data acquisition and provenance

More examples: fairness



source: https://sitn.hms.harvard.edu/flash/2020/racial-discrimination-in-face-recognition-technology/



UK police use of live facial recognition unlawful and unethical, report finds

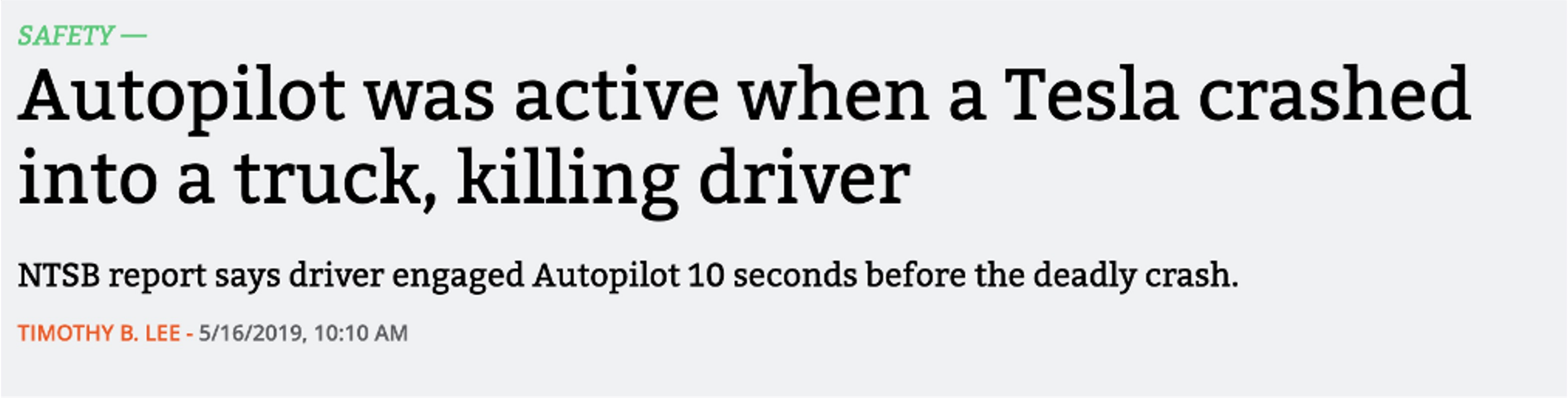
Study says deployment of technology in public by Met and South Wales police failed to meet standards

There are concerns about privacy and racial bias in police deployment of live facial recognition. Photograph: Stefan Rousseau/PA

Source: https://www.theguardian.com/technology/2022/oct/27/live-facial-recognition-police-study-uk

More examples: accountability

When technologies are involved, it is more difficult to pinpoint who is responsible when accidents happen.



SAFETY —

Autopilot was active when a Tesla crashed into a truck, killing driver

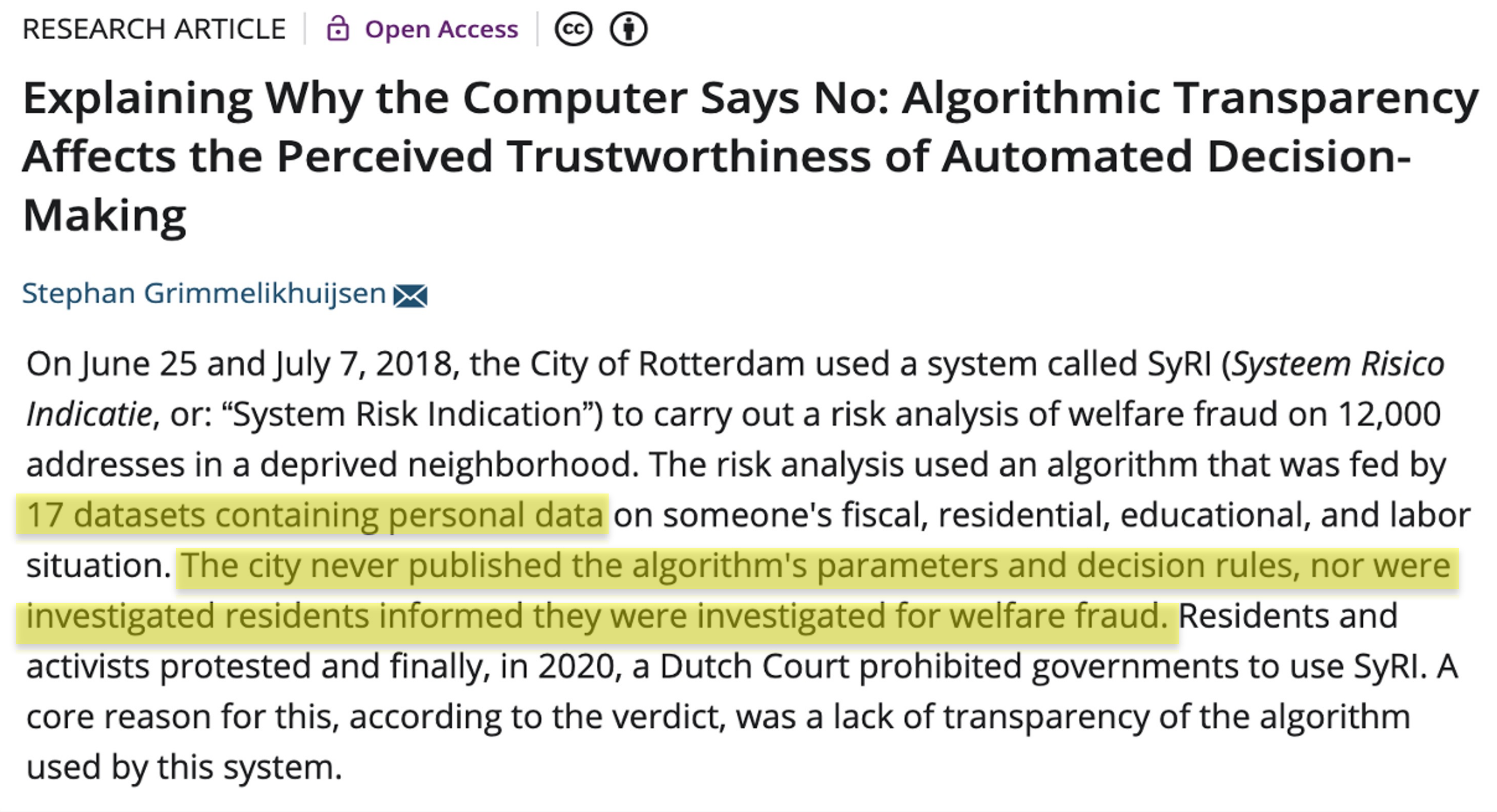
NTSB report says driver engaged Autopilot 10 seconds before the deadly crash.

