LEADDEV NEW YORK 2025

TALK

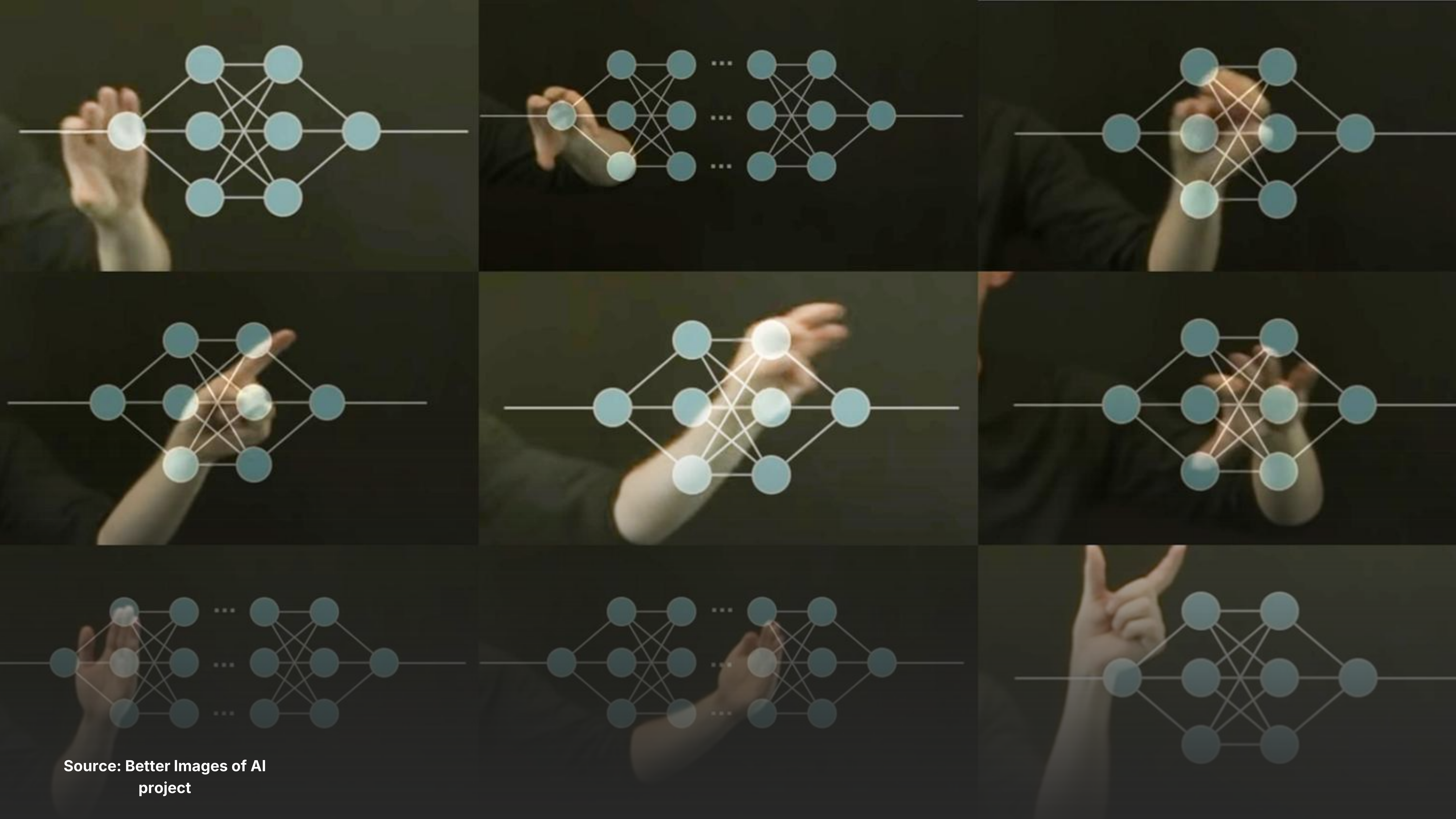
Building ethical AI: Transparency, bias mitigation, and trust

CARLA VIEIRA

Get to Know Me

I'm Carla Vieira, Senior Data Engineer and Google Developer Expert in Machine Learning. I hold a master's degree in Artificial Intelligence, with research focused on building trustworthy and explainable AI systems. Listed as a Rising Star in Women in AI Ethics.

