LEADDEV SAN FRANCISCO 2022

TALK

Towards global and human-centered explanations for machine learning models

CARLA VIEIRA

DATA ENGINEER AND AI ETHICS RESEARCHER

Get to Know Me

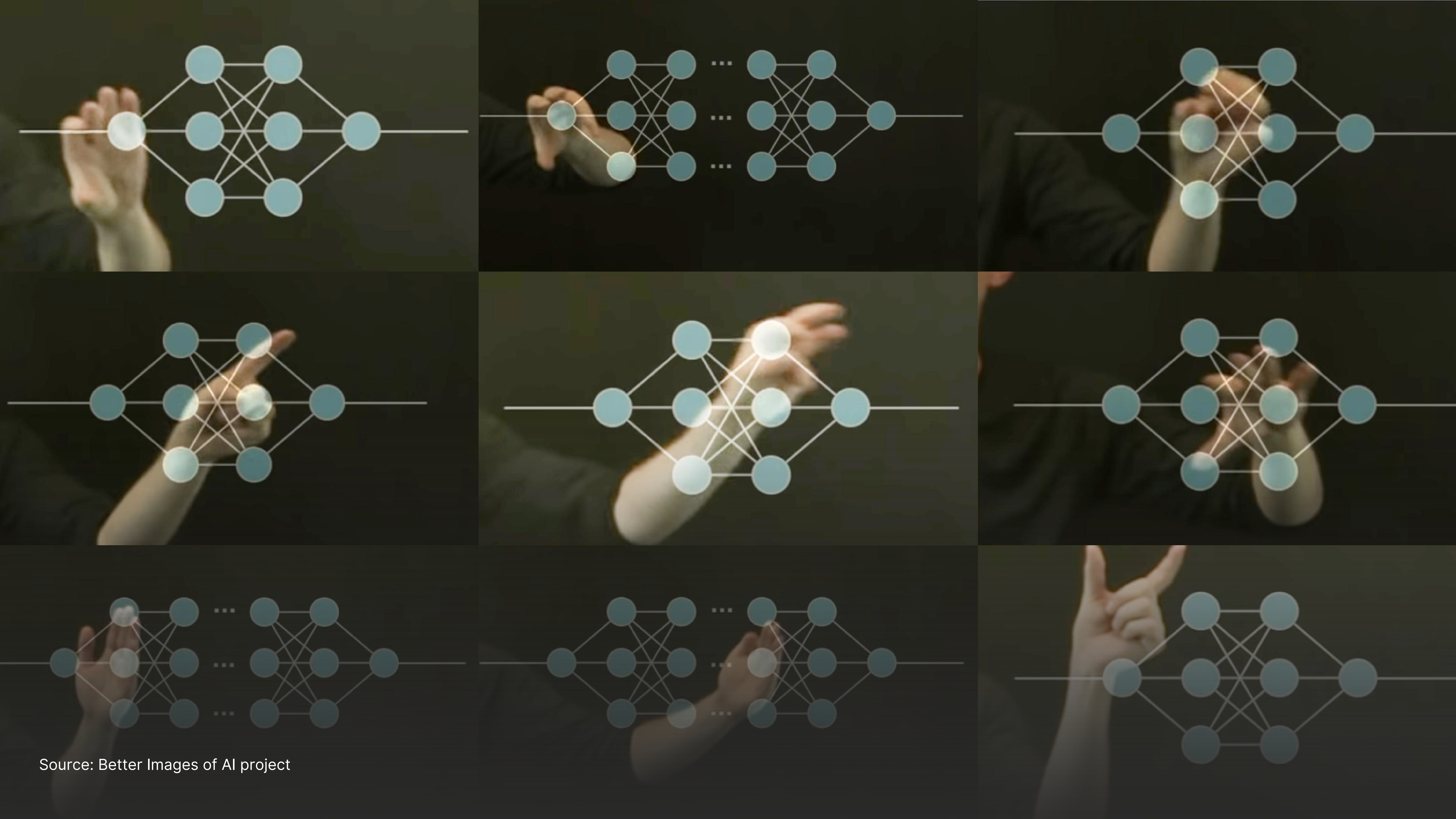
I'm Carla, Data Engineer and Google Developer Expert in Machine Learning. Master student in Artificial Intelligence.

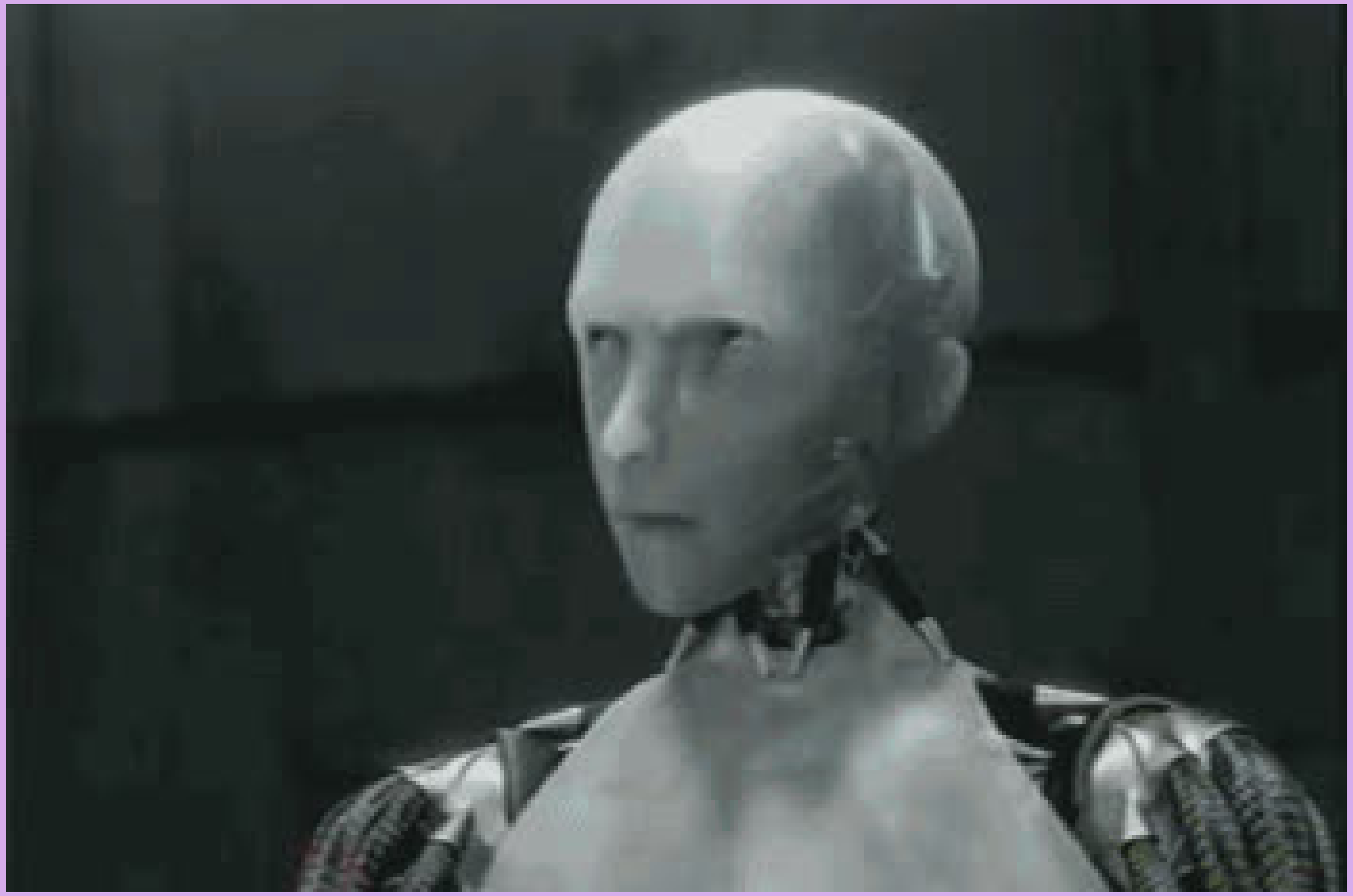
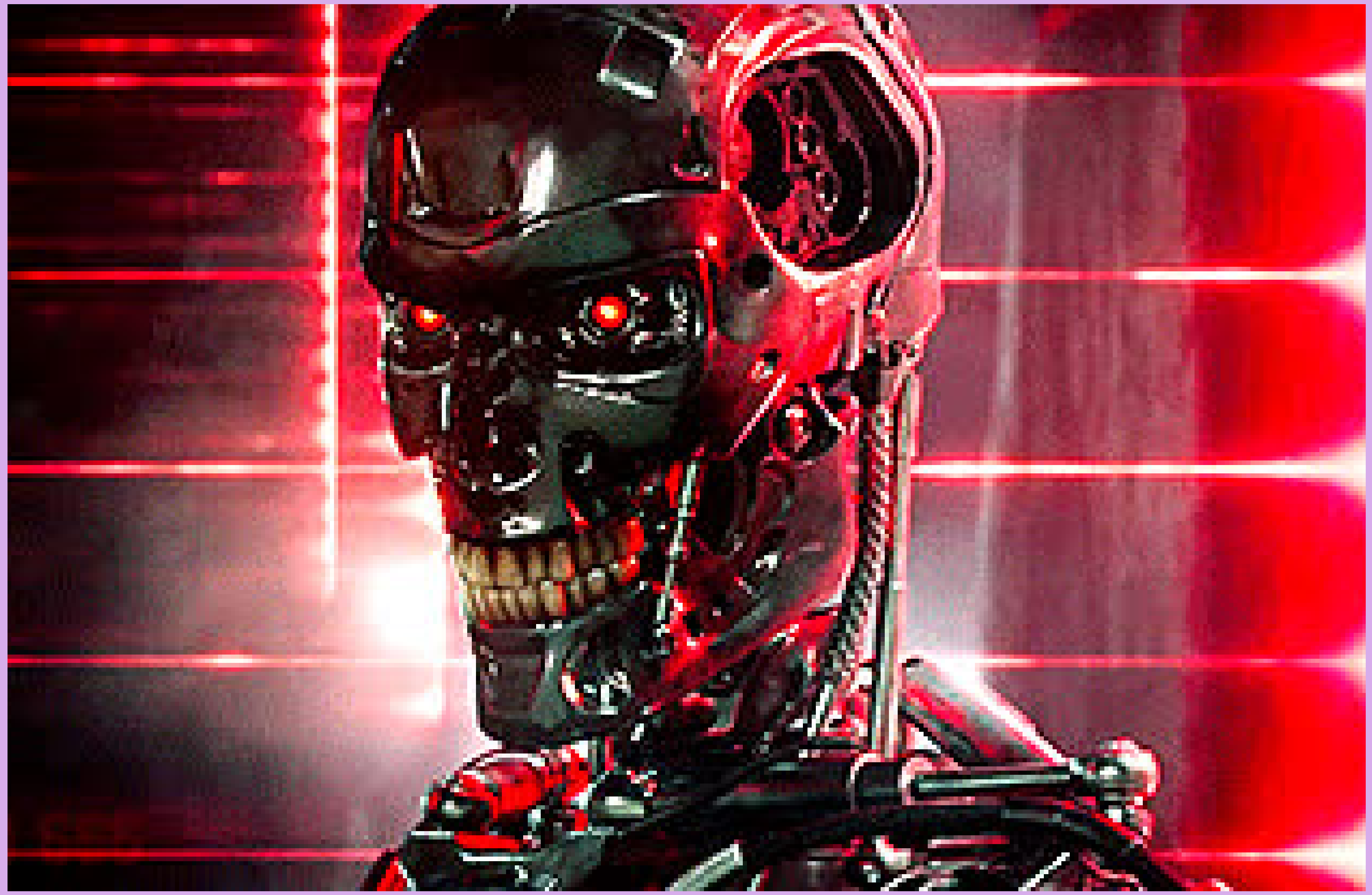
Fun facts:

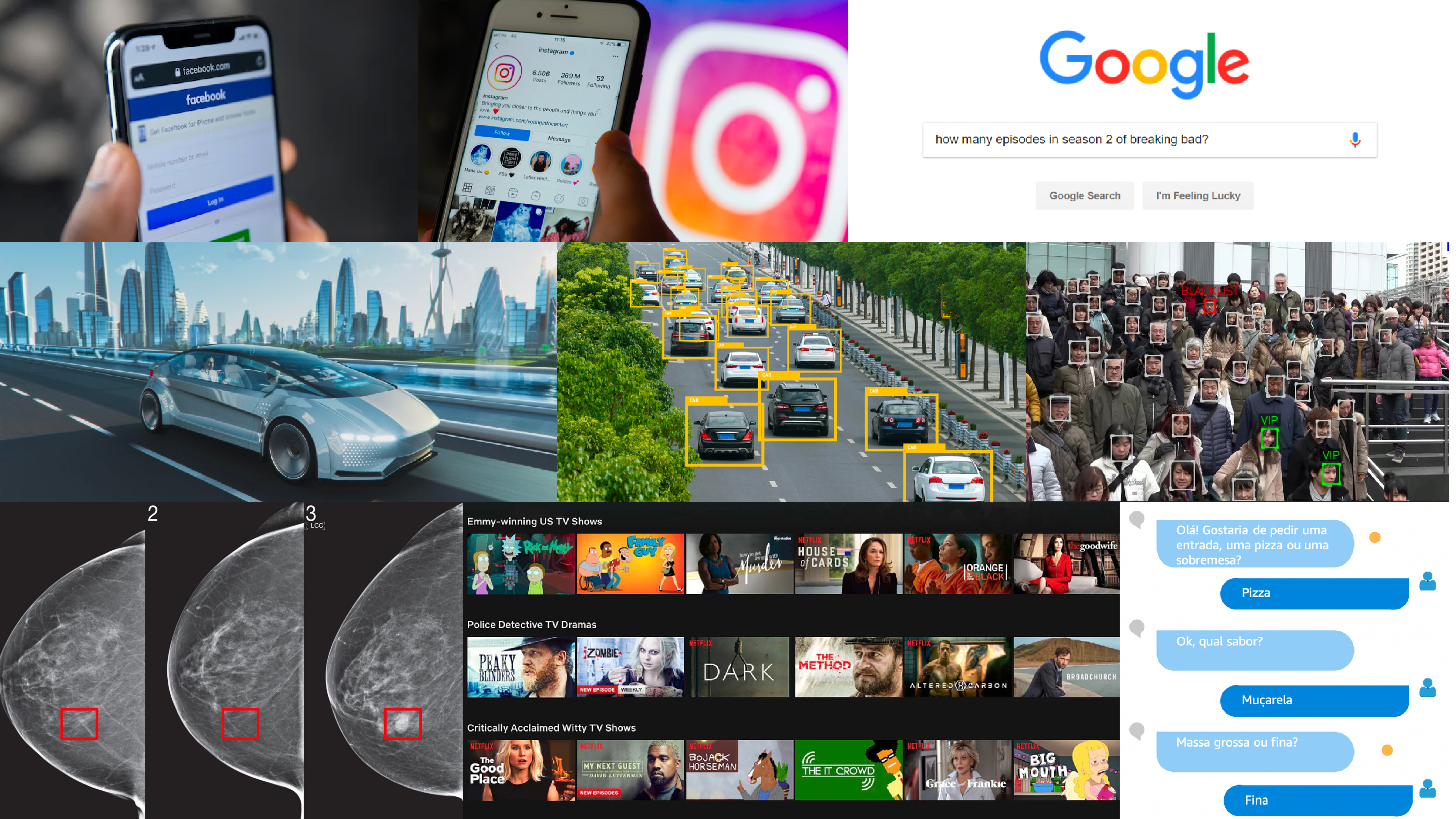
- First time in the U.S.A
- First time speaking in an international conference
- First LeadDev Event

@carlaprvieira / carlavieira.dev

Experts
2022
Machine Learning


Source: Better Images of AI project


facebook.com
facebook
Log In
instagram
6.506 Posts
369 M Followers
52 Following
Instagram
Bringing you closer to the people and things you love.
www.instagram.com/votinginfocenter/
Follow
Message
Google
how many episodes in season 2 of breaking bad?
Google Search
I'm Feeling Lucky
BLACKLIST
VIP
VIP
2
3
Emmy-winning US TV Shows
Police Detective TV Dramas
Critically Acclaimed Witty TV Shows
Olá! Gostaria de pedir uma entrada, uma pizza ou uma sobremesa?
Pizza
Ok, qual sabor?
Muçarela
Massa grossa ou fina?
Fina

Potential Harms Caused by AI Systems

Leslie, D. (2019). Understanding artificial intelligence ethics and safety: A guide for the responsible design and implementation of AI systems in the public sector. The Alan Turing Institute.

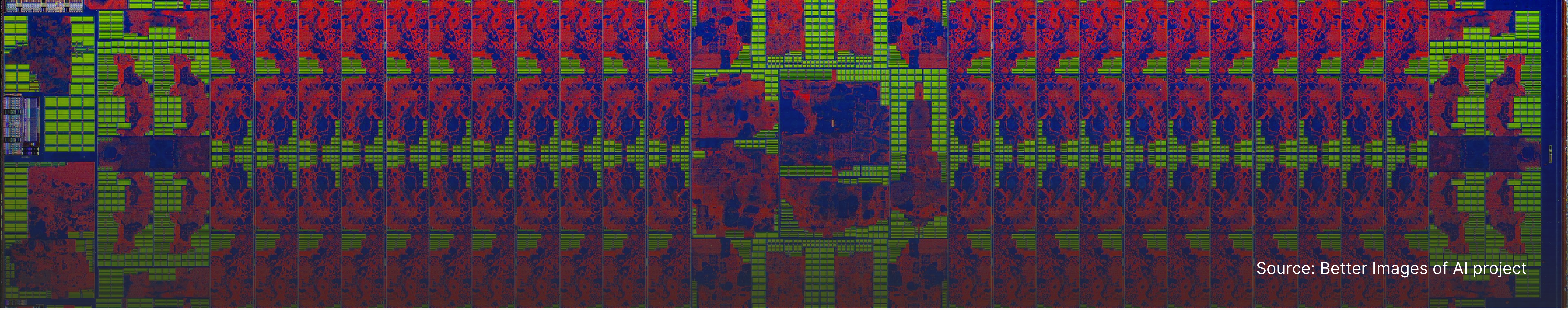
01 **BIAS AND DISCRIMINATION**

02 DENIAL OF INDIVIDUAL AUTONOMYAND RIGHTS

03 **NON-TRANSPARENT, UNEXPLAINABLE, OR UNJUSTIFIABLE OUTCOMES**

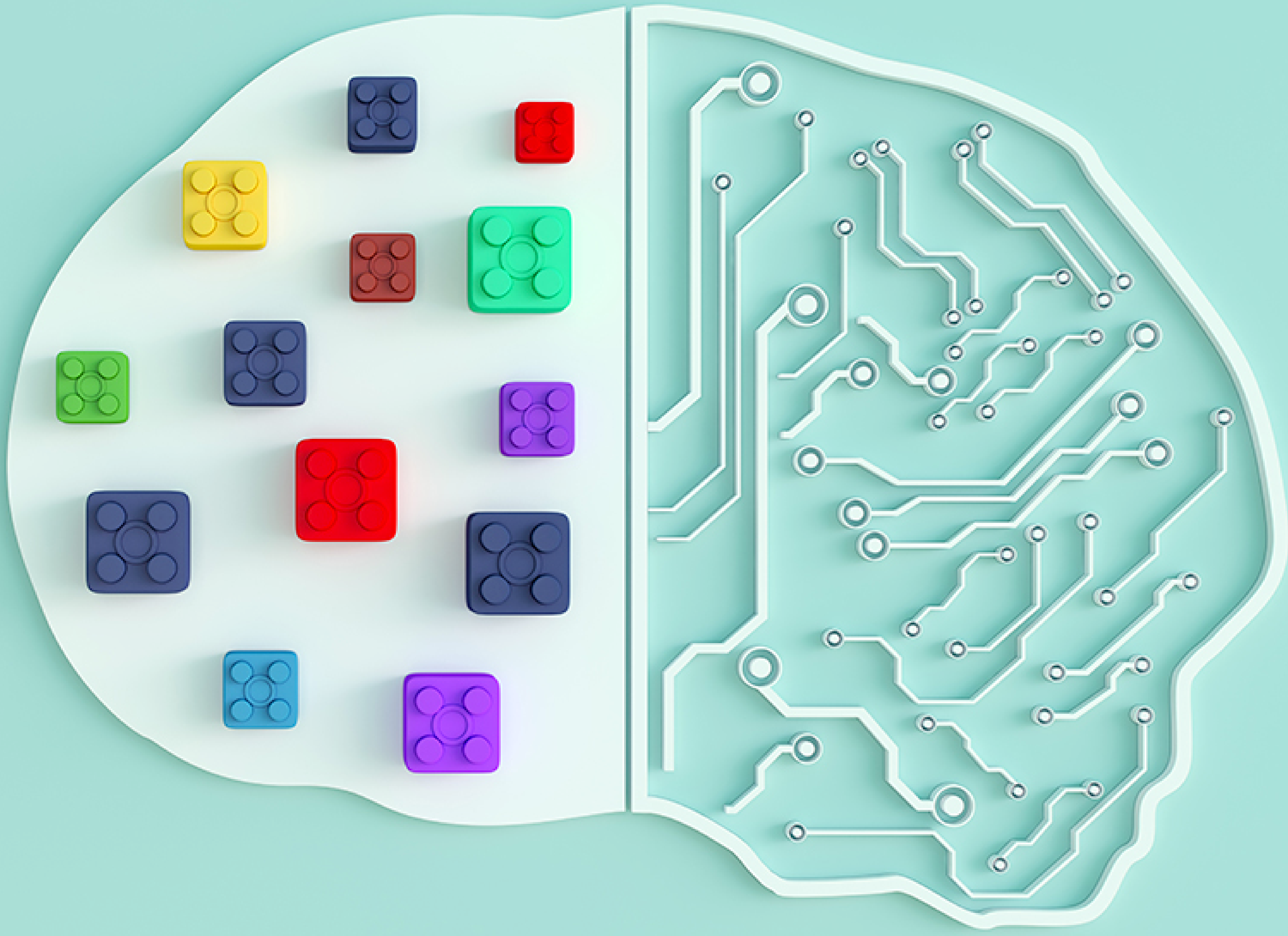
04 INVASIONS OF PRIVACY

05 UNRELIABLE, UNSAFE, OR POOR-QUALITY OUTCOMES



What is bias in ML/AI?