TIMOTHY B. LEE - 5/16/2019, 10:10 AM

This is a legislative gap. Unfortunately, technology move at a much faster pace than the law, so these episodes are not infrequent.

source: https://arstechnica.com/cars/2019/05/feds-autopilot-was-active-during-deadly-march-tesla-crash/

The issue of transparency



RESEARCH ARTICLE | Open Access

Explaining Why the Computer Says No: Algorithmic Transparency Affects the Perceived Trustworthiness of Automated Decision-Making

Stephan Grimmelikhuijsen

On June 25 and July 7, 2018, the City of Rotterdam used a system called SyRI (*Systeem Risico Indicatie*, or: “System Risk Indication”) to carry out a risk analysis of welfare fraud on 12,000 addresses in a deprived neighborhood. The risk analysis used an algorithm that was fed by 17 datasets containing personal data on someone's fiscal, residential, educational, and labor situation. The city never published the algorithm's parameters and decision rules, nor were investigated residents informed they were investigated for welfare fraud. Residents and activists protested and finally, in 2020, a Dutch Court prohibited governments to use SyRI. A core reason for this, according to the verdict, was a lack of transparency of the algorithm used by this system.

source: https://onlinelibrary.wiley.com/doi/full/10.1111/puar.13483

AI and the environment

We have talked about how AI applications are becoming ubiquitous in society. Few people, however, including few of the developers of these systems, spend time thinking about their environmental impact. Even fewer try to evaluate this impact.

The environmental cost of training and maintaining AI algorithms is not immediately evident, but it includes:

- The massive amount of electricity required.
- Depending on how that electricity is produced, there is also a variable cost in terms of CO2 released in the atmosphere.
- The cost of building and maintaining the hardware on which AI algorithms run.

Recent estimates of AI's carbon footprint range from 2.1% to 3.9% of the total greenhouse gas emissions. Still a small fraction of what is produced by more polluting industries such as manufacturing (24%) and transportation (27%), but not insignificant.

Source: The Carbon Footprint of Artificial Intelligence

Accuracy-Efficiency Paradox

"There is a recognised trade-off between model accuracy and energy efficiency. In fact, the relationship has been shown to be logarithmic. That is, ***in order to achieve a linear improvement in accuracy, an exponentially larger model is required.*** *A recent study [...] confirmed the existence of the accuracy-energy trade off [...] and indicated a 30-50% saving in energy for training related to a 1% reduction in accuracy."*

Mill *et al.*, Managing Sustainability Tensions in Artificial Intelligence: Insights from Paradox Theory

Strategies for Greener AI

Optimizing AI algorithms to improve energy efficiency and reduce computational demands.

Transitioning to renewable energy sources for data centers and AI infrastructure.

Exploring more sustainable hardware solutions designed for energy efficiency and reduced environmental impact.

Data Bias

Bias in data

AI algorithms learn through the examples they are provided with, which together form what we call the **training set**.

Bias in the training set will affect the algorithm’s behavior, likely amplifying existing problems and mistakes.

Let’s explore a few ways in which a training set can be biased.

Representation bias

Representation bias happens when the training set is, essentially, incomplete, and a poor representation of the population we wish to apply the algorithm on.

Other times, representation bias happens when trying to use existing samples from a different time or place, instead of collecting new ones.

Dr. Joy Buolamwini demonstrating failures in face recognition.

Measurement bias

Measurement bias has to do with the way we try to create a numerical representation of the problem that we are trying to solve. It may arise in a couple of different ways, such as:

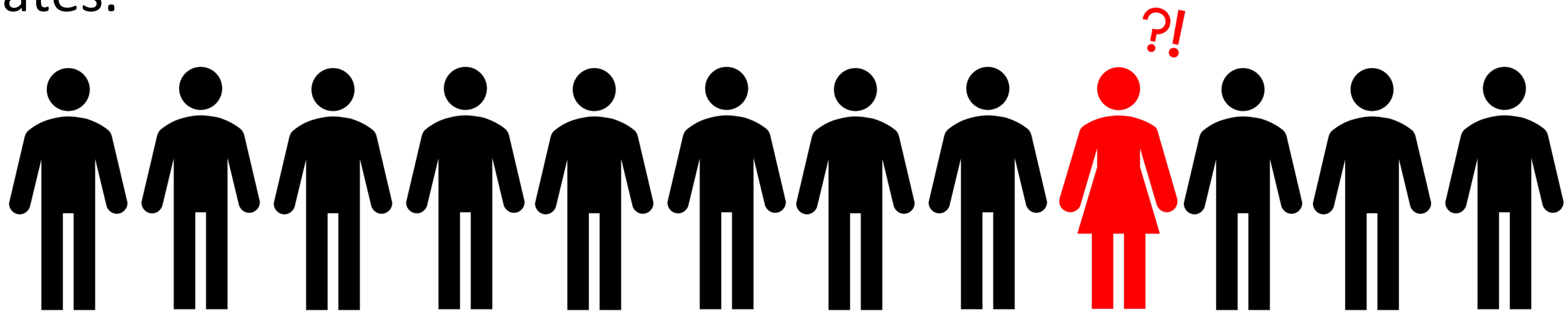
The features or labels used are an oversimplification of the problem

Different groups are measured in different ways

Historical bias

Historical bias refers to a misalignment between what we wish to model and the actual state of the world.

