Introduction to bias and AI regulations

A journey through model debiasing: from methods to applications

Tutorial for ICIAP 2025

https://a-journey-through-model-debiasing.github.io/

15/09/2025

Enzo Tartaglione

Associate Professor, responsible for the Multimedia group

Télécom Paris, Institut Polytechnique de Paris

enzo.tartaglione@telecom-paris.fr

Outline

- The GDPR, the AI Act, and the (future?) regulatory perspective
- Towards a definition of bias: how we perceive it, what it is
- Bias and fairness: where is the gap?
- Ways to solve model bias: unlearning? What if the bias is unknown?

What is the GDPR?

- EU regulation into effect on 15 May 2018.
- “The GDPR will levy harsh fines against those who violate its privacy and security standards, with penalties reaching into the **tens of millions of euros**”.
- From the EU Convention of Human Rights (1950s) “Everyone has the right to respect for his private and family life, his home and his correspondence.”
- Focus on DATA and its PROTECTION.

Source: https://gdpr.eu/what-is-gdpr/

The EU Artificial Intelligence Act

- EU regulation into force since 1 Aug 2024.

- Covers all types of AI across a broad range of sectors, with exceptions for AI systems used solely for military, national security, research and non-professional purposes.

- For general-purpose AI (foundation model), transparency requirements are imposed, with reduced requirements for open-source models, and additional evaluations for high-capability models.

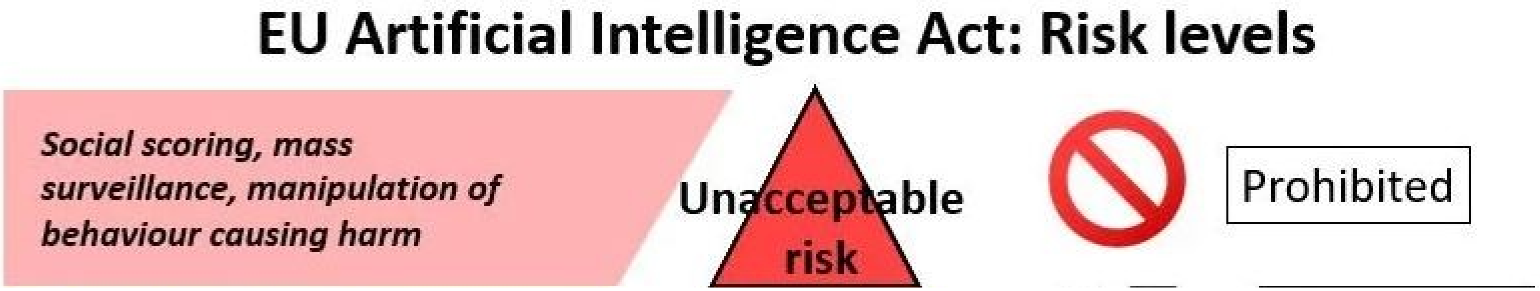
EU Artificial
Intelligence Act

What is the difference between GDPR and AI Act?

- **GDPR**: focus on privacy and data protection, giving penalties for non-compliance.
  AI Act: regulation on developed technologies (so, dictates rules for compliance), ranked on risk levels.

- Both focus on transparency and accountability. Their intersection happens when personal data are used within AI technology.

Risk levels

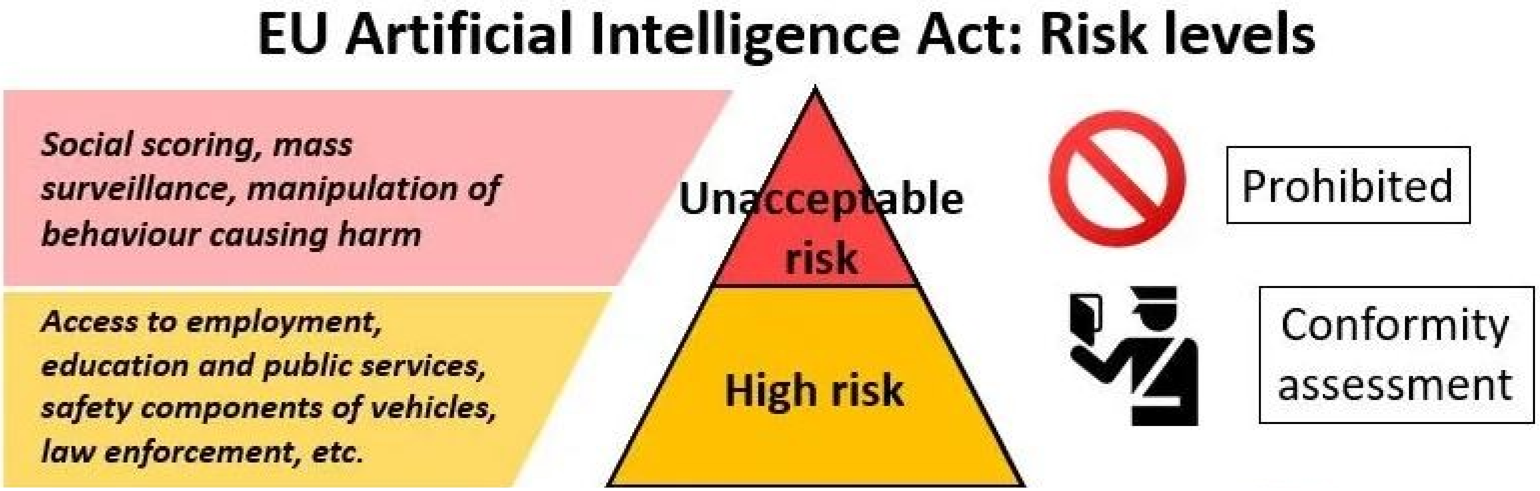


EU Artificial Intelligence Act: Risk levels

Social scoring, mass surveillance, manipulation of behaviour causing harm
Unacceptable risk
Prohibited

Image taken from https://datasciencedojo.com/blog/eu-ai-act/

Risk levels



EU Artificial Intelligence Act: Risk levels

Social scoring, mass surveillance, manipulation of behaviour causing harm

Access to employment, education and public services, safety components of vehicles, law enforcement, etc.