Algorithmic bias is when a computer system reflects the **implicit values of the humans** who created it.

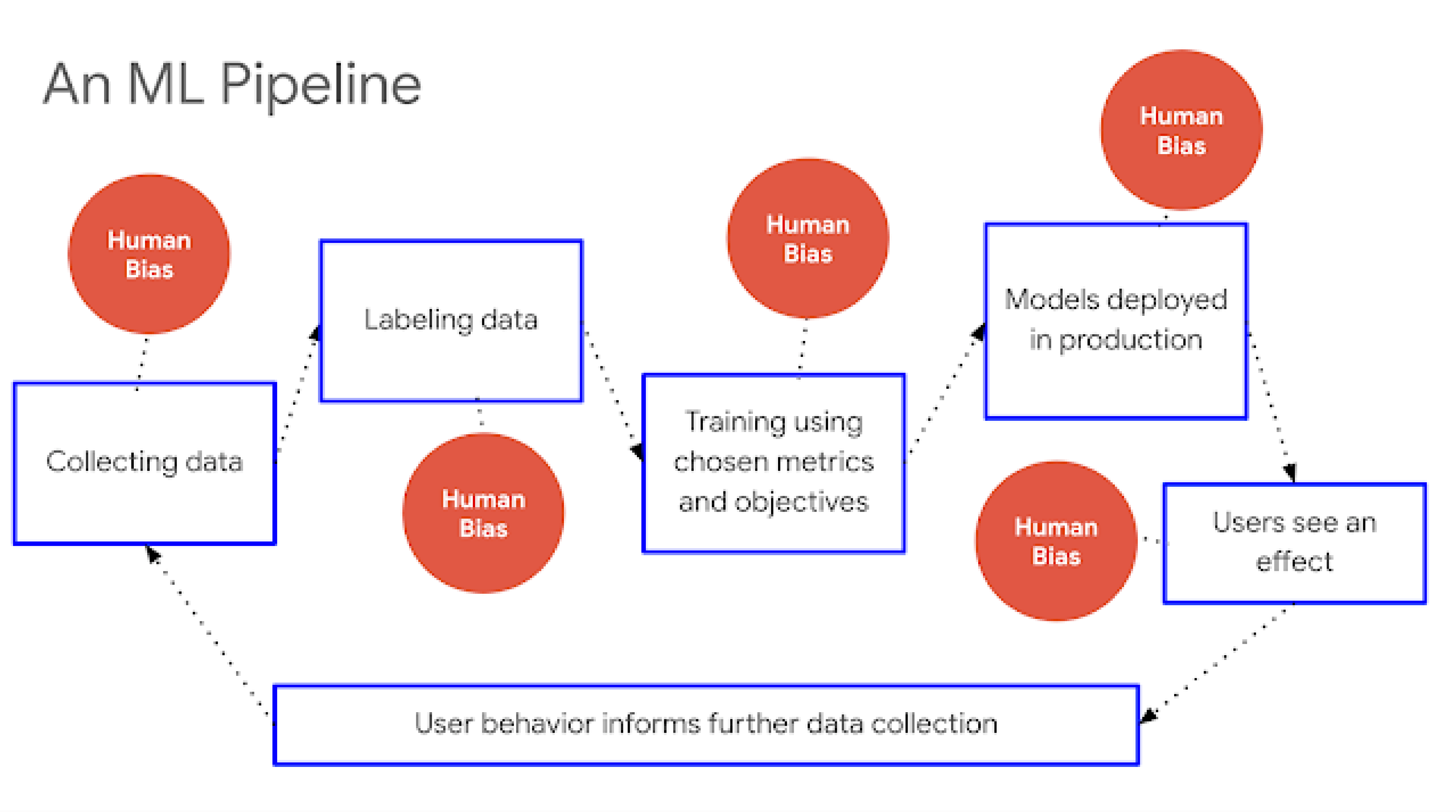


"Despite our aspirations for tech to be better than us, to be more objective than we are, the machines we create are a reflection of both our aspirations and our limitations."

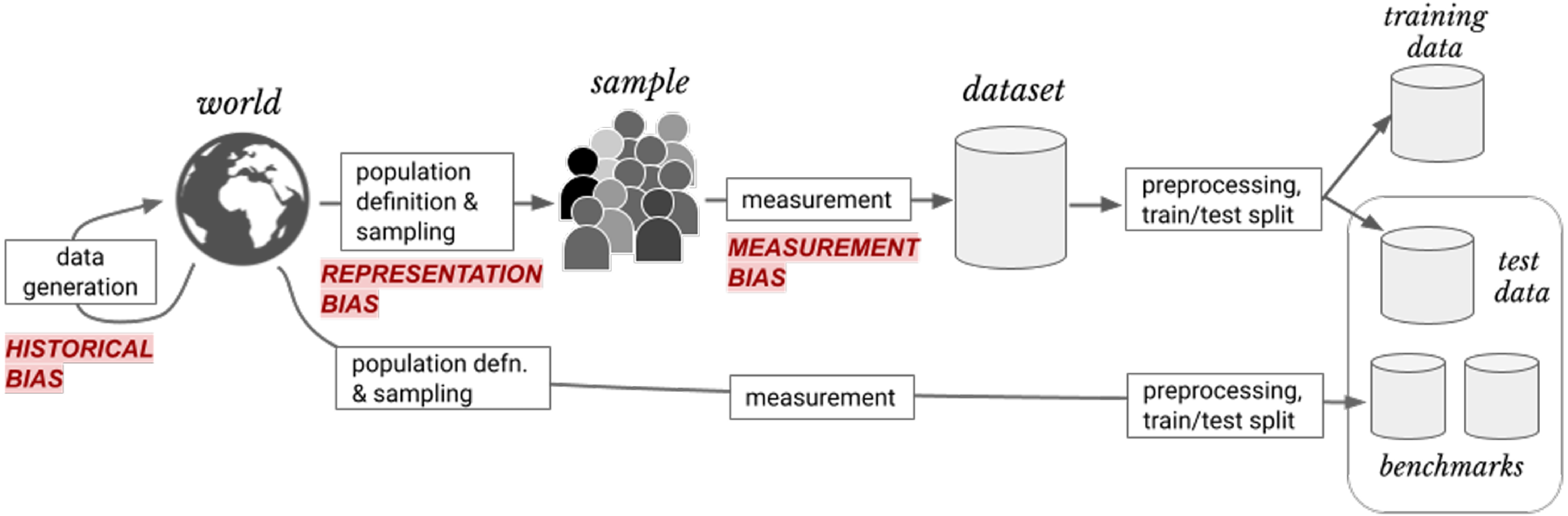
Joy Buolamwini

How bias become part of AI systems?

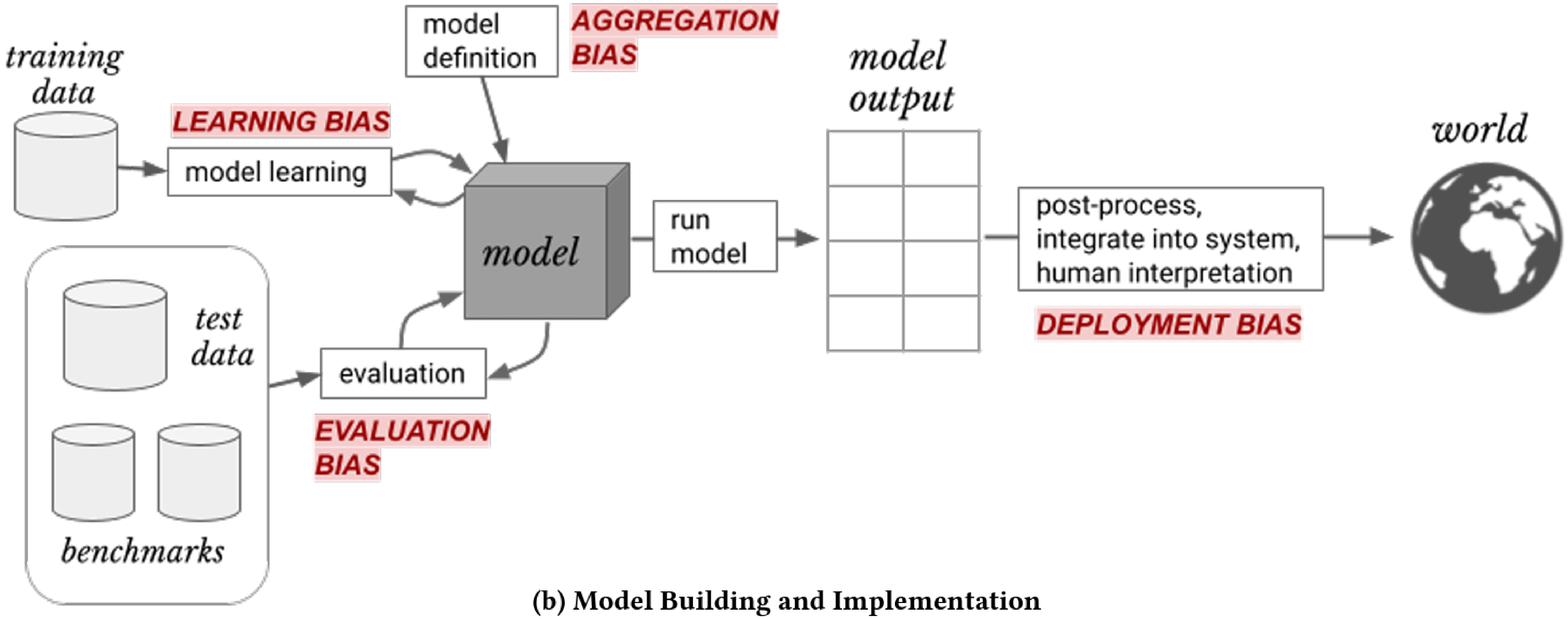
Let's explore how this happens in the **ML Lifecycle**.