@carlaprvieira / carlavieira.dev



Source: Better Images of AI project

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.

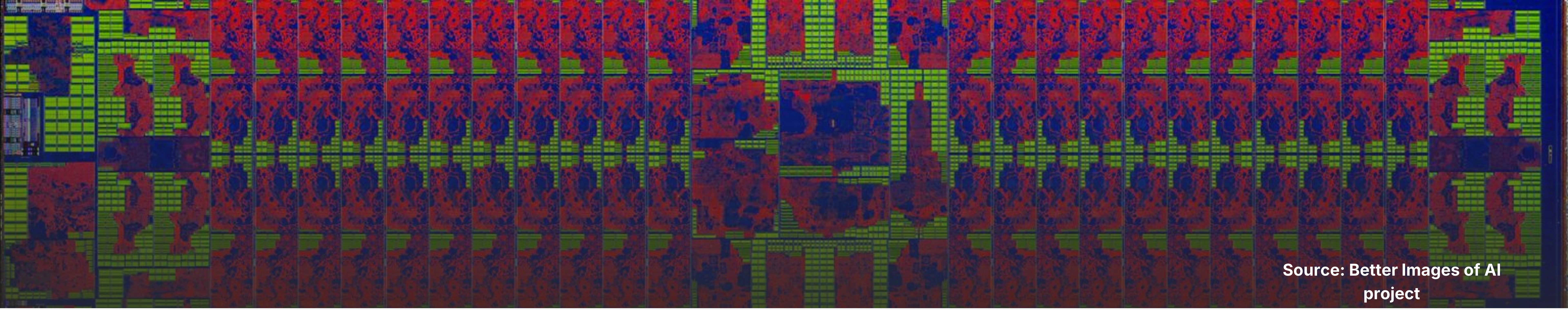
01 BIAS AND DISCRIMINATION

02 DENIAL OF INDIVIDUAL AUTONOMY AND RIGHTS

03 NON-TRANSPARENT, UNEXPLAINABLE, OR UNJUSTIFIABLE OUTCOMES

04 INVASIONS OF PRIVACY

05 UNRELIABLE, UNSAFE, OR POOR-QUALITY OUTCOMES



Source: Better Images of AI project

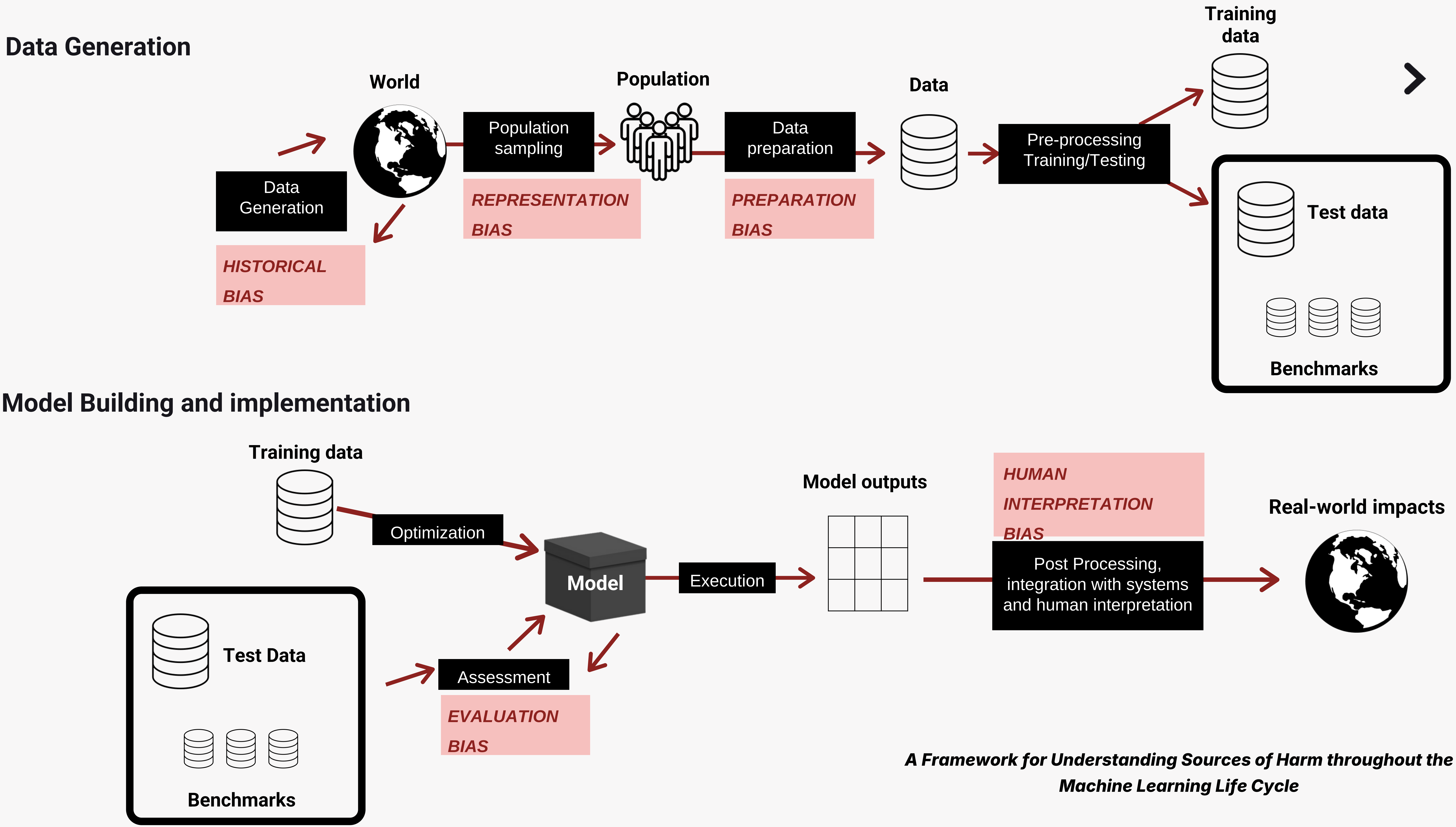
What is bias in ML/AI?

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

How bias become part of AI systems?