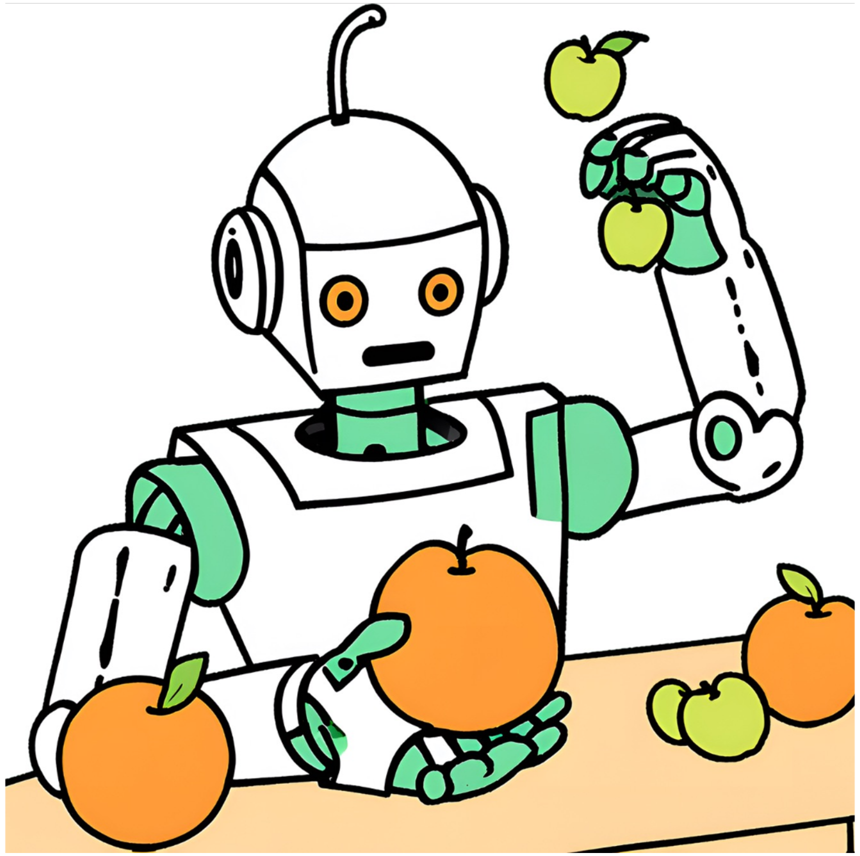
The Amazon AI recruiting tool discussed earlier was affected by historical bias. The algorithm was trained to select people similar to the existent employees, and because men employees were more numerous than women, the algorithm learned that, all things being equal, women were less preferable candidates.



Note that historical bias is particularly insidious, because it is not a sampling or measurement problem, and can not be corrected in the training set – we can only work around it.

Learning Algorithms - Fundamentals

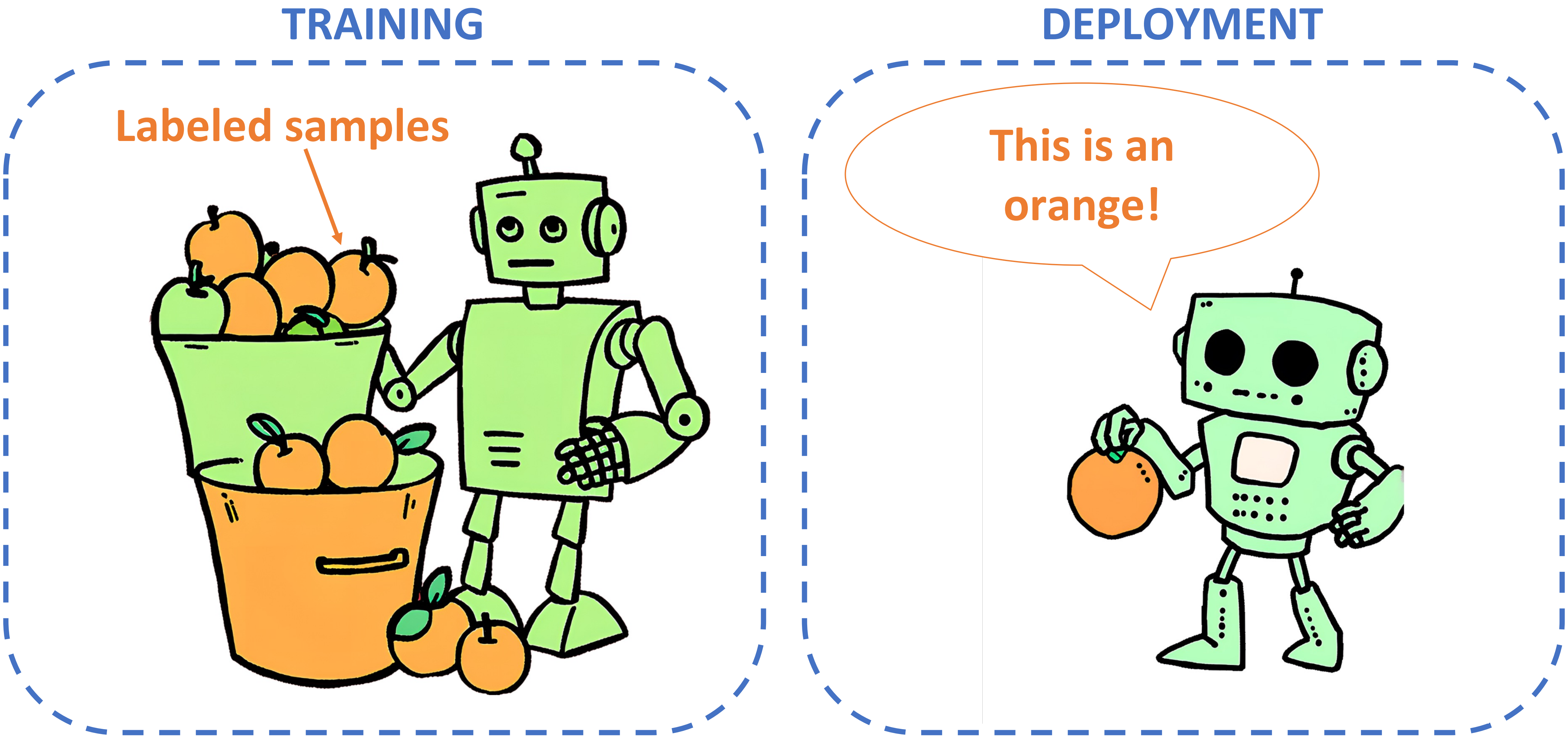
A shift in paradigm



AI is becoming increasingly powerful, thanks to a fundamental shift in the way it is programmed. Imagine wanting to program a machine to recognize and separate apples and oranges:

- Traditionally, a programmer would need to write code including rules to follow to separate apples from oranges based on their characteristics (e.g. color, skin texture, shape…). It was typically hard to come up with a good set of rules, especially for complex problems!
- With recent AI approaches (e.g. neural networks), we see many programs written to learn rather than to follow existing instructions. This way, all one has to do is showing the machine enough apples and oranges, and it will learn how to recognize new samples!

How does a program learn?



TRAINING

Labeled samples

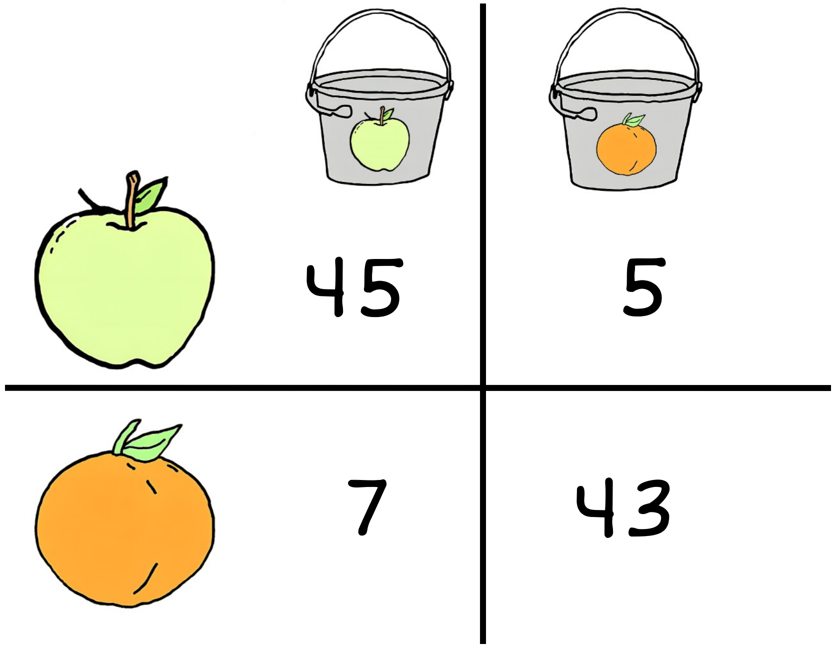
DEPLOYMENT

This is an orange!

Measuring performance

A computer needs a quantitative metric to measure how well they are performing their task. The most common performance metric is called **accuracy**.

Accuracy simply counts how many samples were classified correctly.



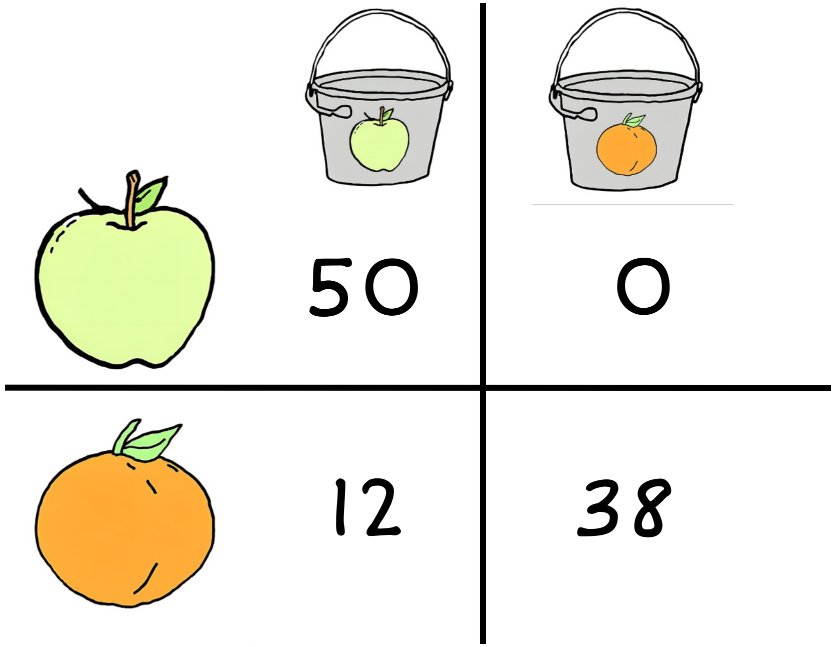
	45	5
	7	43

The chart to the left (called a **confusion matrix**) shows how the machine classified 50 apples and 50 oranges.

Accuracy = 45 + 43 /100 = 0.88, or 88%

Your turn!

Try to evaluate the accuracy of this machine, based on this new confusion matrix:



	🪣🍏	🪣🍊
🍏	50	0
🍊	12	38

When accuracy is not enough

In both examples the resulting accuracy was 88%.

The first case was more balanced, while in the second example more fruits were misclassified as apples.

If the error’s cost is symmetric (it is just as bad to classify an apple as an orange than vice-versa), then both machines are equivalent. But in some cases this is not true.

Can you think of real applications in which making an error in a direction is worse than making it in the other?