Source: https://ai.googleblog.com/2019/12/fairness-indicators-scalable.html



(a) Data Generation

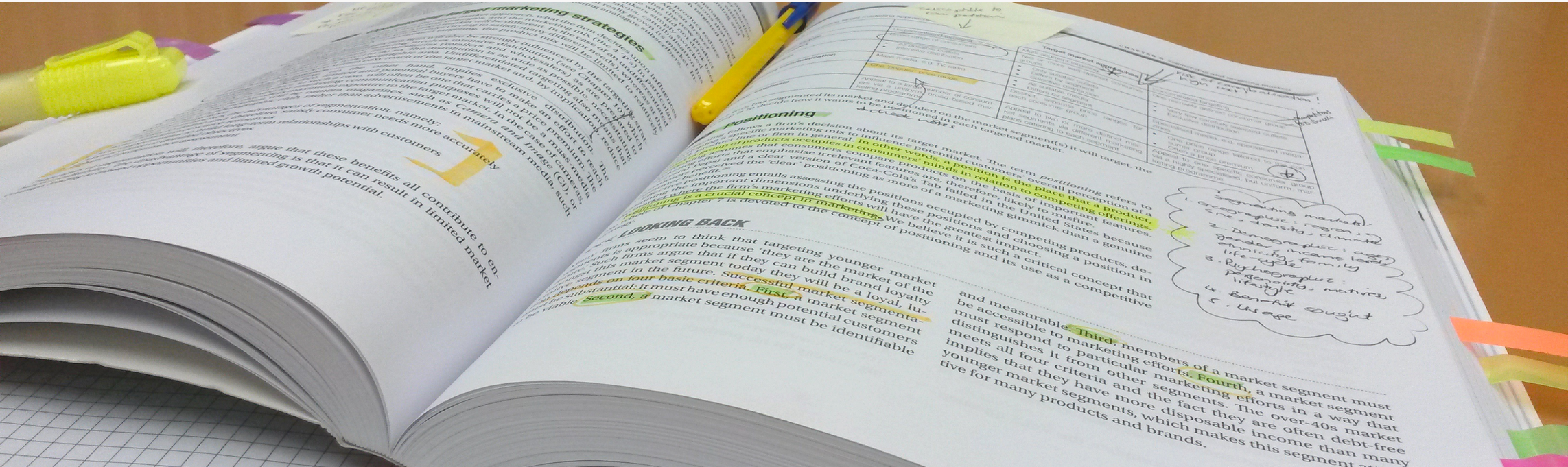


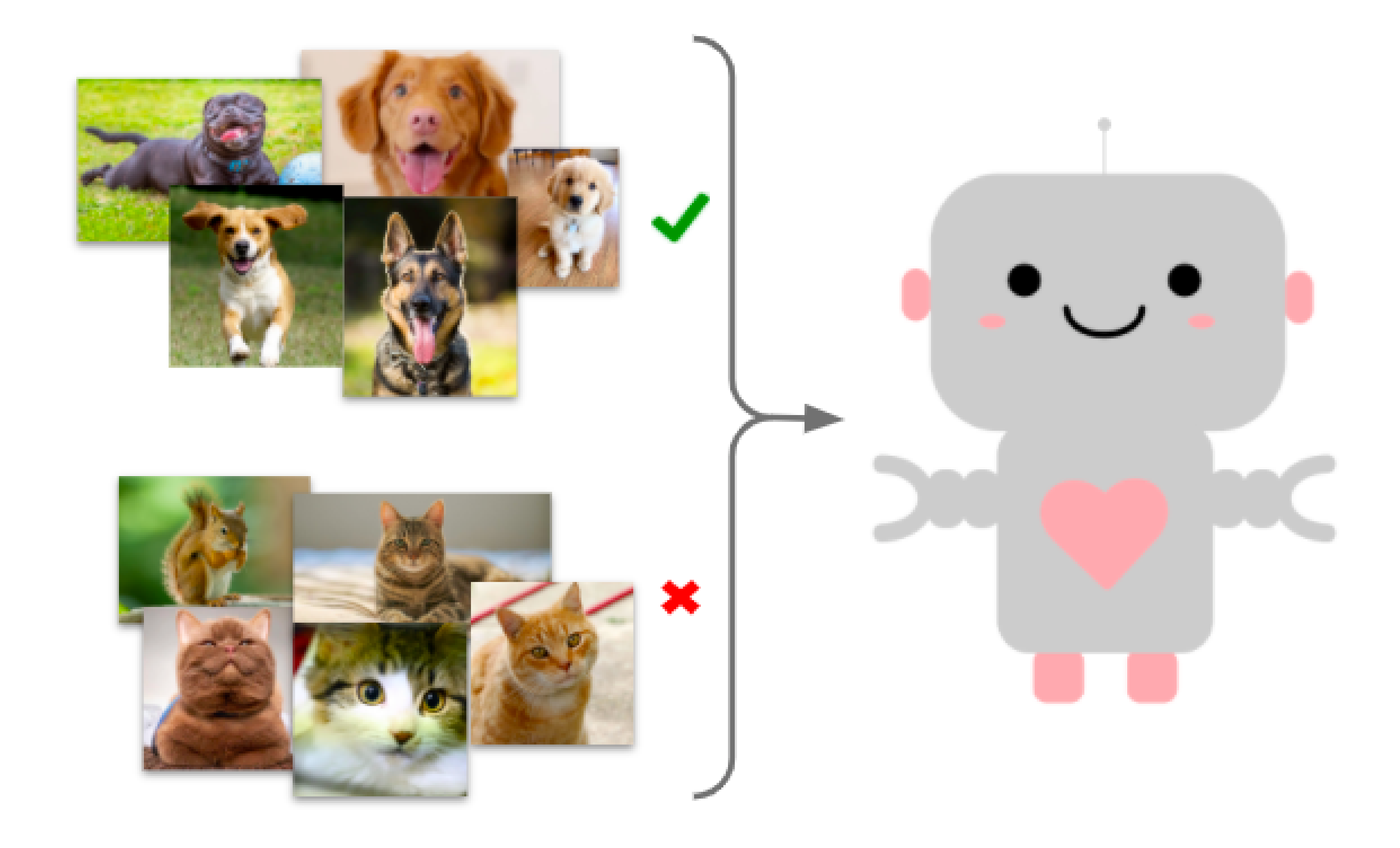
(b) Model Building and Implementation

Source: A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle

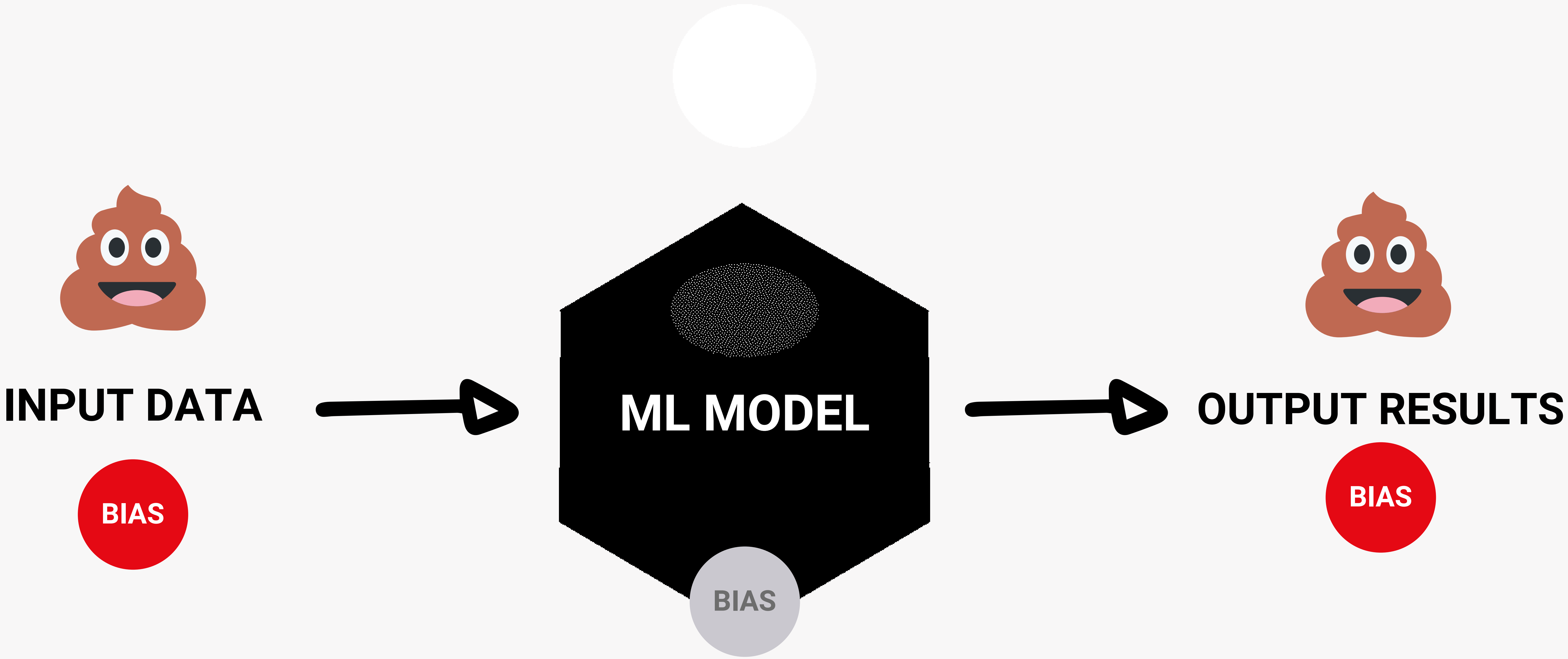
Data generation bias

"**Datasets are like textbooks** for your student to learn from. **Textbooks** have human authors, and so do **datasets**." (Cassie Kozyrkov)



