What is the meaning of "bias" in AI?

Ambre DAVAT

Post-doctoral researcher at the GRESEC
Member of the Ethics&AI chair of the MIAI Institut and Institut of Philosophy of Grenoble
ambre.davat@univ-grenoble-alpes.fr

November 23, 2023

Introduction

Initial observations :

- **The word “bias” is frequently used in the field of AI, but to describe very different situations .**
  Ex : wrong image classification, variable performances in Speech-to-Text technologies, stereotyped content generation, algorithmic discrimination in risk assessment
- **The literature focuses on the various sources of bias, but global definitions of “bias” are few and inconsistent.**

(Hovy, Prabhumoye, 2021) : Bias = "Differences between (a) a “true” or intended distribution (e.g., over users, labels, or outcomes), and (b) the distribution used or produced by the model."

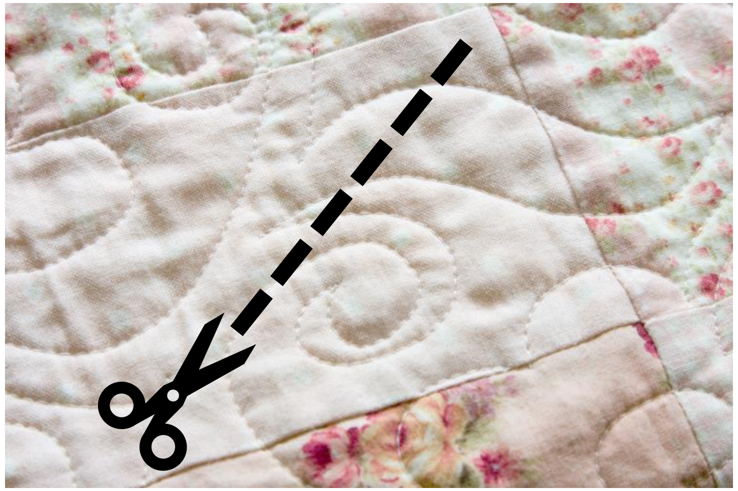
(Loubes, 2022) : Bias = "An unfair/irrelevant information that influences a decision"

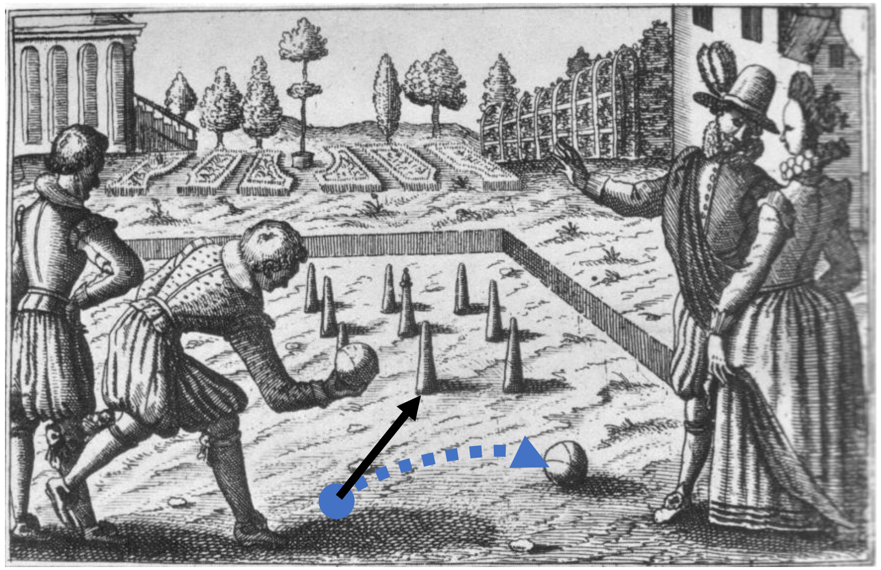
(Mehrabi et al., 2022) : Bias ⇔ "Source of unfairness“

Bias as deviation from a norm

sources: TLFi & Online Etymology Dictionary

- France, 1250: « de biais » = a **sewing** term when the cutting of fabric is not straight





English Life in Tudor Times, Roger Hart
NT: Putnam, 1972

- England, 1560: « bias » = a **bowling** term, used to describe unbalanced balls, which tend to deviate from the intended direction