Image taken from https://datasciencedojo.com/blog/eu-ai-act/

Risk levels

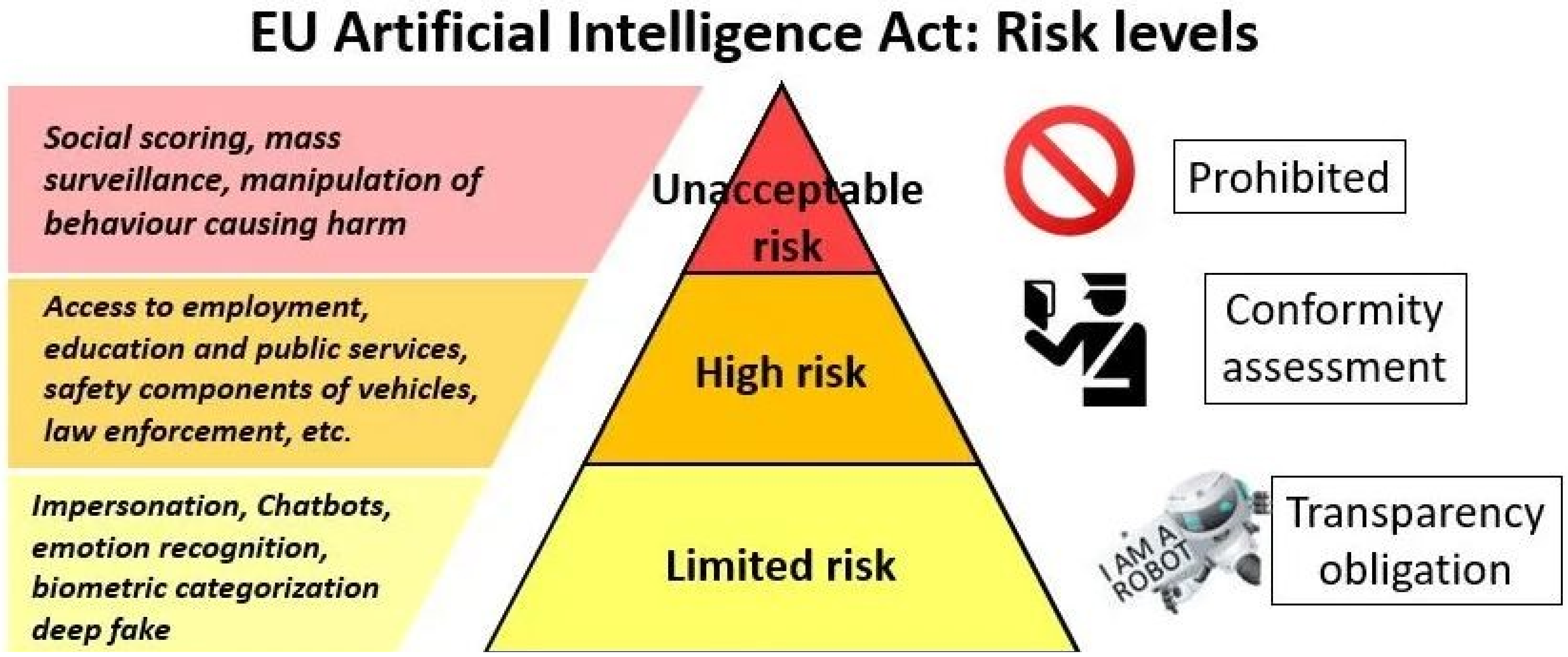
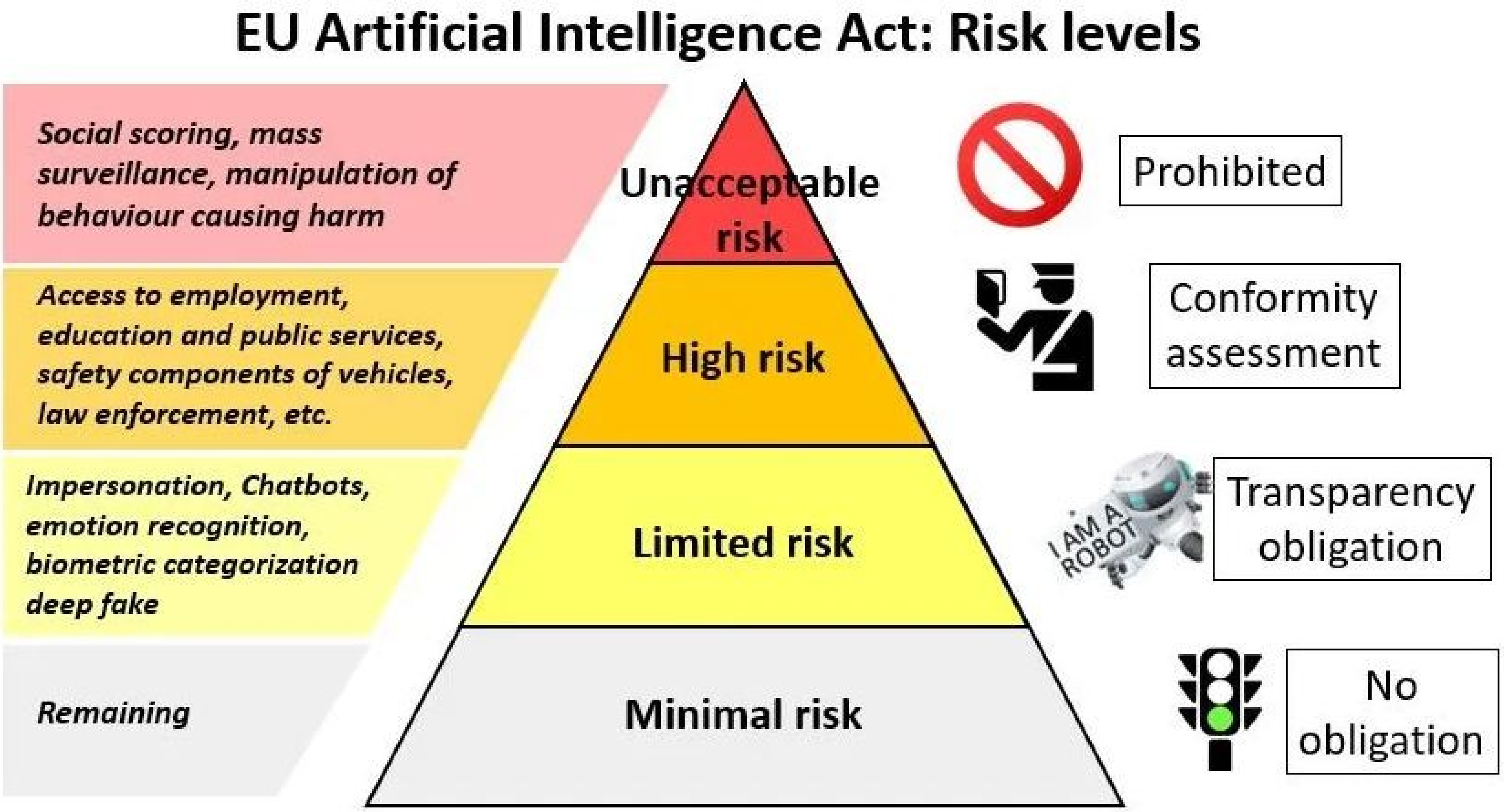

EU Artificial Intelligence Act: Risk levels
Social scoring, mass surveillance, manipulation of behaviour causing harm
Unacceptable risk
Prohibited
Access to employment, education and public services, safety components of vehicles, law enforcement, etc.
High risk
Conformity assessment
Impersonation, Chatbots, emotion recognition, biometric categorization deep fake
Limited risk
I AM A ROBOT
Transparency obligation

Image taken from https://datasciencedojo.com/blog/eu-ai-act/

Risk levels



EU Artificial Intelligence Act: Risk levels

Social scoring, mass surveillance, manipulation of behaviour causing harm

Unacceptable risk

Prohibited

Access to employment, education and public services, safety components of vehicles, law enforcement, etc.