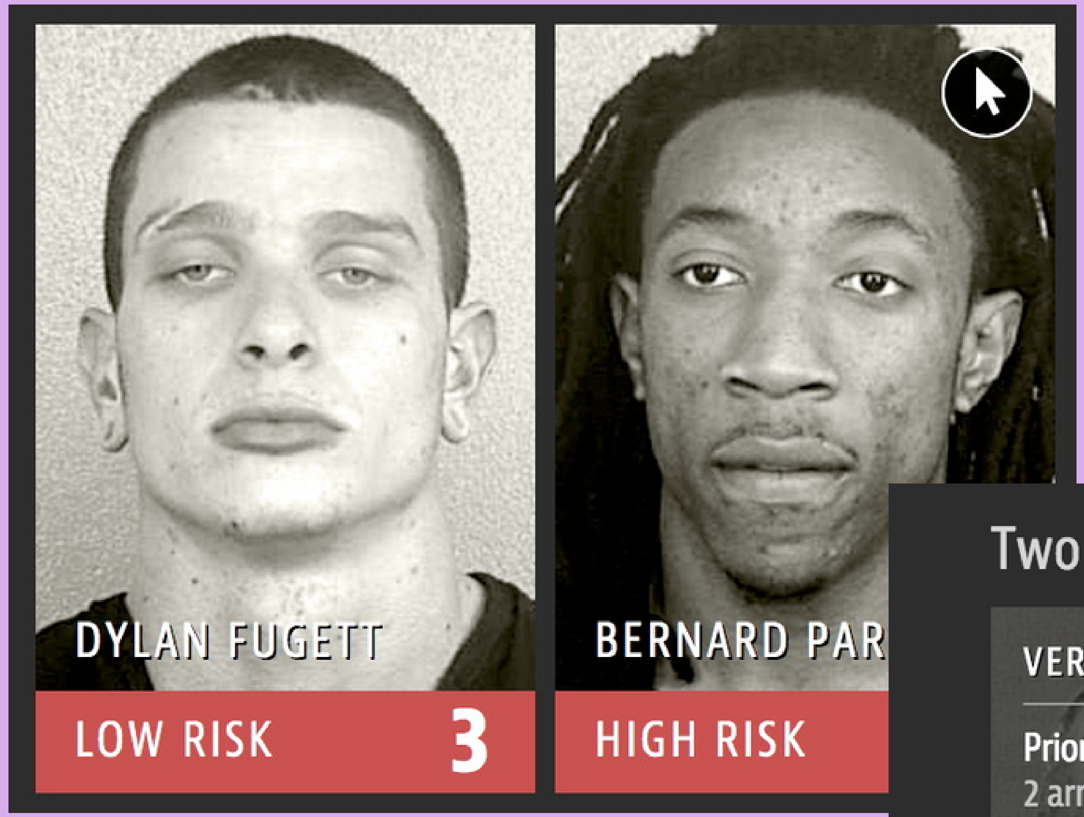


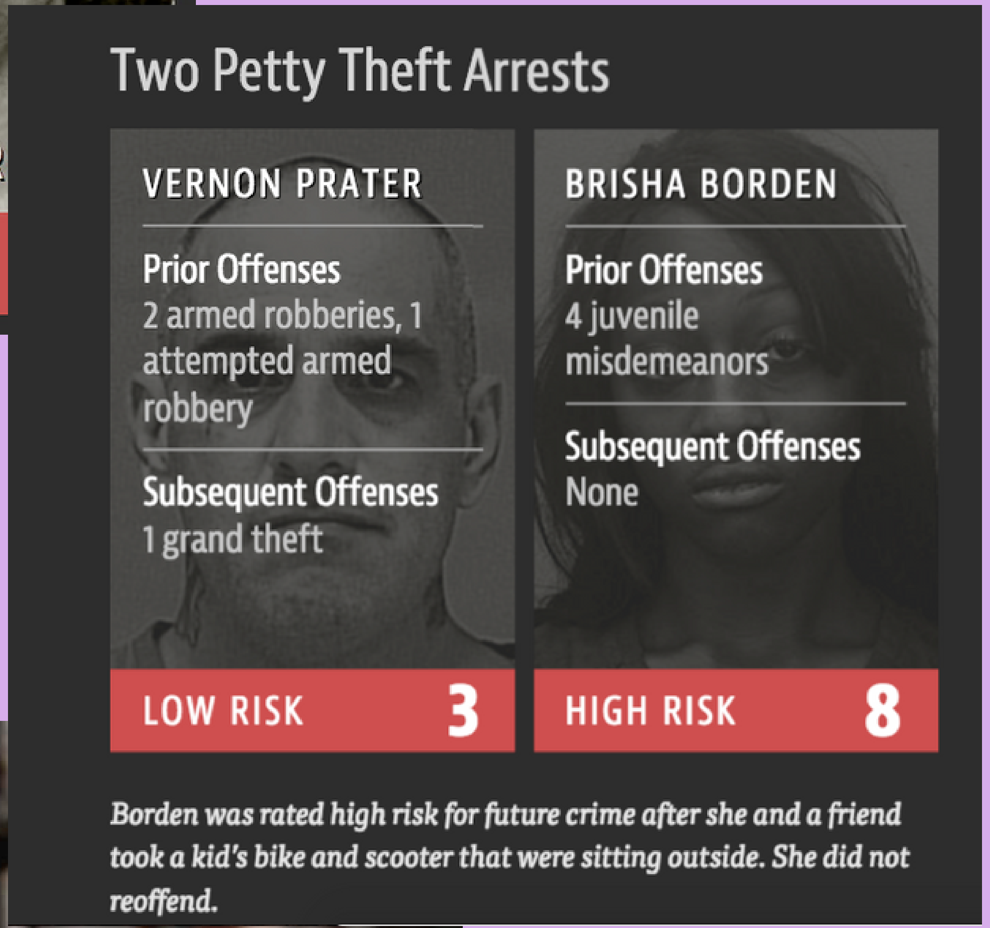
Source: Dogs vs. Not-Dogs: How can a machine learning algorithm learn to tell the difference?

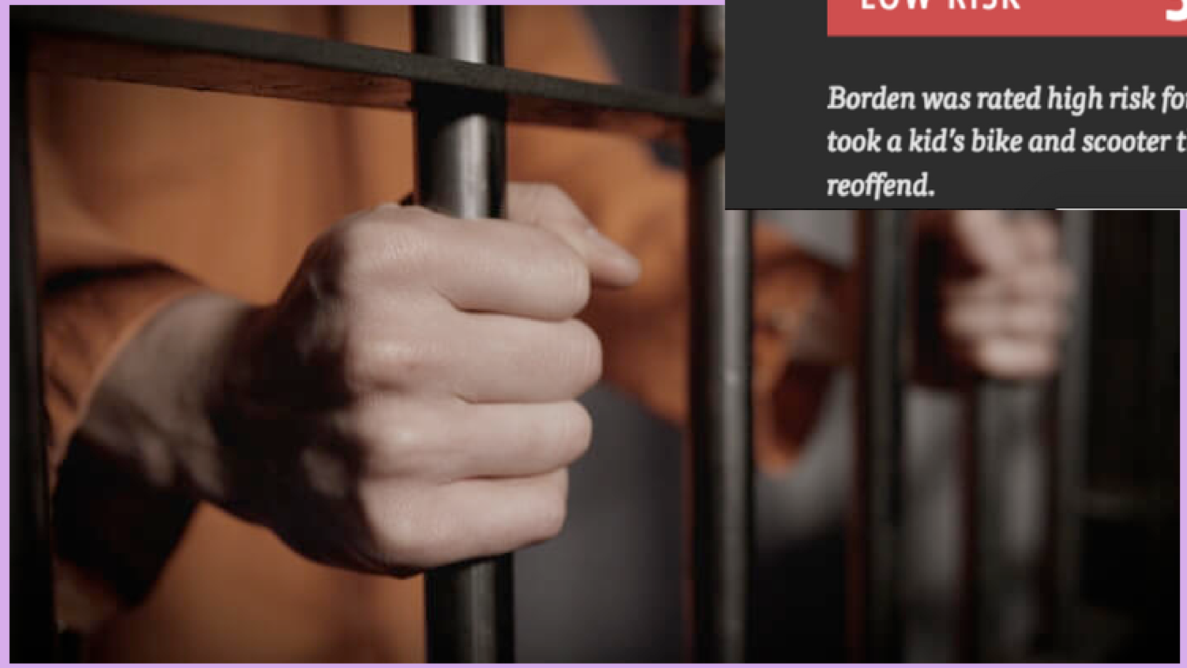

INPUT DATA
BIAS
ML MODEL
BIAS
OUTPUT RESULTS
BIAS

Historical bias

"Historical bias arises even if data is perfectly measured and sampled, if the world as it is or was leads to a model that produces harmful outcomes." (Suresh et. al. 2019)