Figurative use :
"a one-sided tendency of the mind"

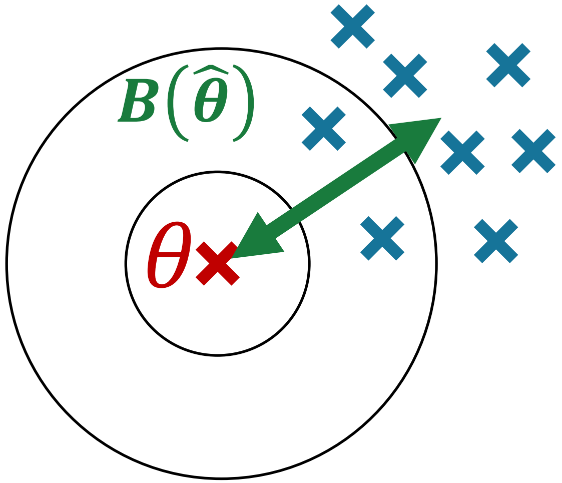
Bias as deviation from a norm

- **In statistics**: bias of an estimator

$$Biais(\hat{\theta}) \overset{\text{def}}{=} \mathbb{E}(\hat{\theta}) - \theta$$

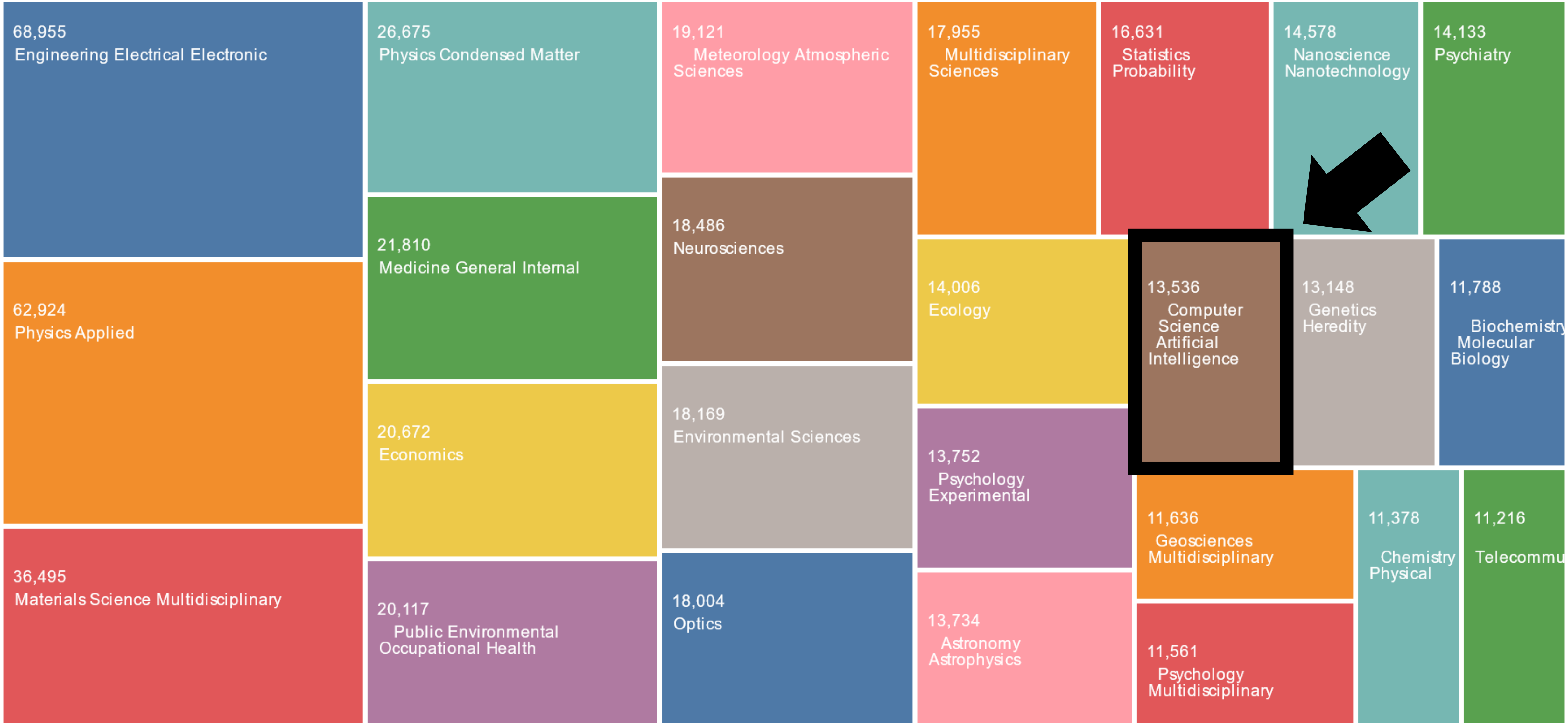
$\mathbb{E}(\hat{\theta})$ **: Expected value of the estimator**
(mean of the **estimations**)