Examples of not symmetric error cost

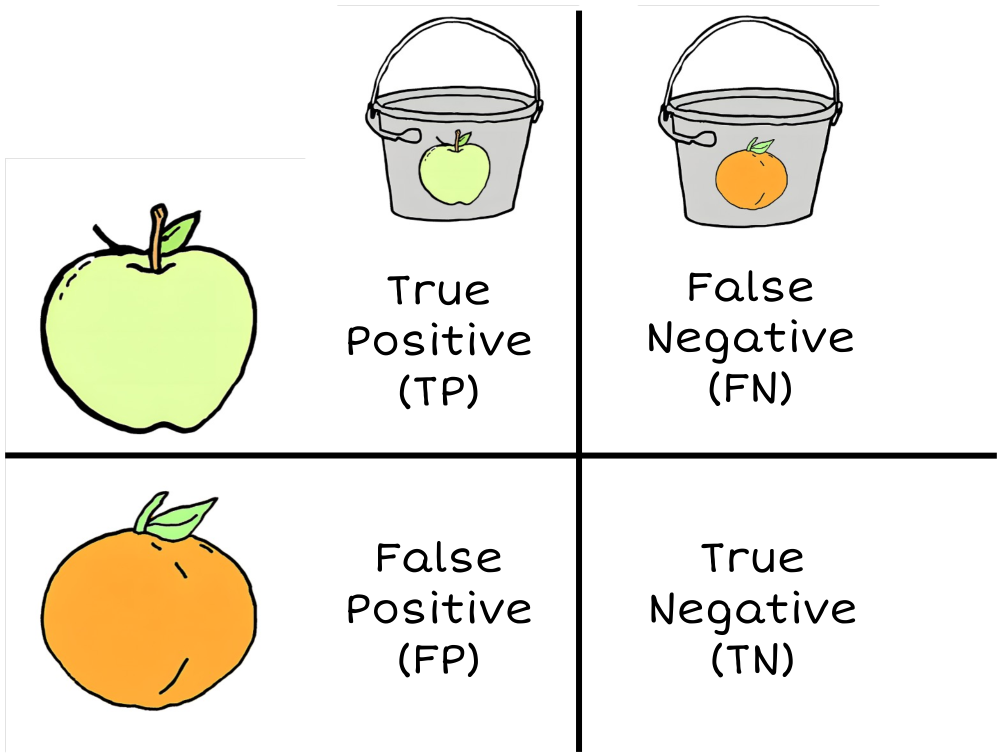
When it’s better to be more careful

In most medical applications it is better to classify a few more patients as having the disease, then follow up.

When it’s better to let it slide

In the case of a spam filter (yes, that’s AI too), it may be better to classify emails as legitimate, rather than losing possibly important, legitimate emails.

Confusion matrix – a closer look



True Positive (TP)

False Negative (FN)

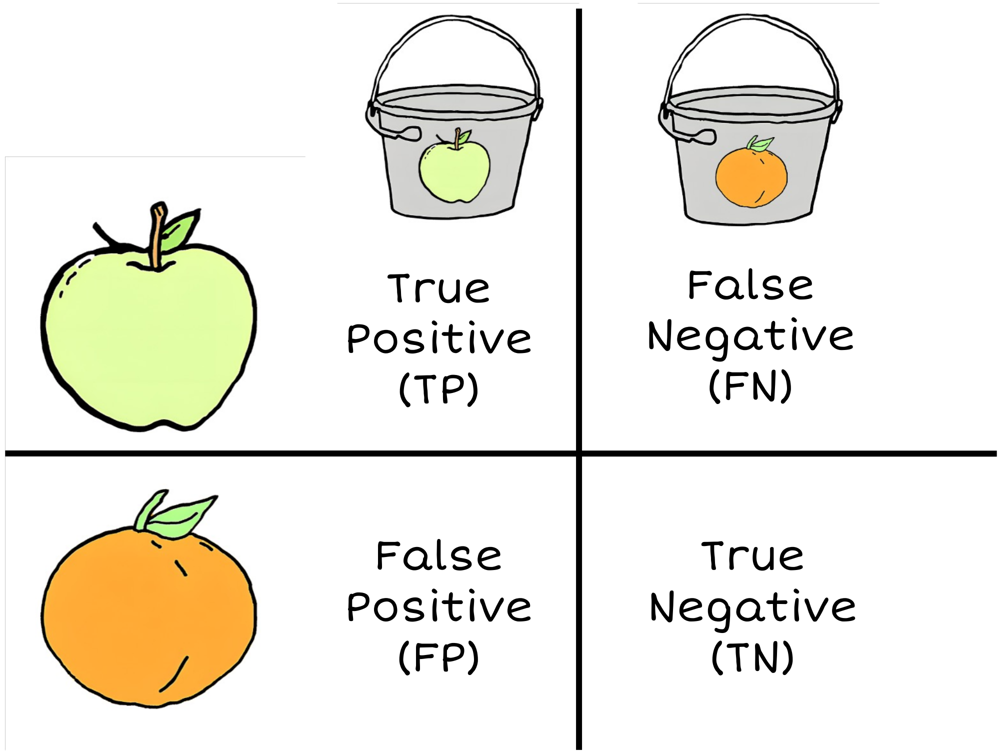
False Positive (FP)

True Negative (TN)

Let's take a closer look to the confusion matrix – this will help us define other performance metrics.

Typically, in a classification problem, we define a class of interest (**positive class**) that we want to separate from the other (**negative class**). Based on our choice, we can label the cells of the confusion matrix as shown in the picture

Other performance metrics



True Positive (TP)	False Negative (FN)
False Positive (FP)	True Negative (TN)

Recall: also called *sensitivity*, it measures how good the algorithm is at finding samples of the positive class. It is expressed as

