

# The multifaceted nature of bias in AI: Implications for generalization, fairness, and robustness

Eirini Ntoutsi

# Bias: A multifaceted concept

- **Terminology drift**
  - Original meaning ([1828](#))
    - “a leaning of the mind”, “to lean or incline from a state of indifference, to a particular object or course”
  - Modern definition ([2024](#))
    - “an inclination of [temperament](#) or outlook - especially a personal and sometimes unreasoned judgment : [PREJUDICE](#)”
- **Overloaded term:** now refers to misuse of data, prejudiced behavior, and favoritism.
- **Bias is not inherently bad**, its meaning depends on context
  - In contemporary AI literature, bias is often used as a synonym for discrimination and unfairness.

## Bias



BI'AS, *noun*

1. A weight on the side of a bowl which turns it from a straight line.
2. A leaning of the mind; inclination; prepossession; propensity towards an object, not leaving the mind indifferent; as, education gives a *bias* to the mind.
3. That which causes the mind to lean or incline from a state of indifference, to a particular object or course.

## bias 1 of 4 noun

bi·as (bī-əs)

[Synonyms of bias >](#)

- 1 **a** : an inclination of [temperament](#) or outlook  
*especially* : a personal and sometimes unreasoned judgment : [PREJUDICE](#)
- b** : an instance of such prejudice



# The many facets of bias in AI

The useful, the problematic, and the harmful

- Bias is a **neutral concept** - it can help, hinder or harm
- Bias can be
  - **Useful:** guides or steers the machine towards desired or efficient solutions
    - Analogy to humans:
      - Preferring healthy food based on past outcomes
      - Starting work early if you are a morning person
  - **Problematic:** reducing technical robustness
    - Analogy to humans:
      - Not taking an umbrella because yesterday was sunny (recency bias)
      - Being overconfident based on past success (overconfidence bias)
  - **Harmful:** causing discrimination and unfairness
    - Analogy to humans:
      - Judging someone's competence based on gender or race
      - Expecting better academic performance from students with higher socioeconomic status