θ **: True value of the parameter being estimated**



Bias in scientific literature

25 fields with higher occurrence of the word « bias » (title, summary or key words) in Web of Science



Bias in scientific literature

Thematic analysis of the first 50 articles from Web of Science => 3 norms

<u>Specific object / field of study</u>: voltage bias, algorithmic bias, bias in peer-review, education, history, politics, clinical research...

<u>Statistical biases</u>: small sample bias, confounding bias, selection bias, bias elicitation, bias awareness, bias testing, bias correction...

➢ Bias as a divergence from **scientific standards** (*methodological bias*)

Difficulty bias, attention bias, optimism bias, memory bias, hinsight bias, ideological bias, confirmation bias...

➢ Bias as a divergence from **the "rational" decision** (*cognitive bias*)

*(**implicit**) racial bias, gender bias, social class bias...*

➢Bias as a divergence from **the ideal society** (*socio-historical bias*)

What about bias in AI?

- **Various norms, sometimes implicit and mixed up:**

(Hovy, Prabhumoye, 2021) : Bias = "Differences between (a) a “**true**” or **intended** distribution (e.g., over users, labels, or outcomes), and (b) the distribution **used** or **produced** by the model."

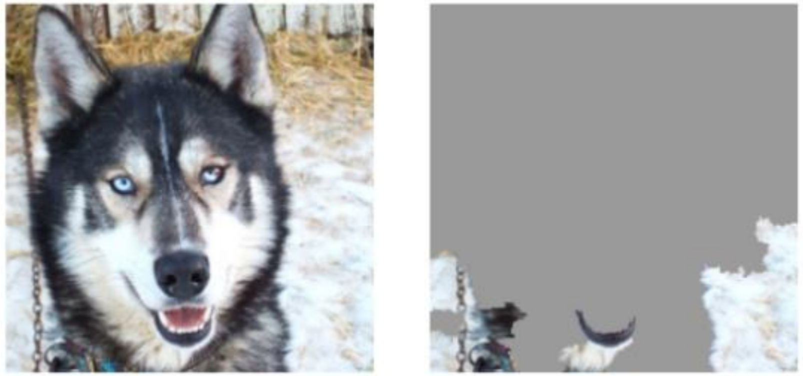
(Loubes, 2022) : Bias = "An **unfair**/**irrelevant** information that influences a decision"

(Mehrabi et al., 2022) : Bias ⬄ "Source of **unfairness**"

- Bias as a divergence from **scientific standards** (*methodological bias*)
- Bias as a divergence from **the “rational” decision** (*cognitive bias*)
- Bias as a divergence from **the ideal society** (*socio-historical bias*)

What does it mean to "unbias AI"?

- **Methodological bias**: to improve data quality and/or quantity
  A motto of ML : "Garbage in, garbage out"
- **Cognitive bias**: to improve explainability of the results
- **Socio-historical bias**: to improve fairness and diversity of the results