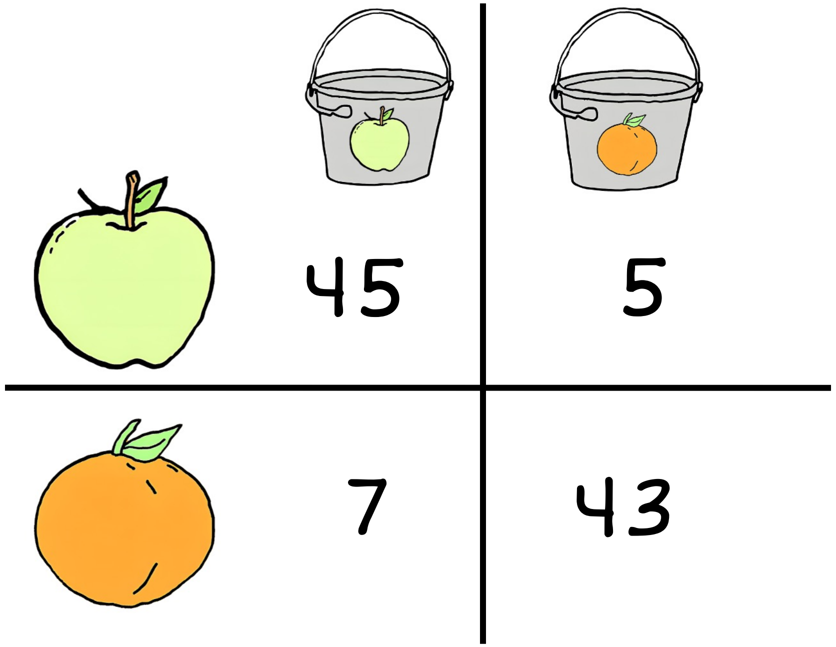
$$\frac{TP}{TP + FN}$$

Precision: also called *specificity*, it measures the ability of the algorithm to incorrect assign samples to the positive class (in other words, how good it is at avoiding false positives. Its formula is

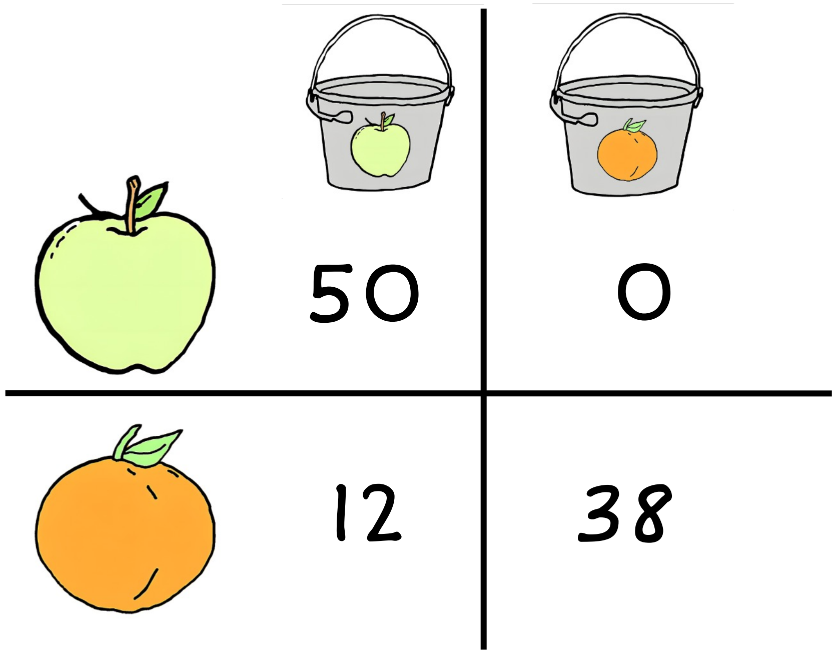
$$\frac{TP}{TP + FP}$$

Exercise

We already know the accuracy of these two algorithms is 0.88. Now, compute their recall and precision.



	🪣🍏	🪣🍊
🍏	45	5
🍊	7	43



	🪣🍏	🪣🍊
🍏	50	0
🍊	12	38

Choosing the right metric

Based on the application, we may want our algorithm to optimize for a metric rather than others.

Understanding questions:

1. In a medical application, where it is really important to not miss patients with a disease, what metric should we optimize: accuracy, recall or precision?
2. What about in the spam filter example?
3. Can you think of an easy way for an algorithm to have perfect recall? Do you think this algorithm would be actually useful in the real world?

Recap

In this section, we have learned:

- Sample applications of AI
- Difference between learning algorithms and traditional programming
- Performance metrics for simple classification tasks: accuracy, recall and precision
- How to read a confusion matrix

Fairness and AI

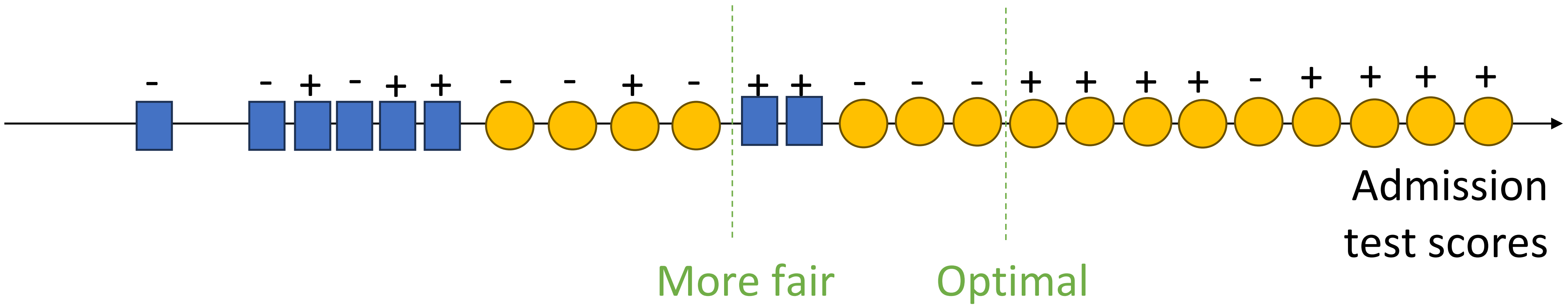
Unfairness in learning algorithms

Scenario: The Data Science University has decided to use an admission test to select which of the students who apply should be admitted. The top N students get in, the others are rejected.

In this world, two races exists: Circles and Squares

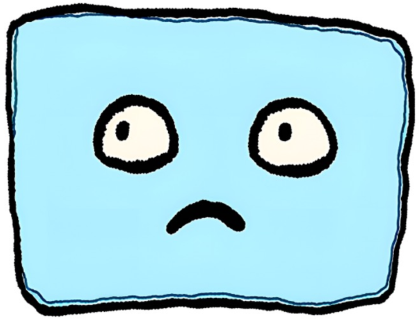
Circles tend to be wealthier; this means their families can pay tutors and preparation centers to help them ace the admission test.