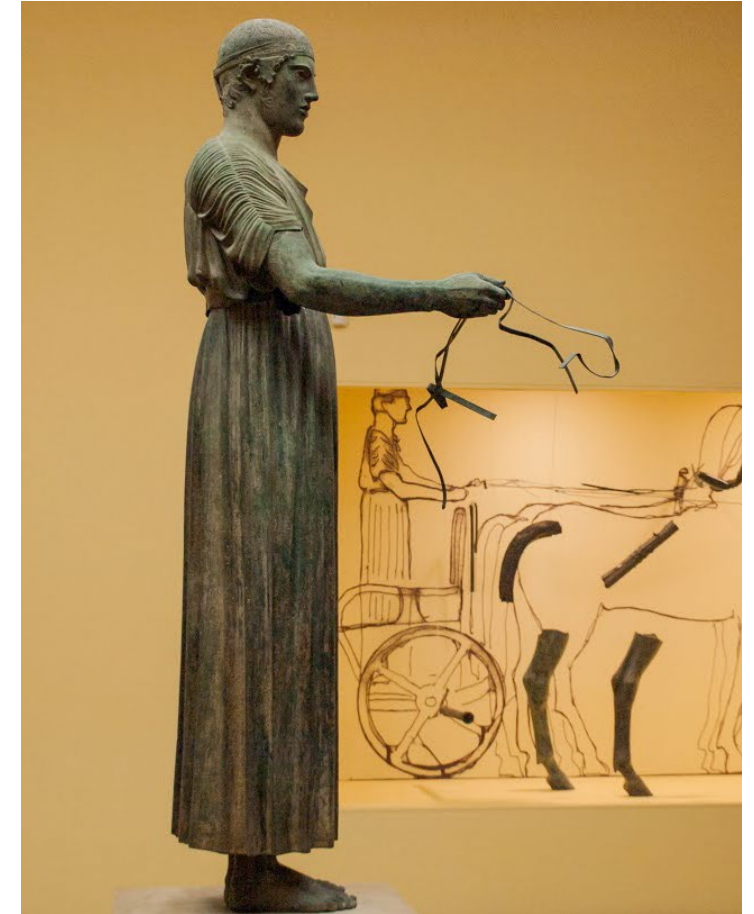
# Useful biases: steering & control

Guiding the machine towards desired or efficient solutions

## In modern AI (Machine Learning)

- **Inductive or learning bias**: Explicit or implicit *assumptions* made by a learning algorithm to perform induction ([Hüllermeier et al, 2013](#)).
- “**Bias-free learning is futile**”: A learner that makes no a priori assumptions regarding the identity of the target concept has no rational basis for classifying any unseen instances.
- Inductive bias is learner-specific, e.g.:
  - Decision trees: axis-aligned splits
  - k-nearest neighbors: locality assumption
  - Naïve Bayes: Class-conditional independence
  - Neural networks: Compositional structure

Here, bias is necessary for learning, as it enables generalization from a finite set of instances.



The Charioteer of Delphi, aka Heniokhos (Greek: Ἡνίοχος, the rein-holder)

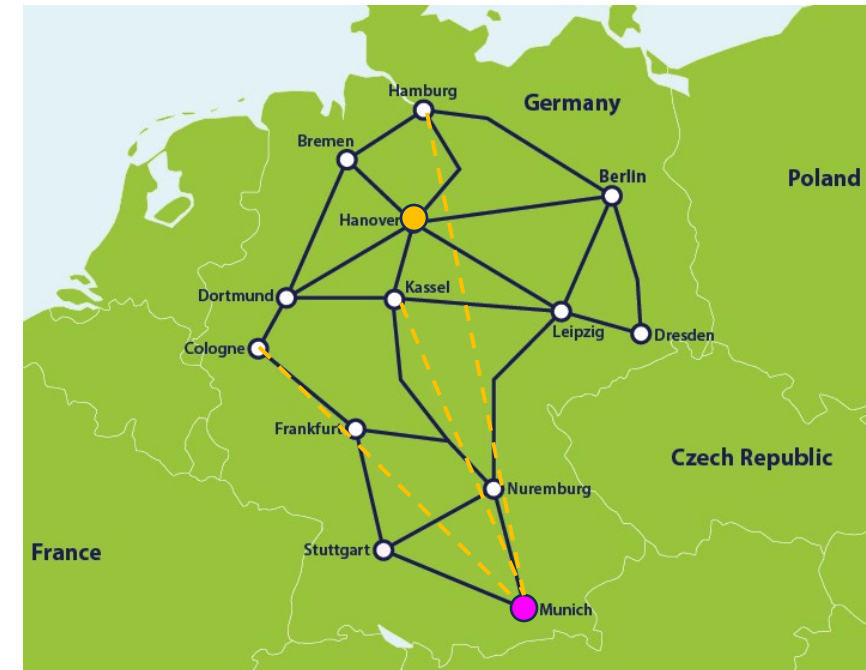
# Useful biases: steering & control

Guiding the machine towards desired or efficient solutions

## In traditional AI (Search algorithms)

- In **AI search**, heuristics guide the exploration of the state space by estimating how close a state is to the goal, avoiding exhaustive search and making problem-solving more efficient.
  - A common heuristic is straight-line distance (SLD), a proxy for actual travel cost.
- **Heuristic bias** arises from the assumptions built into the heuristic function
  - These assumptions **bias** the search toward promising states, reducing the search space and improving efficiency.

Here, bias improves efficiency



Main train connections in Germany ([source](#))