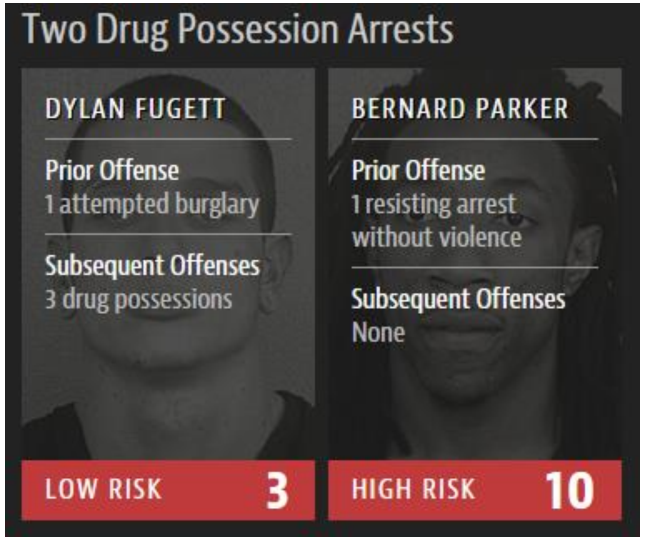
(a) Husky classified as wolf (b) Explanation

(Ribeiro et al., 2016)

Example of image classification (excluding human recognition):
Methodological & cognitive bias:
Issue of data representativeness & classification criteria

What does it mean to "unbias AI"?

- **Methodological bias**: to improve data quality and/or quantity
  A motto of ML : "Garbage in, garbage out"
- **Cognitive bias**: to improve explainability of the results
- **Socio-historical bias**: to improve fairness and diversity of the results



Two Drug Possession Arrests

DYLAN FUGETT

Prior Offense
1 attempted burglary

Subsequent Offenses
3 drug possessions

LOW RISK 3

BERNARD PARKER

Prior Offense
1 resisting arrest without violence

Subsequent Offenses
None

HIGH RISK 10

(Angwin et al., 2016)

Example of COMPAS (NorthPointe) : Socio-historical bias
Representative data (criminal justice records)
But produced in an unequal society
=> The risk assessment tool reinforces existing social inequities and stereotypes by applying group statistics to individuals

(+ Cognitive bias : No open access to the algorithm
& poor predictive results)

The pitfalls of unbiasing AI

- **Concerning methodological bias:**

To mistake the map for the territory
To forget context of production and application
To overlook the existence of spurious correlations (Calude et Longo, 2020)

"Who knows why people do what they do? The point is they do it, and we can track and measure it with unprecedented fidelity. With enough data, the numbers speak for themselves." (*The End of Theory: The Data Deluge Makes the Scientific Method Obsolete*, Chris Anderson, 06/23/2008)

=> A "positivist viewpoint" debated by the field of **critical data studies**
data are never "raw" or "objective" but always "cooked"
(Iliadis, Russo, 2016 ; D'Ignazio, Klein, 2020 ; Zacklad, Rouvroy, 2021)

The pitfalls of unbiasing AI

- **Concerning cognitive bias:**

To reduce human variability and agency (dependent on culture, emotions…)

for the sake of performance or fairness

“It has long been known that predictions and decisions generated by simple statistical algorithms are often more accurate than those made by experts, even when the experts have access to more information than the formulas use. It is less well known that the key advantage of algorithms is that they are noise-free: Unlike humans, a formula will always return the same output for any given input. Superior consistency allows even simple and imperfect algorithms to achieve greater accuracy than human professionals.” (Kahneman, 2016)

“The good news is that we have many computer scientists who care deeply about the fairness of ML algorithms, and have developed methods to make them less biased than humans.” (Jennifer Chayes, Interview to HuffPost, 2017)

The pitfalls of unbiasing AI

- **Concerning socio-historical bias:**

To “hard code” power balance and cultural norms
(Johnson et al., 2020 ; Bender et al., 2021)

To increase “information bubbles” by personalizing the results according to the users’ (real or guessed) preferences
(Mitchell et al., 2020)

Conclusion

- Definitions of “bias in AI” should be **more explicit**

- Instead of “unbiasing AI” we should talk of **“rebiaising AI”** according to a certain norm

- In some cases, the real question should not be:
  "How to unbias AI?“ but **"Should we use AI?"**

Thank you for your attention!

Contact: ambre.davat@univ-grenoble-alpes.fr

DAVAT Ambre, « Biais, intelligence artificielle et technosolutionnisme »,
Éthique, politique, religions, n° 22, 2023 – 1, L'éthique de l'intelligence artificielle à travers les dispositifs et les pouvoirs, p. 67-83