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

Source: Better Images of AI project

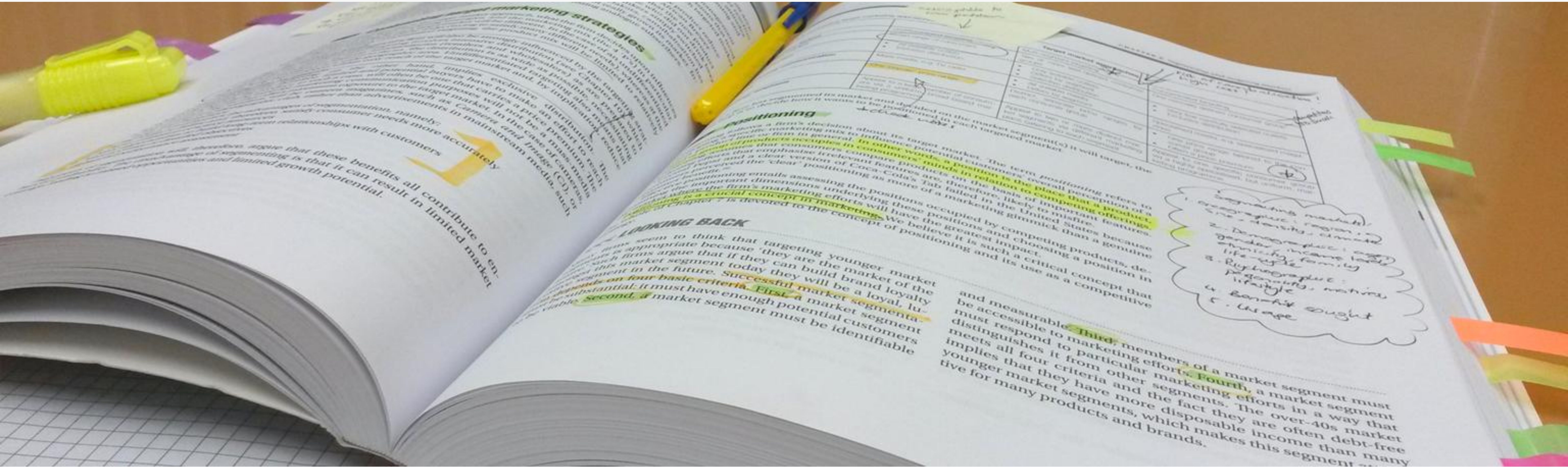