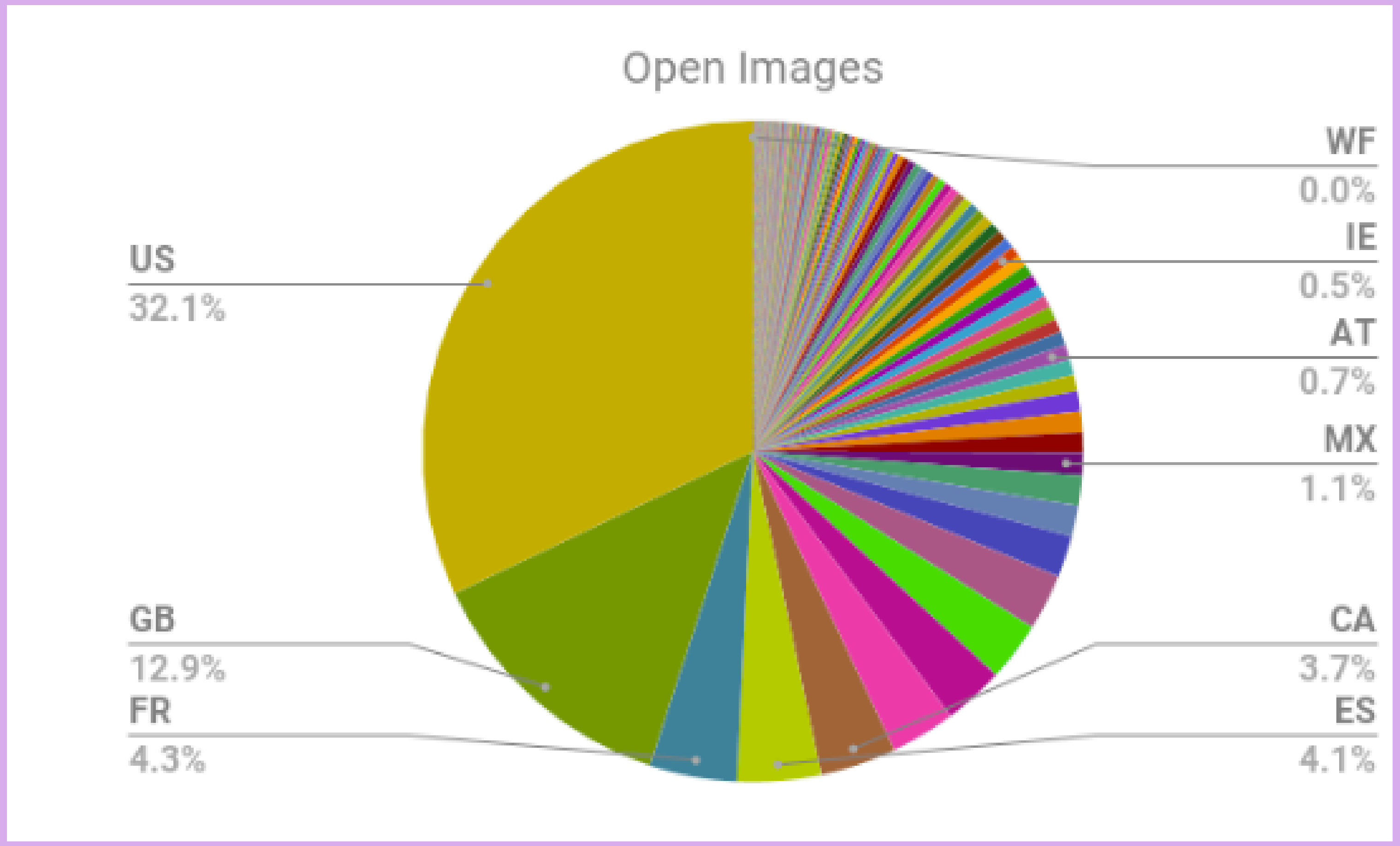


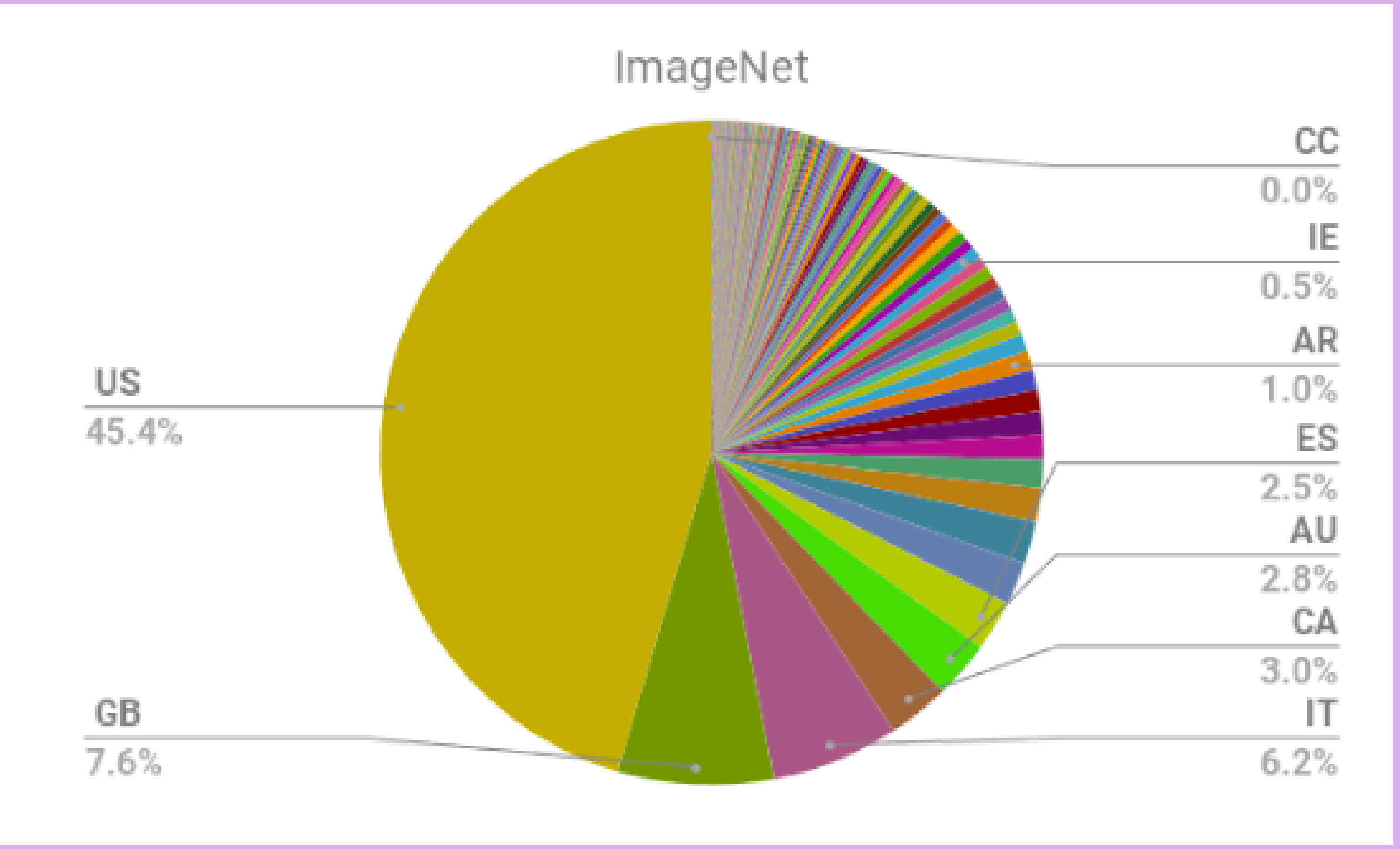




Representation bias

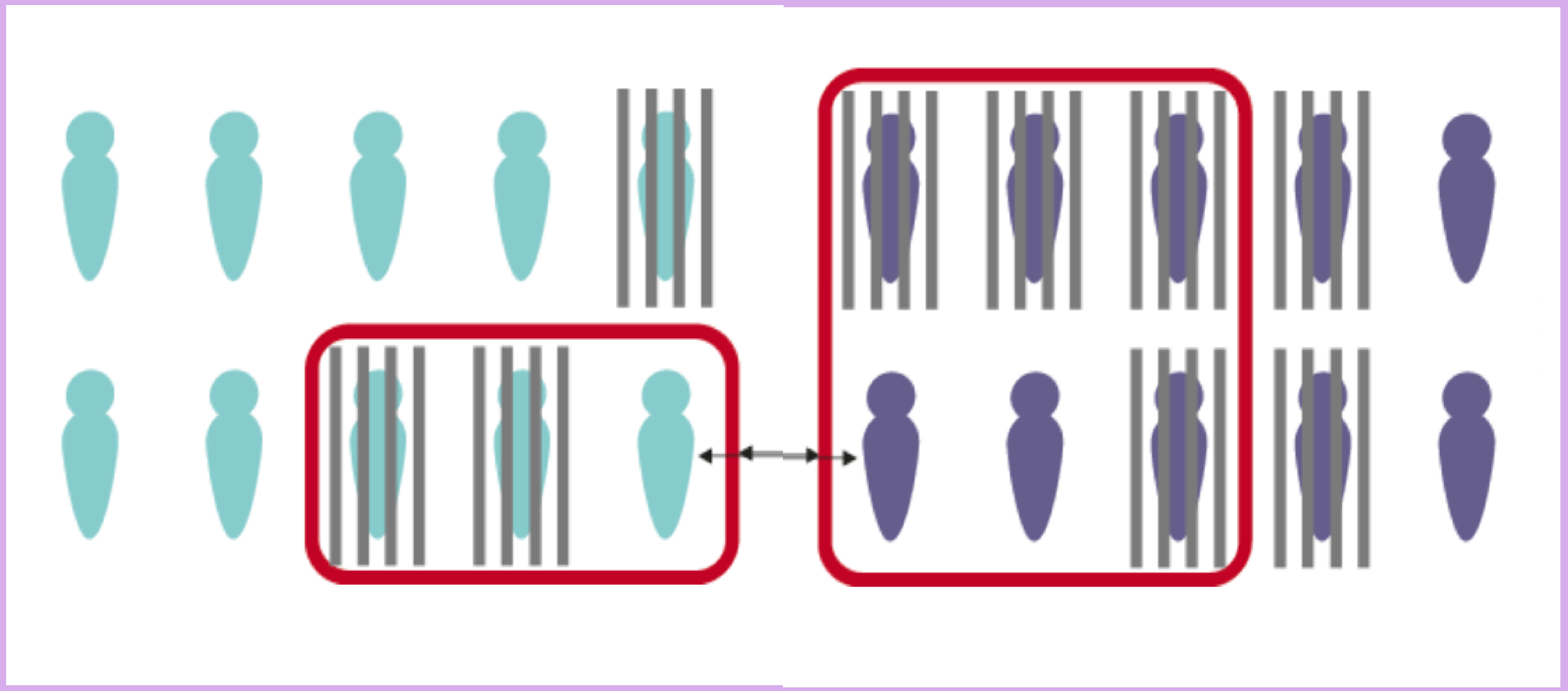
Representation bias occurs when the development sample underrepresents some part of the population.





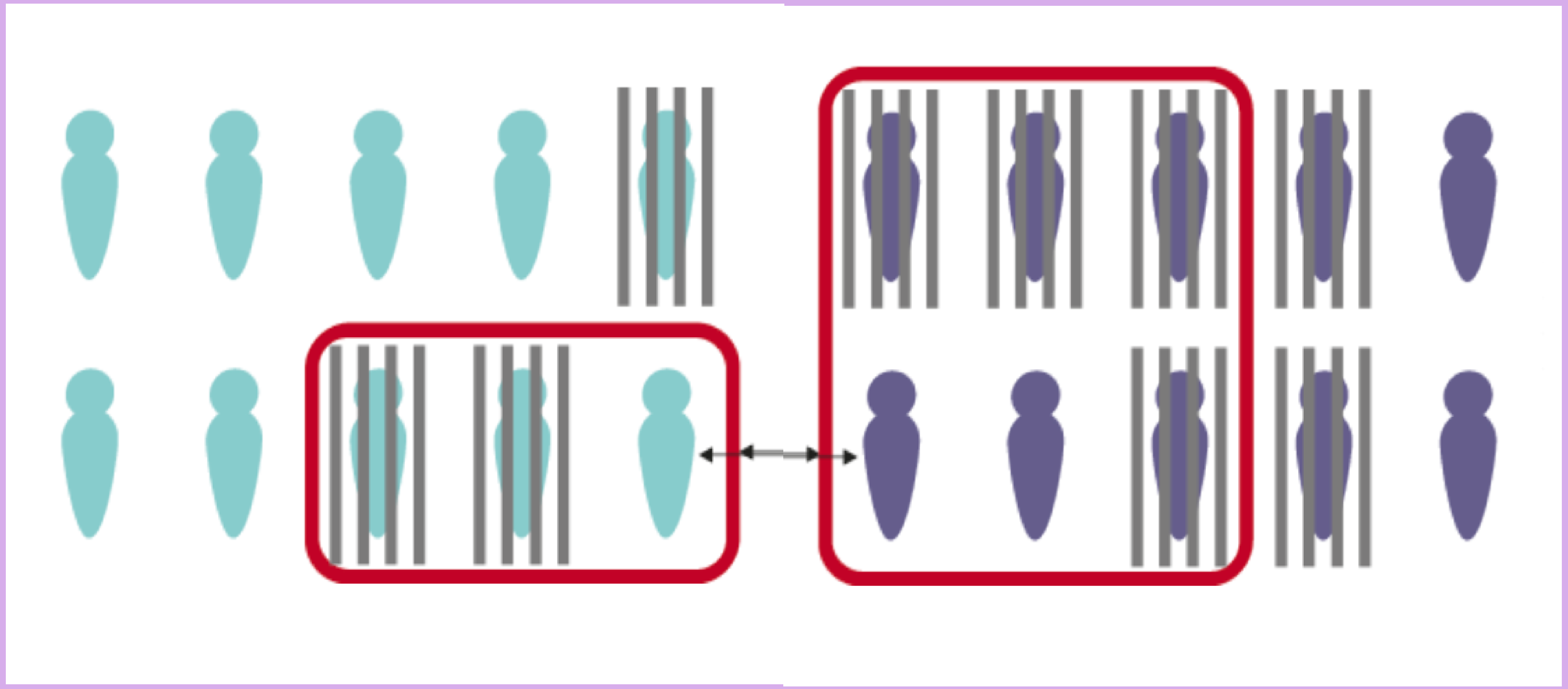
Evaluation bias

"The dominant values in ML are Performance, Generalization, (...) Efficiency, and Novelty. These are often portrayed as innate and purely technical." (Birhane et al., 2021)



Evaluation bias

Recent research has proposed new metrics to evaluate the performance of the model considering notions of bias, fairness and discrimination.