High risk

Conformity assessment

Impersonation, Chatbots, emotion recognition, biometric categorization deep fake

Limited risk

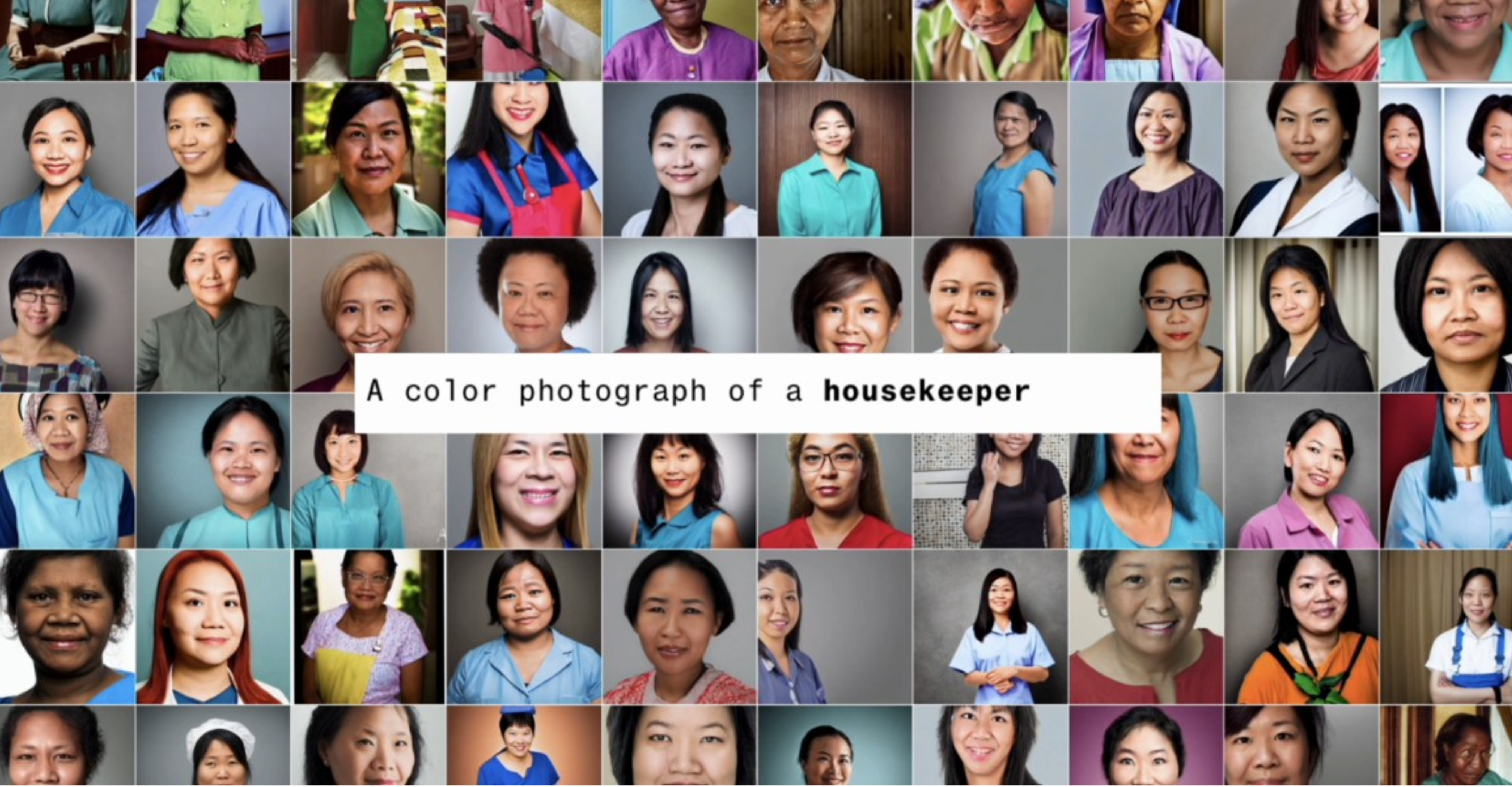
Transparency obligation

Remaining

Minimal risk

No obligation

Image taken from https://datasciencedojo.com/blog/eu-ai-act/



A color photograph of a **housekeeper**

Bias in AI models

- Model Bias occurs when the model itself is not able to accurately represent the underlying relationship between the input features and the output variable.

- Simply: the model captures some **spurious relations**, harming the performance at test time.

- We can solve this problem providing metadata of these correlations. However, this is an expensive and sometimes is even unfeasible (*e.g.*, when these are not known a-priori).

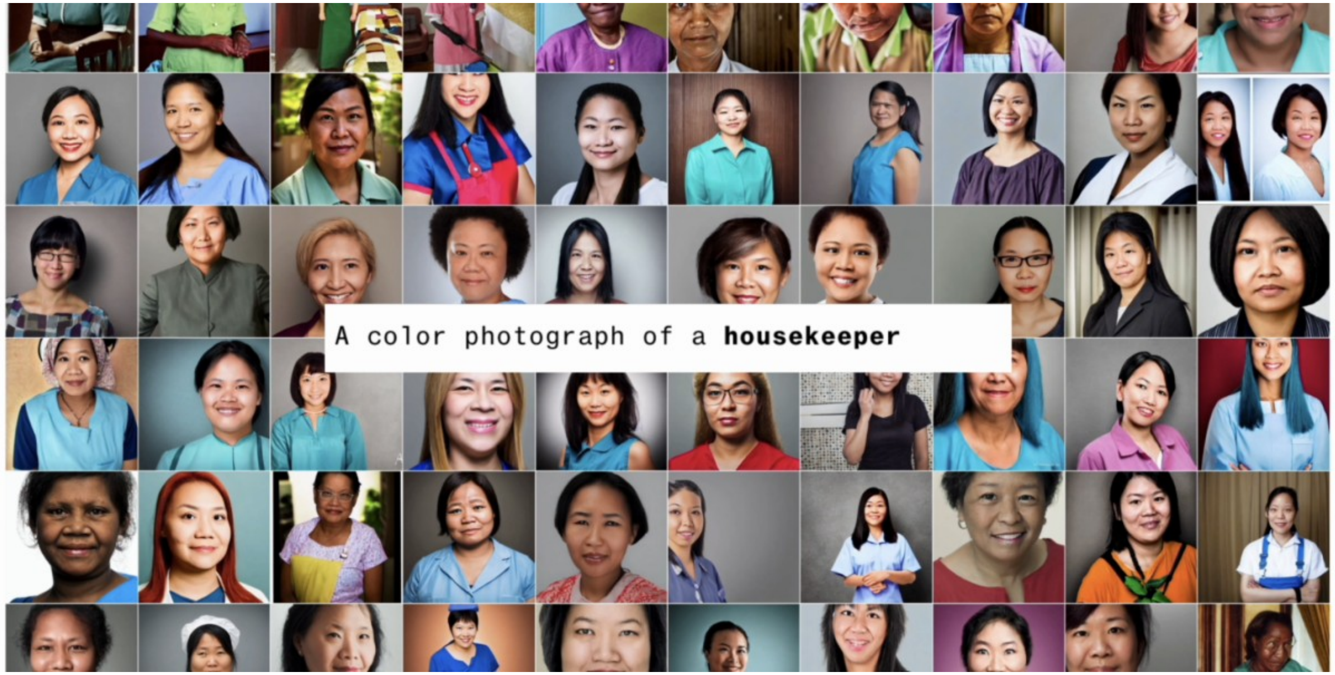


Image taken from https://www.bloomberg.com/graphics/2023-generative-ai-bias/

Bias from the AI Act perspective

From the Article 10, Data and data governance:
Training, validation and testing data sets shall be subject to […] management practices appropriate for the intended purpose of the high-risk AI system. Those practices shall concern in particular: […]