# Problematic biases:

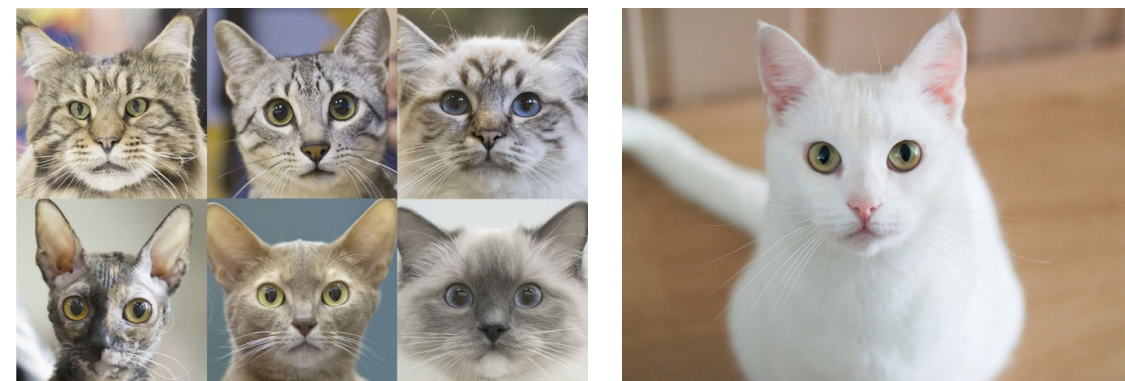
Reducing technical robustness

Biases that affect model generalization and **limit performance** in new or changing contexts

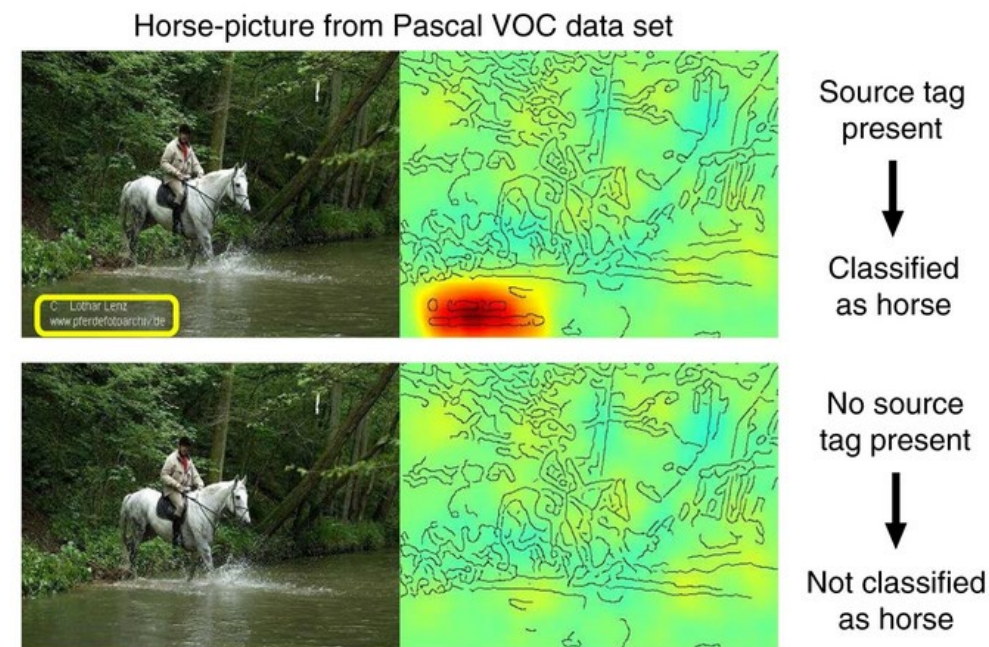
Common sources of problematic biases

- **Unrepresentative** training data → the model learns from an incomplete or skewed sample
- **Spurious correlations in the data** → wrong associations learned (e.g., source tag present ~ class)
- **Distribution shifts** → model fails when context changes (e.g., over time, or across domains)
- ...

Here bias doesn't cause unfairness but undermines technical robustness and must be addressed to ensure AI systems remain reliable across context.