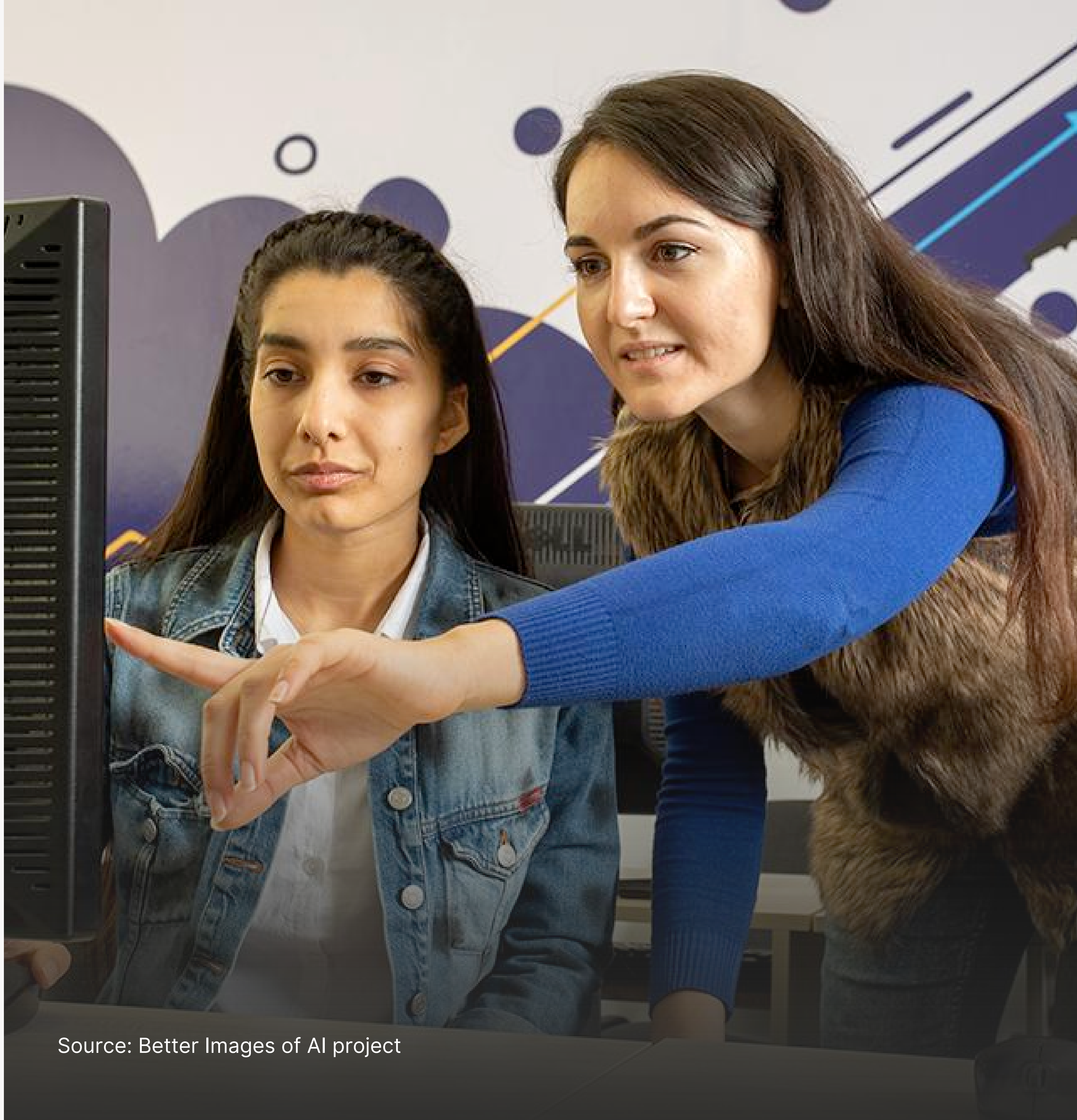


Examples:

- measure the accuracy in the groups separately: a facial recognition model can have an accuracy of 80% on average, but 60% for black women and 90% for white men.
- another way is to assess disproportionate impacts, that is, to assess the balance between false positives for each group;

Deployment Bias

"Deployment bias arises when there is a mismatch between the problem a model is intended to solve and the way in which it is actually used."



Source: Better Images of AI project

Algorithms, the illusion of neutrality

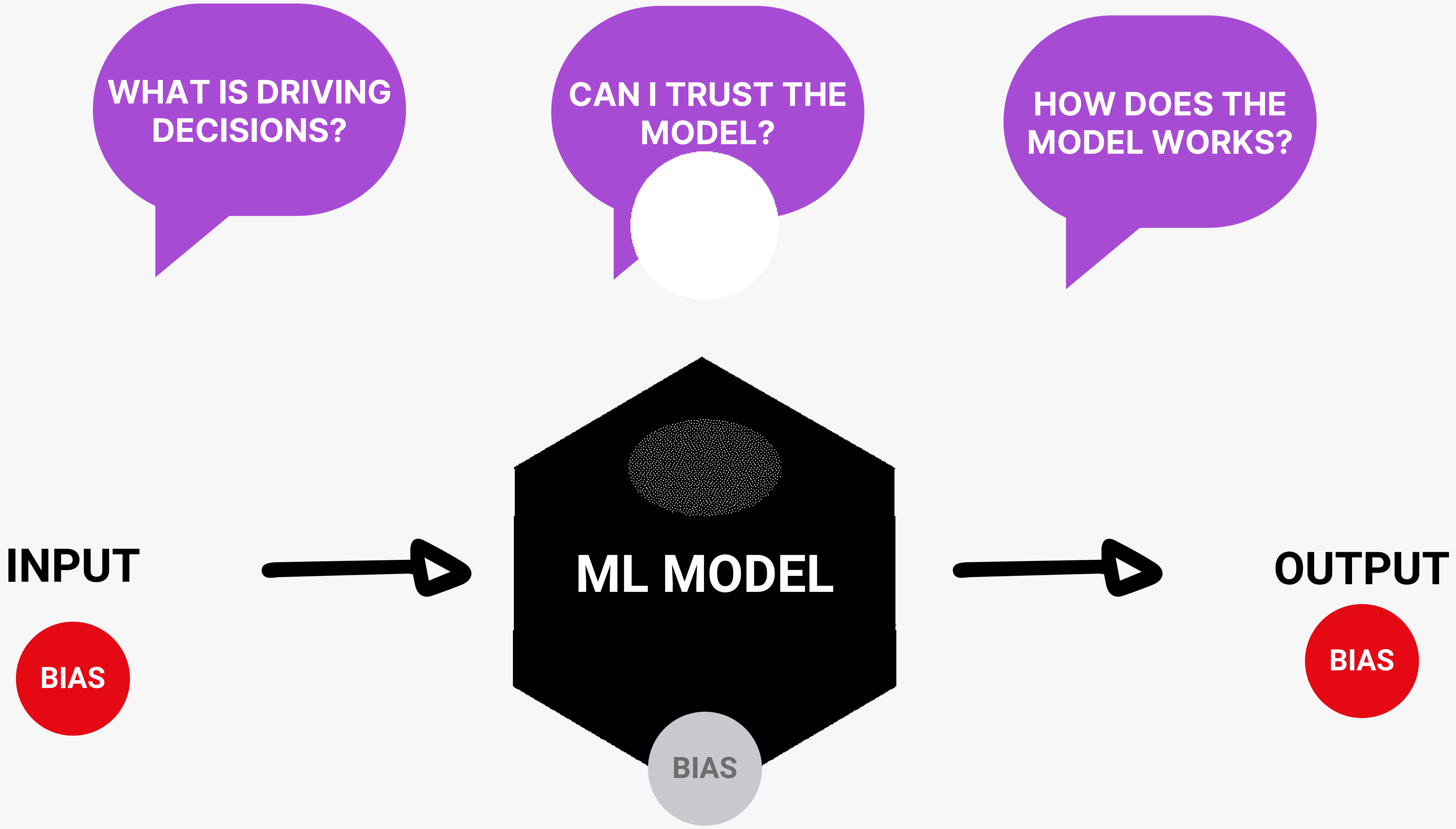
This is called Mathwashing. When power and bias hide behind the facade of "*neutral*" math.

Fred Benenson

Bias doesn't come from AI algorithms, it comes from people.

Black-box problem

The current generation of AI Systems are what we call **black-boxes.**


WHAT IS DRIVING DECISIONS?
CAN I TRUST THE MODEL?
HOW DOES THE MODEL WORKS?
INPUT
BIAS
ML MODEL
BIAS
OUTPUT
BIAS

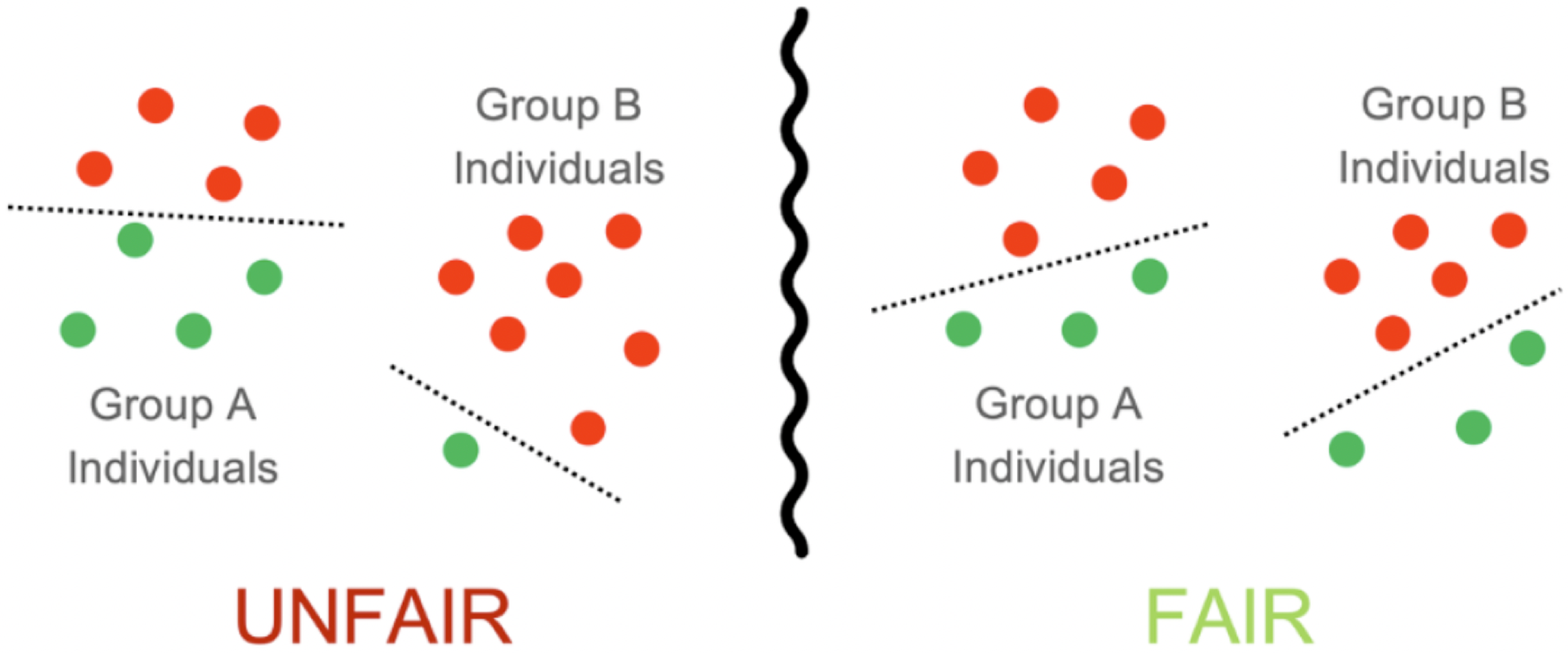
What can we do to solve this?