Source: https://artificialintelligenceact.eu/article/10/

Bias from the AI Act perspective

From the Article 10, Data and data governance:
Training, validation and testing data sets shall be subject to […] management practices appropriate for the intended purpose of the high-risk AI system. Those practices shall concern in particular: […]

- (f) examination in view of possible biases that [...] have a negative impact on fundamental rights or lead to discrimination prohibited under Union law[...];

Source: https://artificialintelligenceact.eu/article/10/

Bias from the AI Act perspective

From the Article 10, Data and data governance:
Training, validation and testing data sets shall be subject to […] management practices appropriate for the intended purpose of the high-risk AI system. Those practices shall concern in particular: […]

- (f) examination in view of possible biases that [...] have a negative impact on fundamental rights or lead to discrimination prohibited under Union law[...];

- (g) appropriate measures to detect, **prevent** and mitigate possible biases identified according to point (f).

Source: https://artificialintelligenceact.eu/article/10/

Algorithmic VS Societal bias

- **Societal Bias** refers to the prejudices, stereotypes, and inequalities that exist in society. These biases are embedded in cultural norms, institutions, and social practices, and can shape the data that algorithms use.

Algorithmic VS Societal bias

- **Societal Bias** refers to the prejudices, stereotypes, and inequalities that exist in society. These biases are embedded in cultural norms, institutions, and social practices, and can shape the data that algorithms use.

- **Algorithmic Bias** occurs also when these societal biases are encoded into algorithmic systems, either through biased data, biased model design, or biased implementation. While societal biases are often the source, algorithmic bias refers to how these biases manifest in automated systems, potentially amplifying or perpetuating existing inequalities.

Algorithmic VS Societal bias

- **Societal Bias** refers to the prejudices, stereotypes, and inequalities that exist in society. These biases are embedded in cultural norms, institutions, and social practices, and can shape the data that algorithms use.

- **Algorithmic Bias** occurs also when these societal biases are encoded into algorithmic systems, either through biased data, biased model design, or biased implementation. While societal biases are often the source, algorithmic bias refers to how these biases manifest in automated systems, potentially amplifying or perpetuating existing inequalities.

Bias: definitions

"Algorithmic bias occurs when the mathematical models and algorithms we rely on reflect the existing biases of the data used to train them, leading to unfair and often harmful outcomes."

O’Neill, C. (2016). Weapons of Math Destruction. Crown books.

Bias: definitions

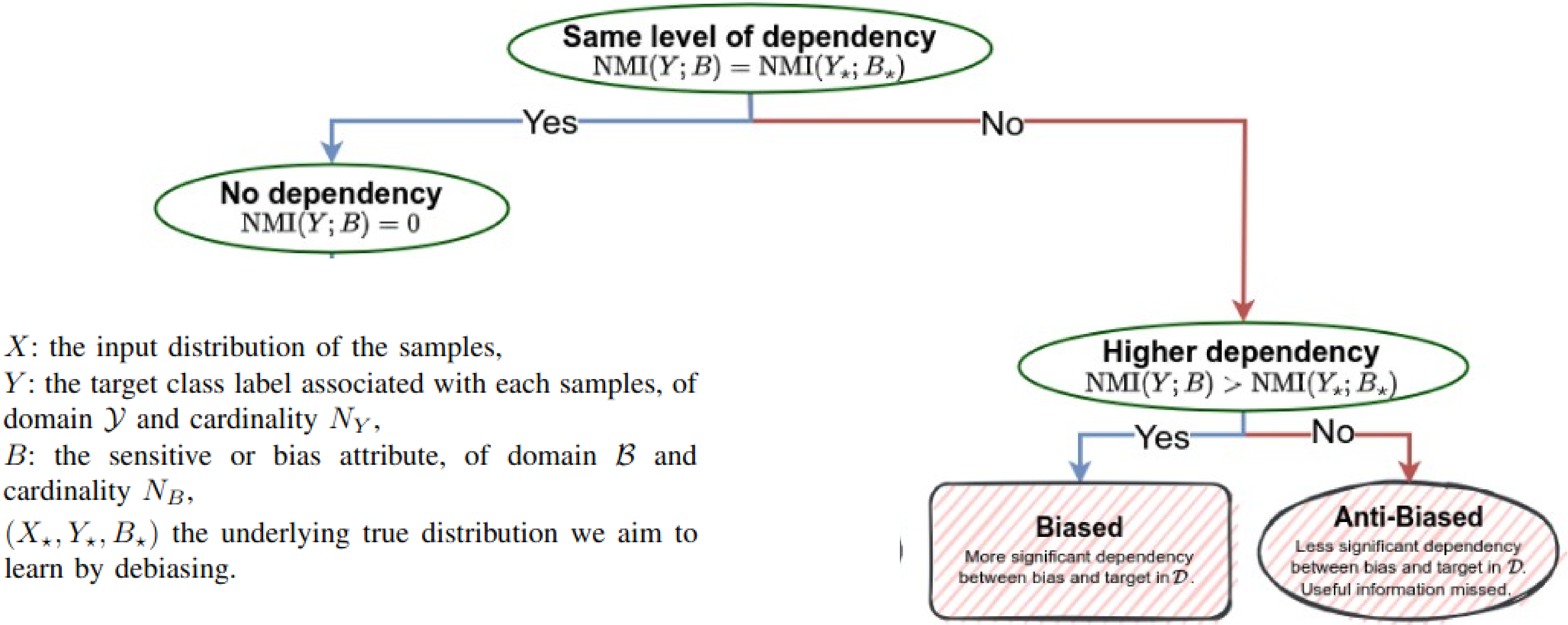
"Algorithmic bias occurs when the mathematical models and algorithms we rely on reflect the existing biases of the data used to train them, leading to unfair and often harmful outcomes."

O’Neill, C. (2016). Weapons of Math Destruction. Crown books.

"Algorithmic bias refers to the systematic and unfair discrimination in the decisions made by algorithms, often resulting from biased data, design choices, or the way an algorithm's outputs are interpreted and used."