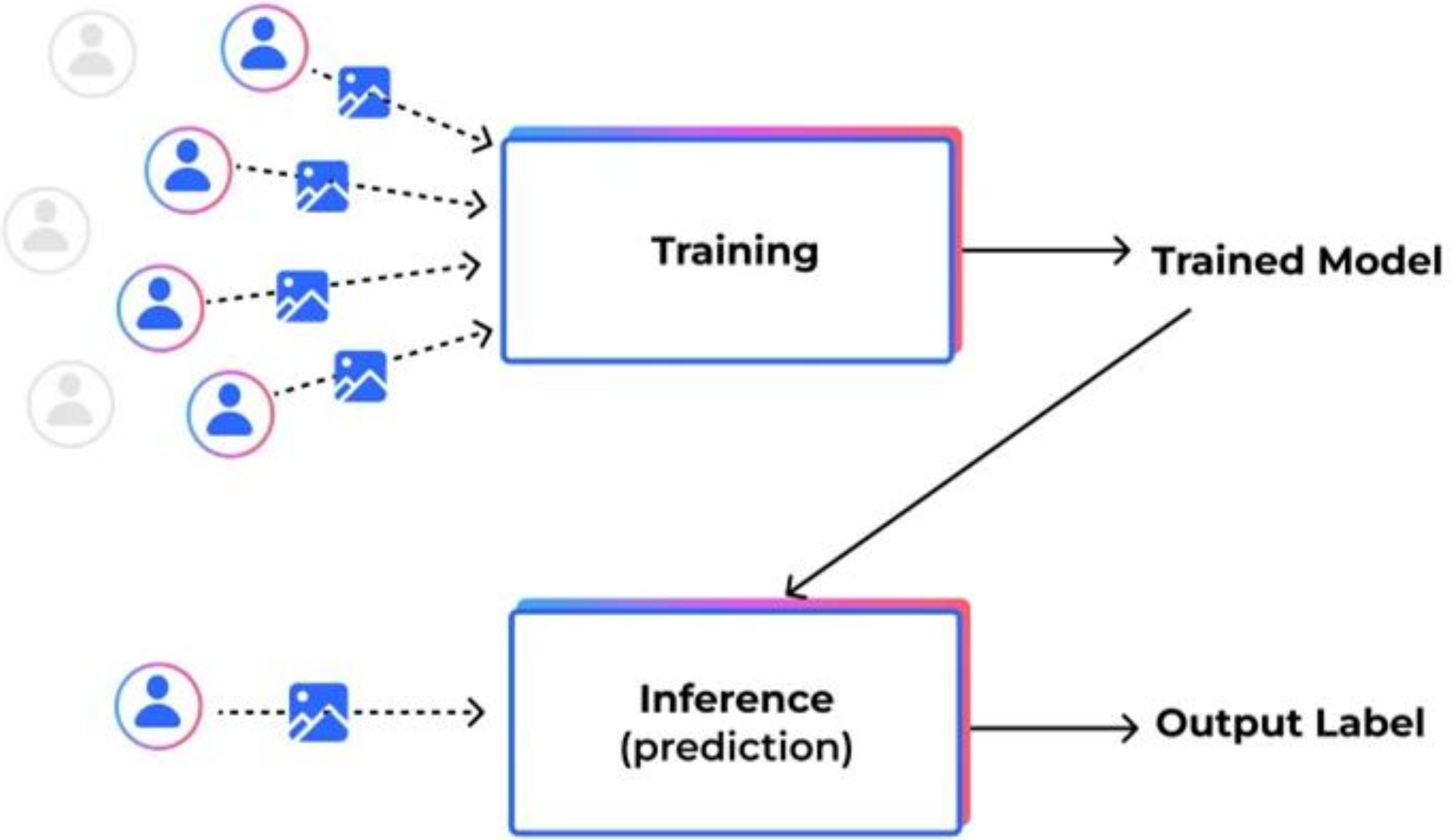
A Framework for Understanding Sources of Harm throughout the Machine Learning Life Cycle