Machine intelligence makes human morals more important.

"We cannot outsource our responsibilities to machines."
(Zeynep Tufekci)

Fairness

“An algorithm is fair if it makes predictions that do not favour or discriminate against certain individuals or groups based on sensitive characteristics.”

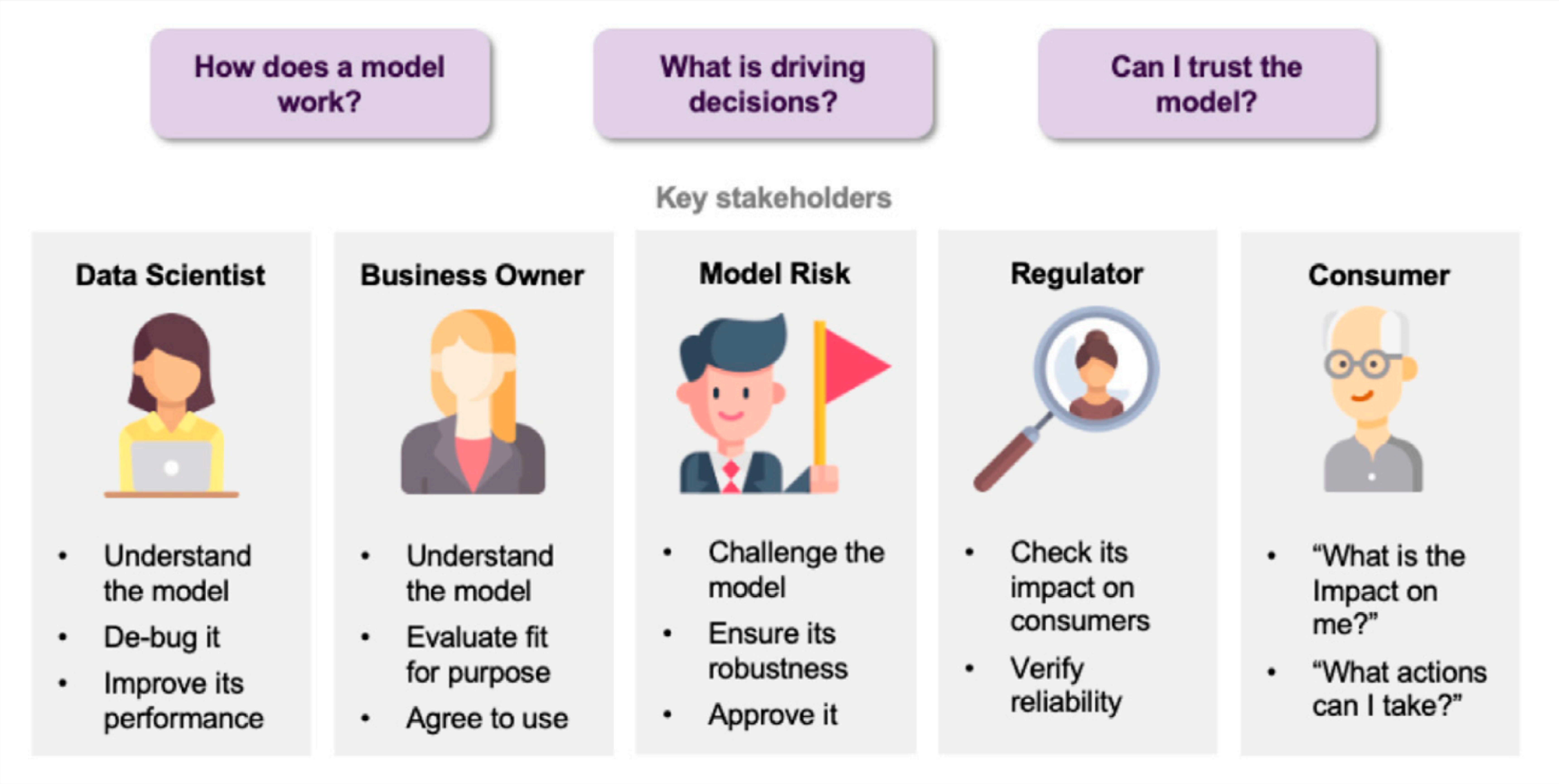


Algorithmic fairness is a topic of great importance, with impact on many applications. The issue requires much further research; even the definition of what "being fair" means for an ML model is still an open research question.

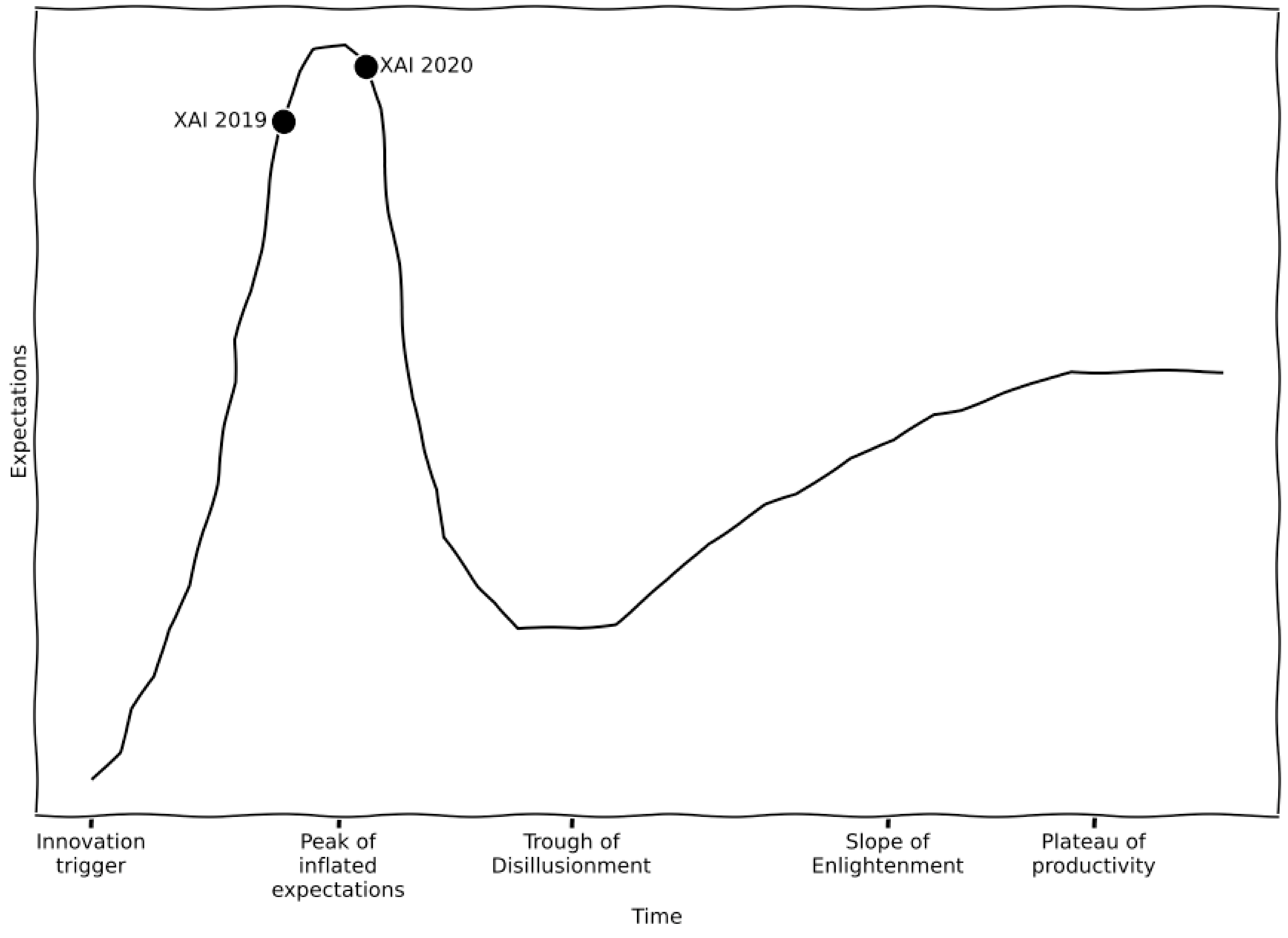
Source: https://www.amazon.science/research-awards/success-stories/algorithmic-bias-and-fairness-in-machine-learning

Explainable and Interpretable AI

Explainability is not a new issue for AI systems. But it has grown along with the success and adoption of deep learning.



Source: Principles and Practice of Explainable Machine Learning (Vaishak and Ioannis, 2019)


XAI 2020
XAI 2019
Expectations
Innovation trigger
Peak of inflated expectations
Trough of Disillusionment
Slope of Enlightenment
Plateau of productivity
Time

Challenges XAI

- Lack of **global explanation** methods
- How to avoid **ground truth unjustification**?
- How can we **better evaluate** explanations?
- Can we do better explanations for **non-expert users**?
- How does fairness interact with interpretability?
- How can we build more **robust** interpretability methods?
- **How to combine and deploy interpretable Machine Learning models?**

Product Thinking approach

Thinking of AI as a product...

Who is your invention for? Who benefits from it?

This is a great time to consult with a UX (user experience) specialist and map out your application’s users.

Is it ethical to proceed?

Just because you can do something, doesn't mean you should.

Think about the humans your creation impacts!

Who benefits and who might be harmed?


Dataset 1
Dataset 2
All swans are white

Diversity of perspective matters!

Applied data science is a team sport that’s highly interdisciplinary

Summary

01 TECHNOLOGY IS NOT FREE OF HUMANS

---

02 MATH CAN OBSCURE THE HUMAN ELEMENT AND GIVE AN ILLUSION OF OBJECTIVITY.

---

03 EVERY SINGLE HUMAN IS BIASED.

Thank you!

@carlaprvieira
carlavieira.dev