Mitchell, M. (2019). Artificial Intelligence: A Guide for Thinking Humans. Farrar, Straus and Giroux. MacMillan.

Towards a more formal definition



Same level of dependency
$\mathrm{NMI}(Y;B) = \mathrm{NMI}(Y_\star;B_\star)$

Yes

No dependency
$\mathrm{NMI}(Y;B) = 0$

No

Higher dependency
$\mathrm{NMI}(Y;B) > \mathrm{NMI}(Y_\star;B_\star)$

Yes

Biased
More significant dependency between bias and target in $\mathcal{D}$.

No

Anti-Biased
Less significant dependency between bias and target in $\mathcal{D}$. Useful information missed.

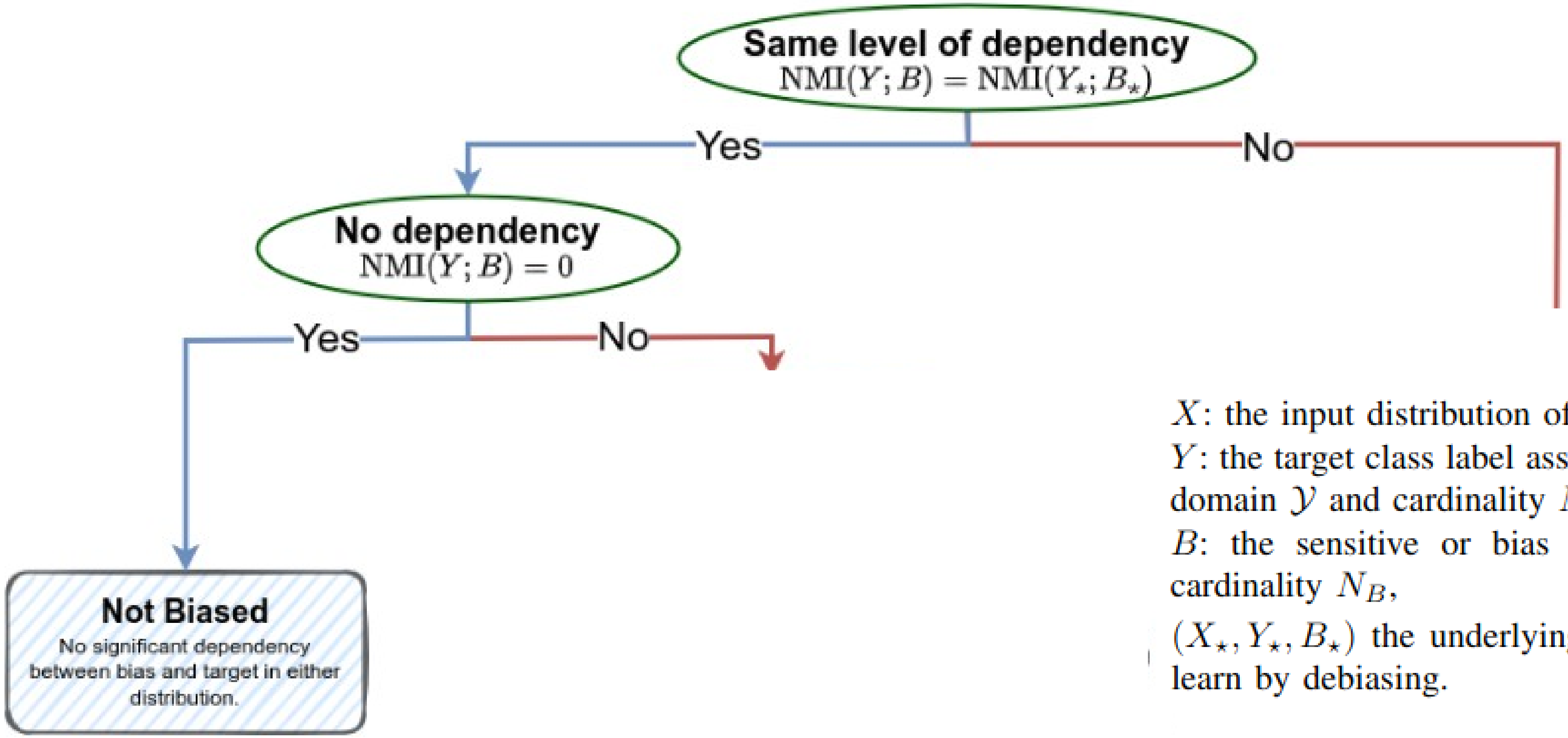
X: the input distribution of the samples,
Y: the target class label associated with each samples, of domain $\mathcal{Y}$ and cardinality N_Y,
B: the sensitive or bias attribute, of domain $\mathcal{B}$ and cardinality N_B,

$(X_\star, Y_\star, B_\star)$ the underlying true distribution we aim to learn by debiasing.

$$\mathrm{NMI}(Y;B) = \frac{2 \times I(Y;B)}{H(Y) + H(B)}$$

Are We Unbiased Yet? A Survey on Model Debiasing for Image Classification. Preprint.