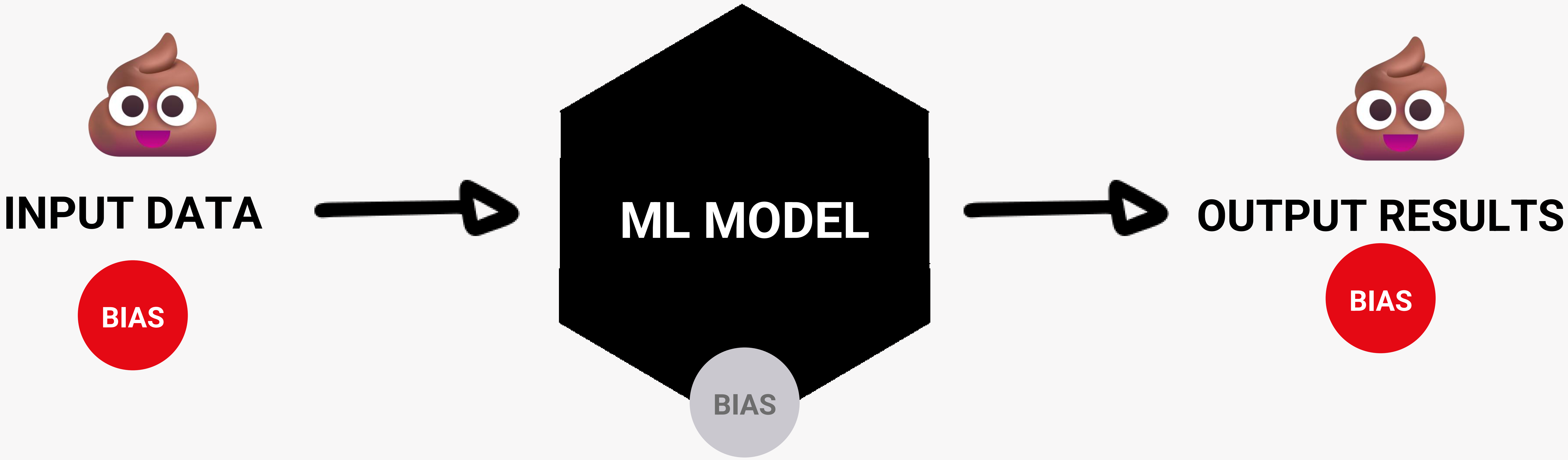
Data generation bias

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

(Cassie Kozyrkov)