Other than this difference in preparation, the two populations are equally likely to succeed.



Example based on “The ethical algorithm”, by Kearns and Roth

Social impact of AI



The previous example shows how AI applications can end up being **biased**, and treat different populations differently.

Note that this often happens involuntarily: the algorithm was not trained to be unfair to a group, it has simply learned based on the examples it was shown.

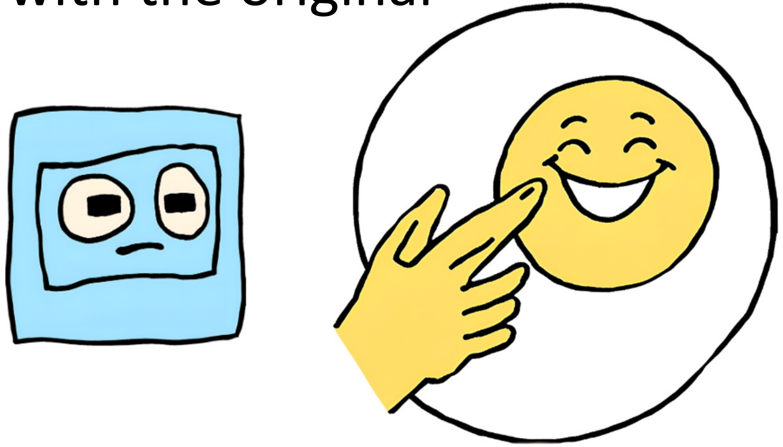
In complex problems, it is difficult to achieve a good representation of the information we are trying to classify, and we are likely dealing with a lot of noise. Coming up with a good algorithm is challenging, and the consequences for real people can be severe.

Other kinds of bias - algorithmic bias

A biased dataset will most likely produce a biased algorithm.

We talk about **algorithmic bias**[*] when an algorithm systematically and unfairly favors a group over others, in a way that has nothing to do with the original intention of the algorithm.

In our example about university admissions, the goal was to admit the best students, but the algorithm is biased against Squares for reasons other than their academic abilities.



[*]Note that, in Machine Learning, we also talk about "high bias" in an algorithm when the algorithm is too simple and inflexible to capture the trend in the data.

Other kinds of bias - evaluation bias

After training, a classification algorithm must be tested on a different set of data, to measure its performance on samples it has not seen before.

If the **testing set** is biased in a similar way as the training set, we will remain unaware of possible issue with fairness. We may also report incorrect, overly-confident measure of performance, and deploy an algorithm that will end up failing when used in the real world.

To avoid evaluation bias, it is important that the algorithm is tested on a dataset that well represents the population it will be used by (or on).

Why does this matter?

Learning algorithms are becoming increasingly popular and their application widespread, because they are:

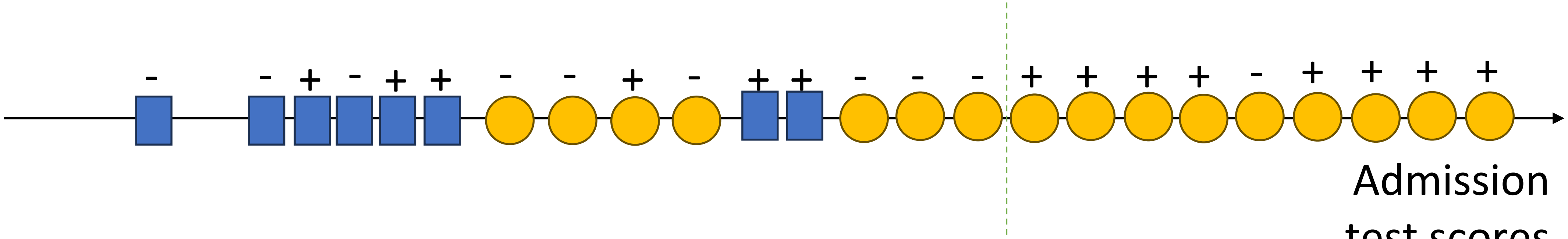
- Cheap
- Scalable
- Automated

Unfortunately, they are also:

- Seemingly objective - for years, the idea of bias in algorithms was rejected, because it was thought that, because machines do not have emotions, they could not be sexist, racist, etc…
- Often lacking appeals processes, because the responsibility of a wrong prediction is difficult to assign (is it the programmer’s fault? Or the client’s?)
- Not just predicting but also causing the future – decision algorithms can significantly influence the world on which future algorithms will be based, creating a vicious cycle.

Measuring fairness

As seen before, high accuracy does not guarantee fairness



Admission test scores

Highest accuracy

Other metrics can be used to measure fairness of an algorithm, for example:

- **statistical parity:** whether the rates of positive predictions are on par across groups
- **equal opportunity:** whether the ratios of false positives to predicted positives are on par across groups
- **predictive equality:** whether the ratios of false negatives to predicted negatives are on par across groups