Towards a more formal definition



Same level of dependency
$\mathrm{NMI}(Y;B) = \mathrm{NMI}(Y_\star;B_\star)$

Yes / No

No dependency
$\mathrm{NMI}(Y;B) = 0$

Yes / No

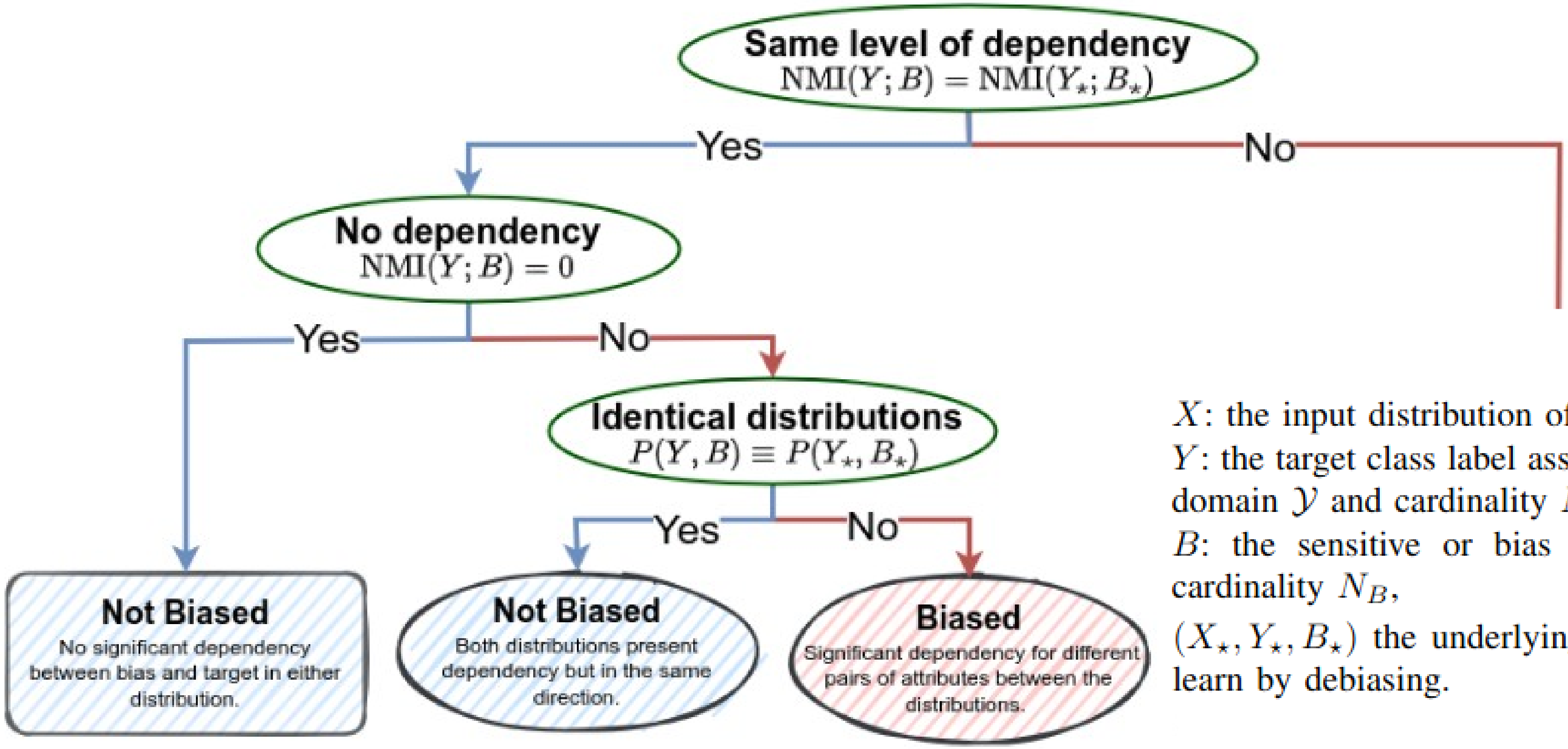
Not Biased
No significant dependency between bias and target in either distribution.

X: the input distribution of the samples,
Y: the target class label associated with each samples, of domain $\mathcal{Y}$ and cardinality N_Y,
B: the sensitive or bias attribute, of domain $\mathcal{B}$ and cardinality N_B,
$(X_\star, Y_\star, B_\star)$ the underlying true distribution we aim to learn by debiasing.

$$\mathrm{NMI}(Y;B) = \frac{2 \times I(Y;B)}{H(Y) + H(B)}$$

Are We Unbiased Yet? A Survey on Model Debiasing for Image Classification. Preprint.

Towards a more formal definition

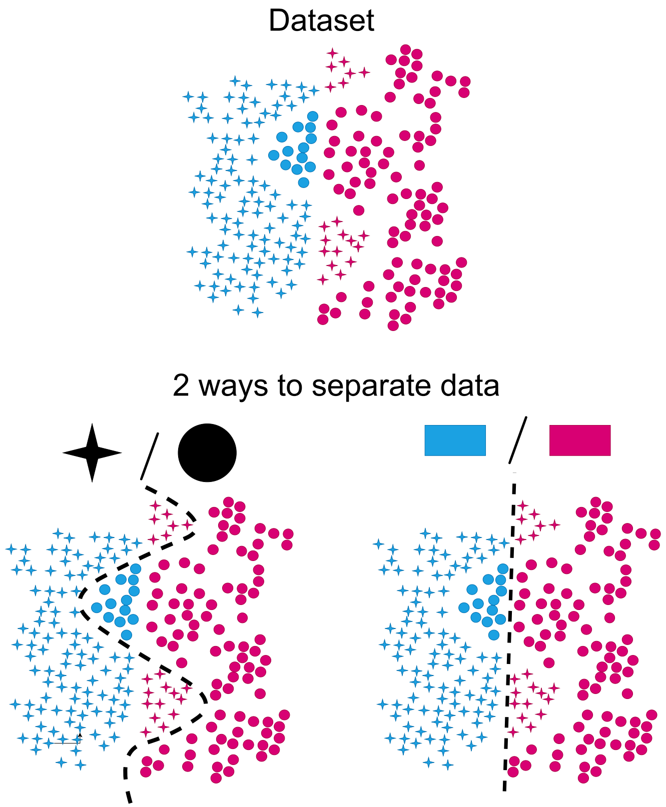

Same level of dependency
$\mathrm{NMI}(Y;B) = \mathrm{NMI}(Y_\star;B_\star)$
Yes
No
No dependency
$\mathrm{NMI}(Y;B) = 0$
Yes
No
Identical distributions
$P(Y,B) \equiv P(Y_\star,B_\star)$
Yes
No
Not Biased
No significant dependency between bias and target in either distribution.
Not Biased
Both distributions present dependency but in the same direction.
Biased
Significant dependency for different pairs of attributes between the distributions.

X: the input distribution of the samples,
Y: the target class label associated with each samples, of domain $\mathcal{Y}$ and cardinality N_Y,
B: the sensitive or bias attribute, of domain $\mathcal{B}$ and cardinality N_B,
$(X_\star, Y_\star, B_\star)$ the underlying true distribution we aim to learn by debiasing.

$$\mathrm{NMI}(Y;B) = \frac{2 \times I(Y;B)}{H(Y) + H(B)}$$

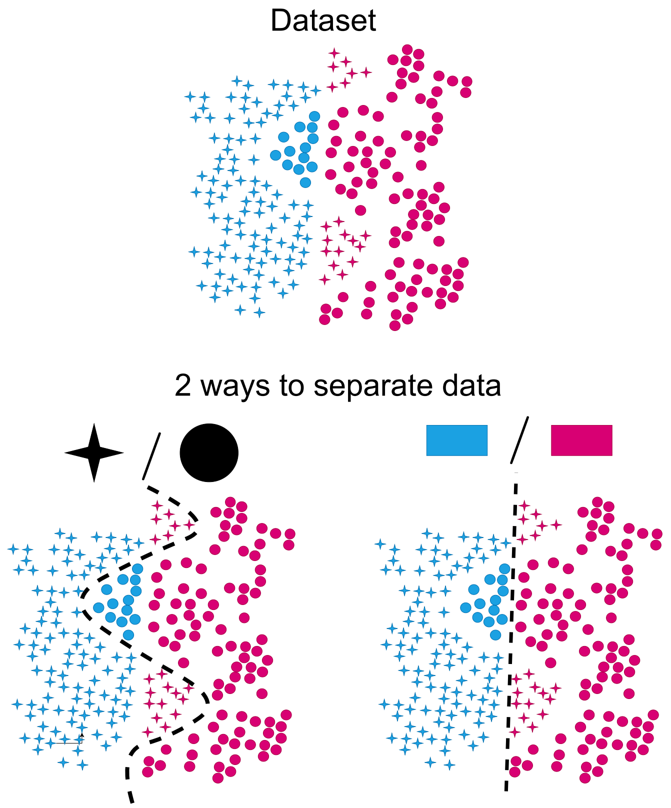
Are We Unbiased Yet? A Survey on Model Debiasing for Image Classification. Preprint.