Overfitting on gray cats, failing to generalize to white cats



Clever Hans effect

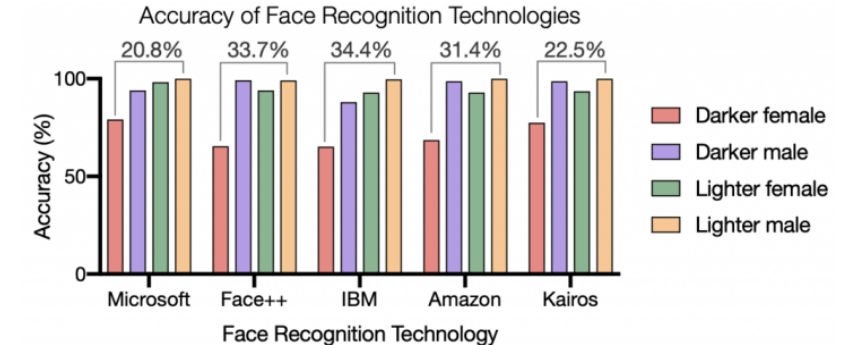
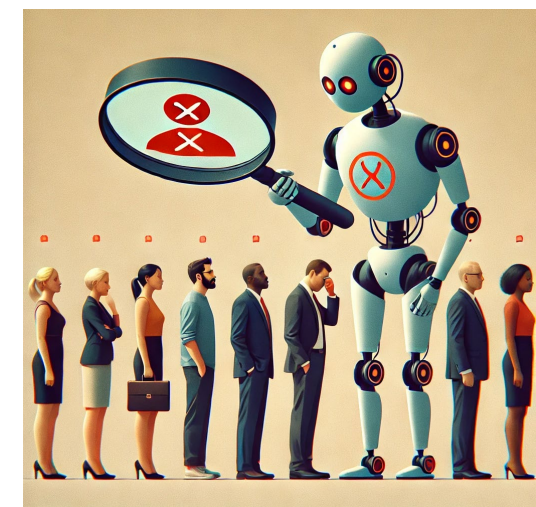
# Harmful biases:

Causing discrimination and unfairness

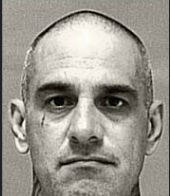



AI systems can **reflect**, **amplify**, or even **create new** social inequalities.

- Examples
  - Hiring algorithms favoring male resumes
  - Facial recognition software recognizing White males better than Black females
  - Criminal risk assessment reinforcing racial stereotypes
- Automated decisions influenced by **protected attributes** (e.g., gender, race, age, sexual orientation)
  - Indirect bias via **proxy attributes**, e.g.:
    - Zip code as a proxy for race
    - Name or hobbies as proxies for gender

Here bias can lead to discrimination and unfairness and cause harm to individuals or groups.



Auditing face recognition technologies. The [Gender Shades](#)

Two Petty Theft Arrests		Two Petty Theft Arrests	
			
VERNON PRATER	BRISHA BORDEN	VERNON PRATER	BRISHA BORDEN
Prior Offenses 2 armed robberies, 1 attempted armed robbery	Prior Offenses 4 juvenile misdemeanors	Prior Offenses 2 armed robberies, 1 attempted armed robbery	Prior Offenses 4 juvenile misdemeanors
Subsequent Offenses 1 grand theft	Subsequent Offenses None	Subsequent Offenses 1 grand theft	Subsequent Offenses None
LOW RISK 3	HIGH RISK 8	LOW RISK 3	HIGH RISK 8
Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.		Borden was rated high risk for future crime after she and a friend took a kid's bike and scooter that were sitting outside. She did not reoffend.	

The multifaceted nature of bias in AI: Implications for generalization, fairness, and robustness

# Disentangling the facets of bias

Clarity in terminology is essential

- Conflating bias with unfairness, risks overlooking that bias has many facets
- **Not all bias is unfairness**
  - Some biases are **useful** for learning and generalization
  - Others are related to technical **robustness**.
- These facets are often **interdependent**
  - Different learners yield different robustness-fairness trade-offs
  - Limited robustness can lead to discriminatory outcomes
  - **Robustness and fairness are linked**, but improving technical robustness does not always improve fairness.
- A balanced understanding of bias is necessary to build responsible AI:
  - Leverage useful biases
  - Mitigate problematic ones
  - Eliminate harmful ones

**Know thy biases**



# Thank you for your attention! Questions?



Funded by  
the European Union



Contact me:

- [eirini.ntoutsi@unibw.de](mailto:eirini.ntoutsi@unibw.de)
- <https://www.unibw.de/aiml>
- <https://aiml-research.github.io/>