Data Ownership in the Age of AI

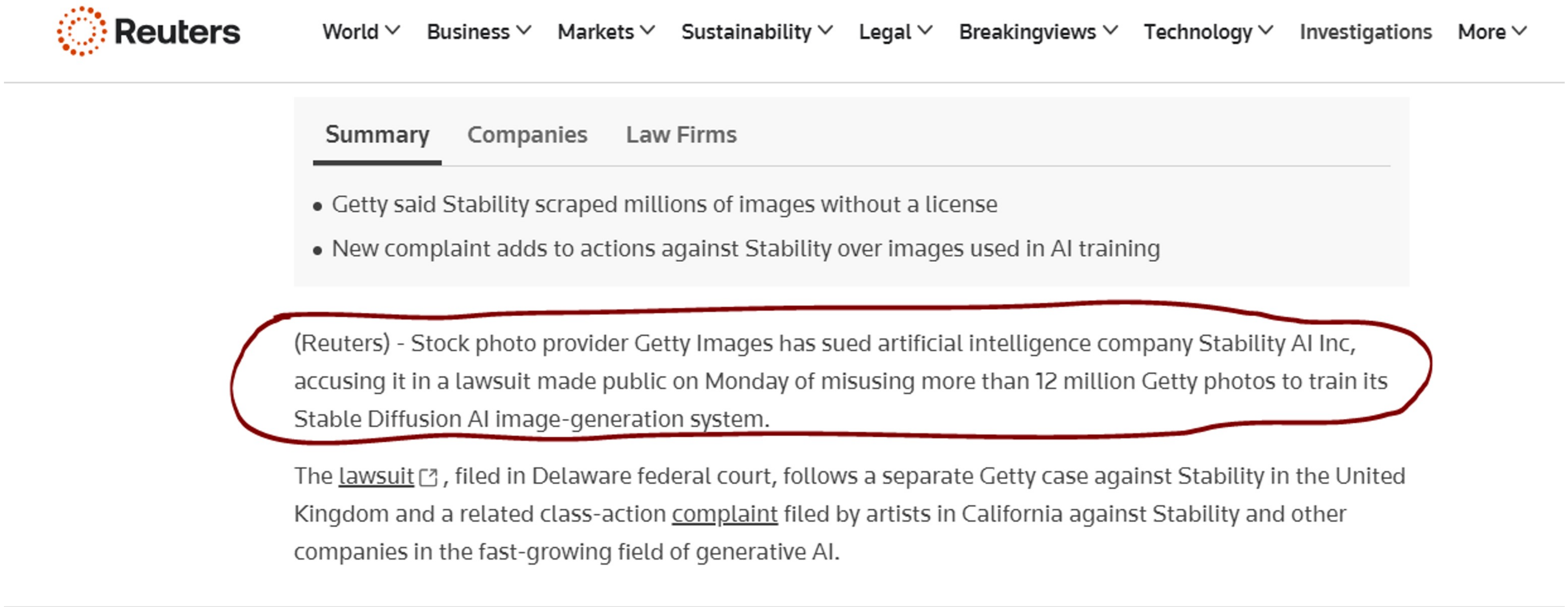
Data ownership in AI age: Why does it matter?

- We live in an age where data fuels innovation; data is the backbone of AI and machine learning algorithms
- Understanding data ownership goes beyond legal matters; decisions made from data and AI based algorithms are expected to have the broader societal and ethical implications.
- Companies with ownership of high-quality data are better positioned to unlock domain specific insights increasing their chances for innovations, monetizing the models and expanding competitive advantage.

What is data ownership?

- Data ownership determines who has the legal rights to control, access, and use data.
- As such, data ownership deals with the issues of intellectual property, and the roles of various stakeholders, including individuals, organizations, and governments and regulatory bodies.

Scenario



- Getty said Stability scraped millions of images without a license
- New complaint adds to actions against Stability over images used in AI training

(Reuters) - Stock photo provider Getty Images has sued artificial intelligence company Stability AI Inc, accusing it in a lawsuit made public on Monday of misusing more than 12 million Getty photos to train its Stable Diffusion AI image-generation system.

The lawsuit, filed in Delaware federal court, follows a separate Getty case against Stability in the United Kingdom and a related class-action complaint filed by artists in California against Stability and other companies in the fast-growing field of generative AI.

Question: How can AI companies develop models that respect data ownership while still advancing innovation?

https://www.reuters.com/legal/getty-images-lawsuit-says-stability-ai-misused-photos-train-ai-2023-02-06/

The Challenges of Data Ownership

Balancing Innovation and Responsible AI: How do we create the balance between the use of AI to innovate with responsible and ethical considerations.

The Dilemma of Ownership in AI Outputs: Who owns the data used to train AI models, and who owns the outputs generated by those models?

Data flow between jurisdictions: How should the integrity and data ownership be maintained?

Scenario

- **Situation**: Google's Maps and other location-based services collect extensive data on users' movements and locations to provide real-time traffic updates and personalized recommendations.
- **Issue**: Privacy advocates raise concerns about continuous location tracking and the potential misuse of location data. Users demand more control over their location data and clarity on how it is used.
- **Question:** How does Google ensure compliance with international data protection regulations, such as GDPR and CCPA, with respect to location data?

Data Monopolies: A Growing Concern

- Data monopolies occurs when a single entity e.g. company, government etc. control data ownership overshadowing the small players.
- The company that has the data has an upper hand on how the data is used.
- **What is the risk introduced by monopolies:** The monopoly entities can create unfair advantages including unfair competition, intentionally limiting innovation and hindering inclusive AI models and decisions.

Monopolies and Legal Frameworks: GDPR and the AI Act

Recognizing the threats that are posed by this monopolies, the governments, institutions, organizations have developed the regulations to reduce the effects of monopolies.

For examples

- General Data Protection Regulation (GDPR): Empower individuals to have control on how their data is collected and processed thus creating spaces for higher accountability and transparency.
- EU AI Act: Focuses on regulations for safety critical systems i.e. AI enabled video games to ensure that they are deployed responsibly and poses minimum risks to the users.

Strategies for Responsible Data Governance

On top of legal frameworks robust data governance is essential for creating Responsible AI.

Some of these strategies companies can adopt include:

- Minimized data collection: Collect data only when is necessary
- Algorithmic transparency: Ensuring that they AI models and algorithms are able to provide data on how they are making the decisions
- Human in the loop: Embracing human insights and feedback on the performance of the models

Questions for reflection

1. The data ownership process can be made more inclusive and equitable by accounting for the diverse perspectives of the involved parties, including individuals, organizations, government entities, and legal stakeholders. How can the involved parties design a collaborative approach that ensures the needs of all relevant stakeholders are addressed?

1. As the use of ChatGPT grows in popularity, the issue of data ownership and control is of particular relevance. Users may have concerns about the confidentiality and security of the data they provide to the tool, for example in tasks such as paraphrasing and formatting. How can users be given a stronger sense of agency over their data?

Key takeaways

Data ownership is complex: There are legal, intellectual property, and stakeholder considerations that make defining data ownership challenging.

The Need for Transparency and User Control: Users need to be informed about how their data is being collected, used and given control over their data.

Navigating Legal and Cultural Differences: AI companies operating globally must contend with varying regulations, cultural variances and rapidly growing innovations