Towards a more formal definition

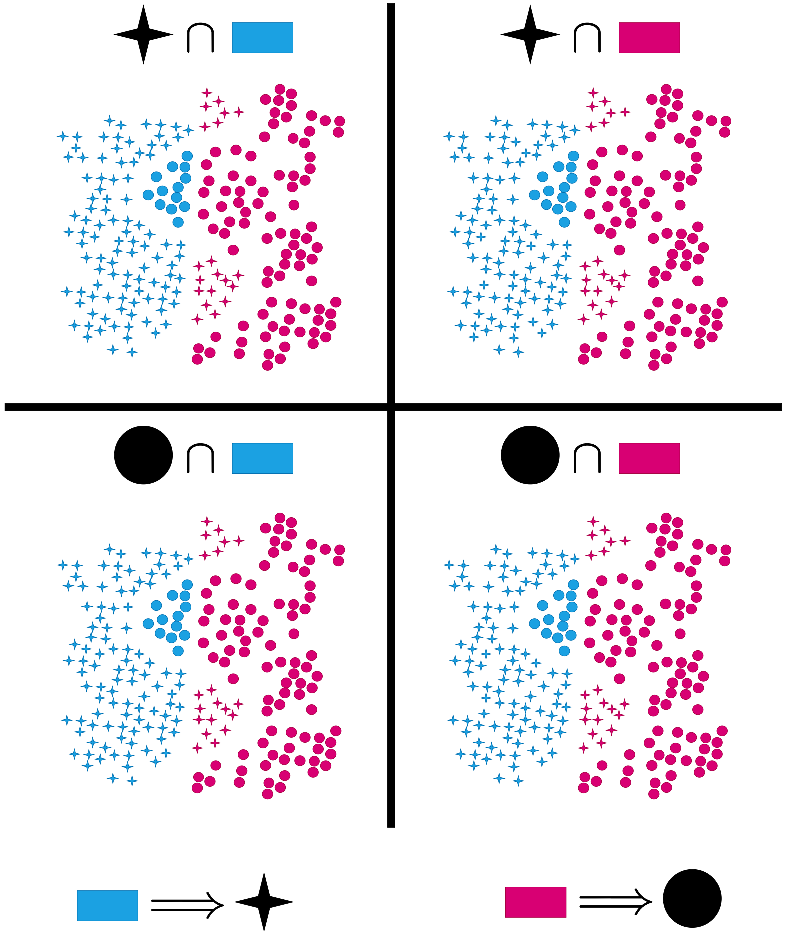


Are We Unbiased Yet? A Survey on Model Debiasing for Image Classification. Preprint.

Towards a more formal definition



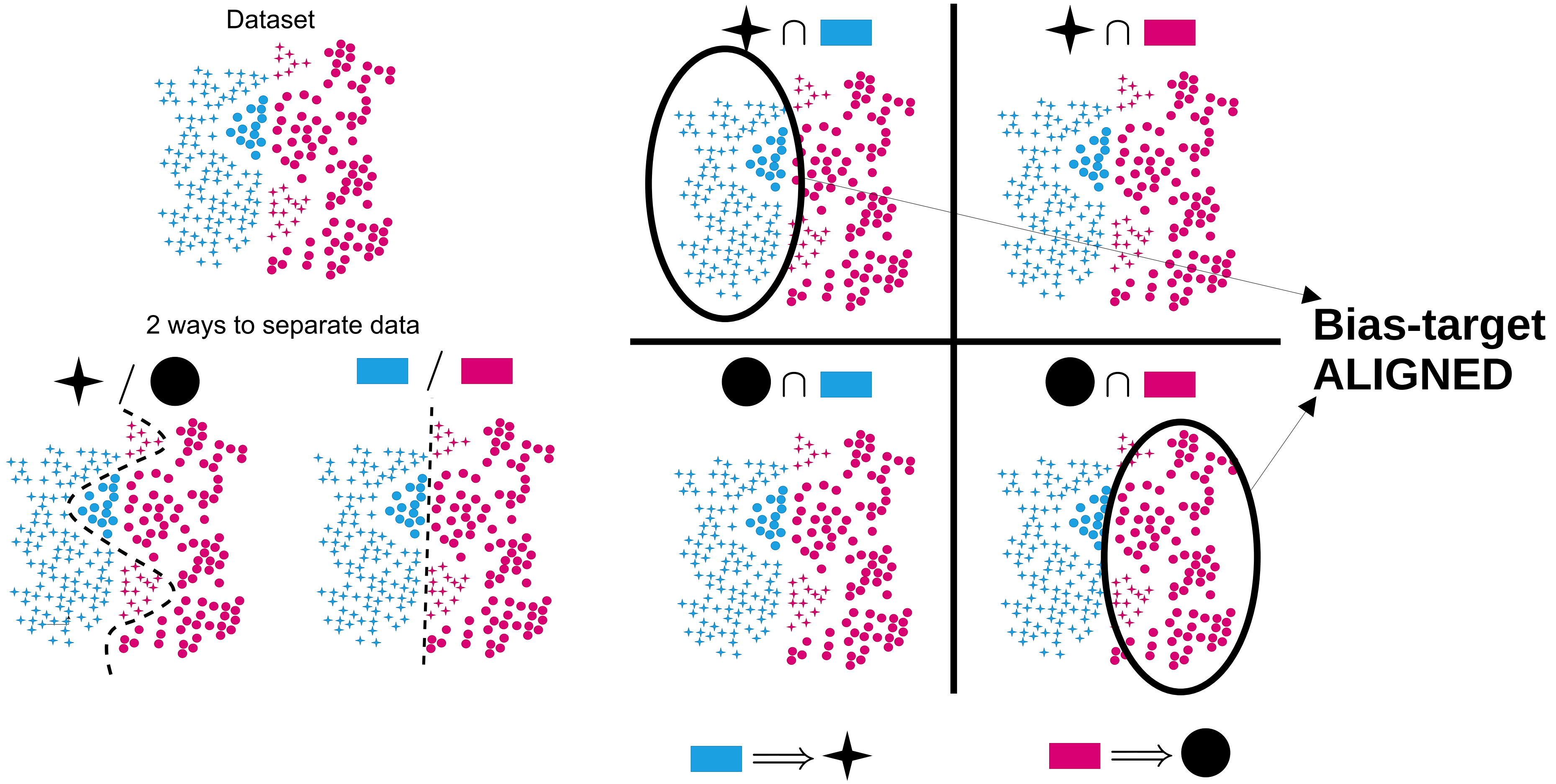
Overlap between subgroups:
Can one infer the target (✦/●) from the bias (🟦/🟥) ?



Are We Unbiased Yet? A Survey on Model Debiasing for Image Classification. Preprint.

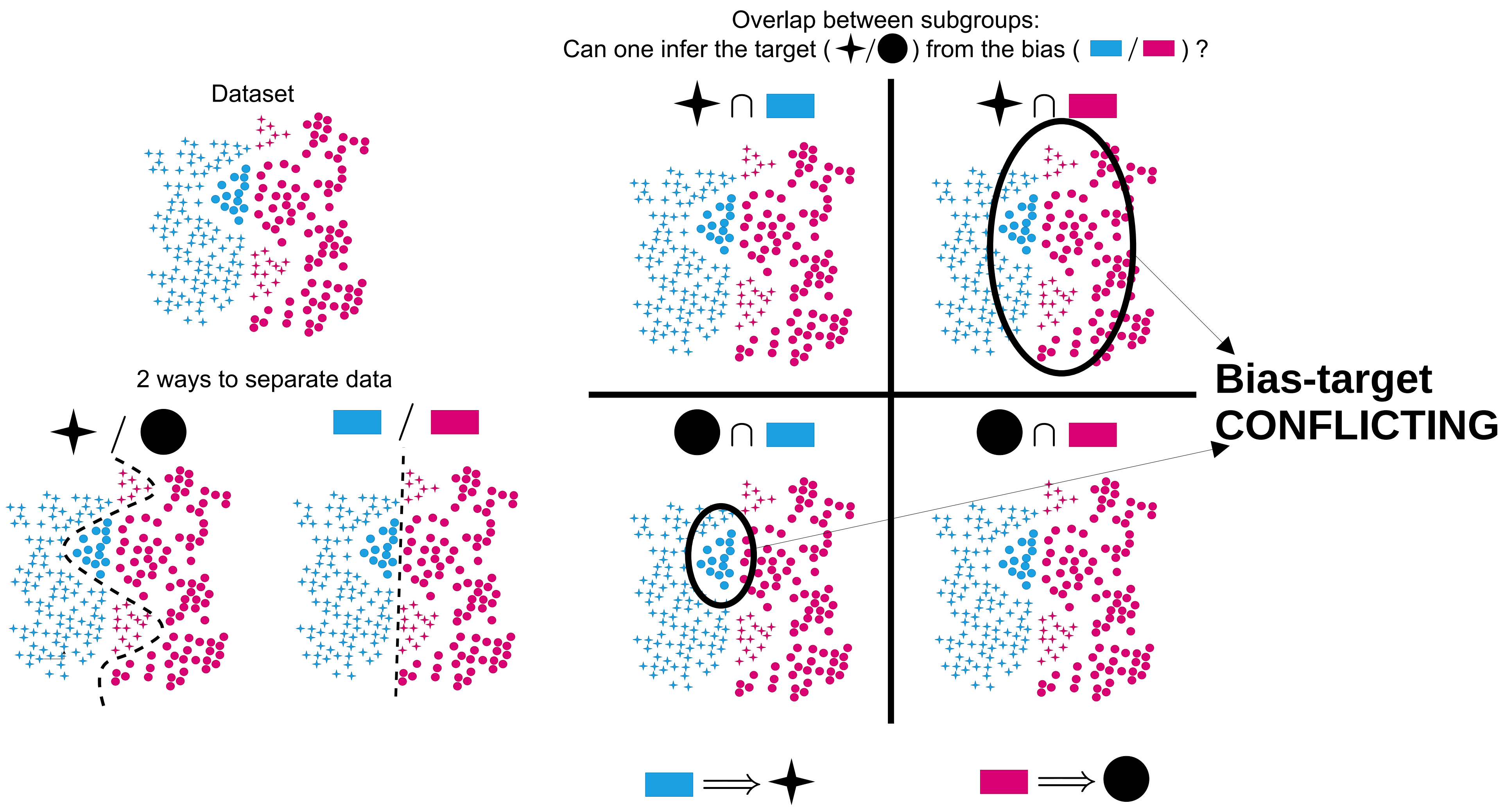
Towards a more formal definition

Overlap between subgroups:
Can one infer the target (✦/●) from the bias (🟦/🟥) ?



Are We Unbiased Yet? A Survey on Model Debiasing for Image Classification. Preprint.

Towards a more formal definition



Overlap between subgroups:
Can one infer the target (✦/●) from the bias (🟦/🟥) ?

Dataset

2 ways to separate data

✦ / ●

🟦 / 🟥

✦ ∩ 🟦

✦ ∩ 🟥

● ∩ 🟦

● ∩ 🟥

Bias-target CONFLICTING

🟦 ⟹ ✦

🟥 ⟹ ●

Are We Unbiased Yet? A Survey on Model Debiasing for Image Classification. Preprint.

Link with fairness

- Importance of the data distribution: fairness (generally) looks for balance regardless the alignment to the true distribution, debiasing targets specific attributes that are making the distribution deviating from the true.

Are We Unbiased Yet? A Survey on Model Debiasing for Image Classification. Preprint.

Link with fairness

- Importance of the data distribution: fairness (generally) looks for balance regardless the alignment to the true distribution, debiasing targets specific attributes that are making the distribution deviating from the true.

- Fairness’ goal is to achieve parity (to some extents), debiasing targets achievement of a natural distribution alignment, accepting populations imbalances.

- Where is the link?

Are We Unbiased Yet? A Survey on Model Debiasing for Image Classification. Preprint.

Link with fairness

- Importance of the data distribution: fairness (generally) looks for balance regardless the alignment to the true distribution, debiasing targets specific attributes that are making the distribution deviating from the true.

- Fairness' goal is to achieve parity (to some extents), debiasing targets achievement of a natural distribution alignment, accepting populations imbalances.

- Where is the link? The "true" distribution considered in debiasing is many times a balanced distribution.

Are We Unbiased Yet? A Survey on Model Debiasing for Image Classification. Preprint.

Link with unlearning?

- Ignoring certain features that are “spurious” is the objective of debiasing.
- Forgetting a part of information (features) is the objective of unlearning (for legal issues, safety, etc.)
- “Biases arise because they are learned before the actual features” → are we doomed to fit biases before learning the true features?

Are We Unbiased Yet? A Survey on Model Debiasing for Image Classification. Preprint.

Supervised vs Unsupervised debiasing

- **Supervised debiasing** refers to approaches that remove the bias when the information related to the bias is already provided.
  - Dataset cleanup approaches
  - Model post-processing
  - In-model approaches (features balance, gradient inversion etc.)

Supervised vs Unsupervised debiasing

- **Supervised debiasing** refers to approaches that remove the bias when the information related to the bias is already provided.
  - Dataset cleanup approaches
  - Model post-processing
  - In-model approaches (features balance, gradient inversion etc.)

- **Unsupervised debiasing** refers to techniques that identifies and removes the bias without provided information. Some assumptions are always taken:
  - Specific biases are searched for in the dataset (you use a proxy model to find potential biases) – Bias-Tailored approaches (BT)
  - Biases are learned earlier in the training process, and the model fits them better than the target ones (our assumption).

Do you want to reach me by email?
enzo.tartaglione@telecom-paris.fr

Curious about my research? https://enzotarta.github.io/

This slides are downloadable at the link provided in the QRCode here below