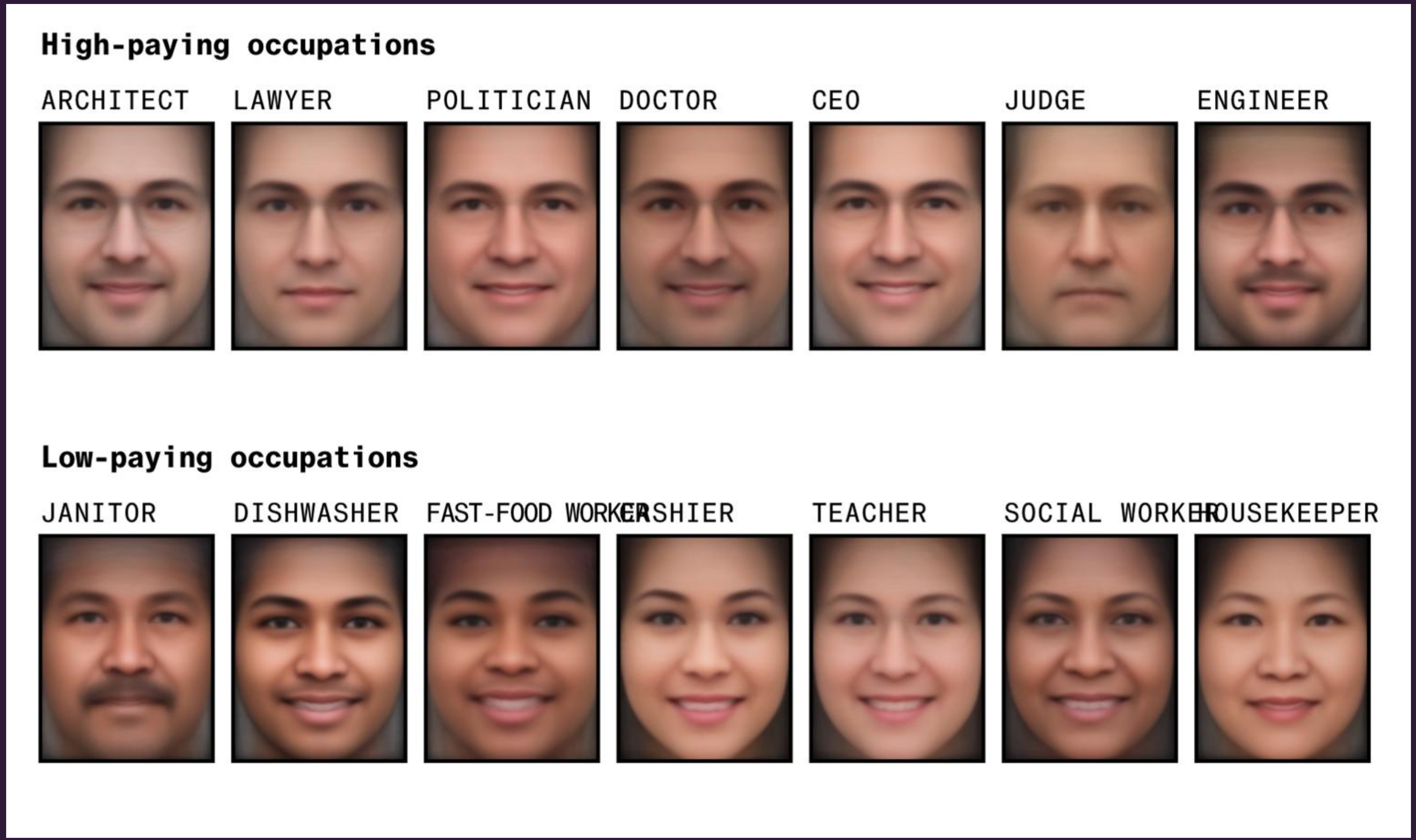



Training
Trained Model
Inference
(prediction)
Output Label


INPUT DATA
BIAS
ML MODEL
BIAS
OUTPUT RESULTS
BIAS

Historical bias

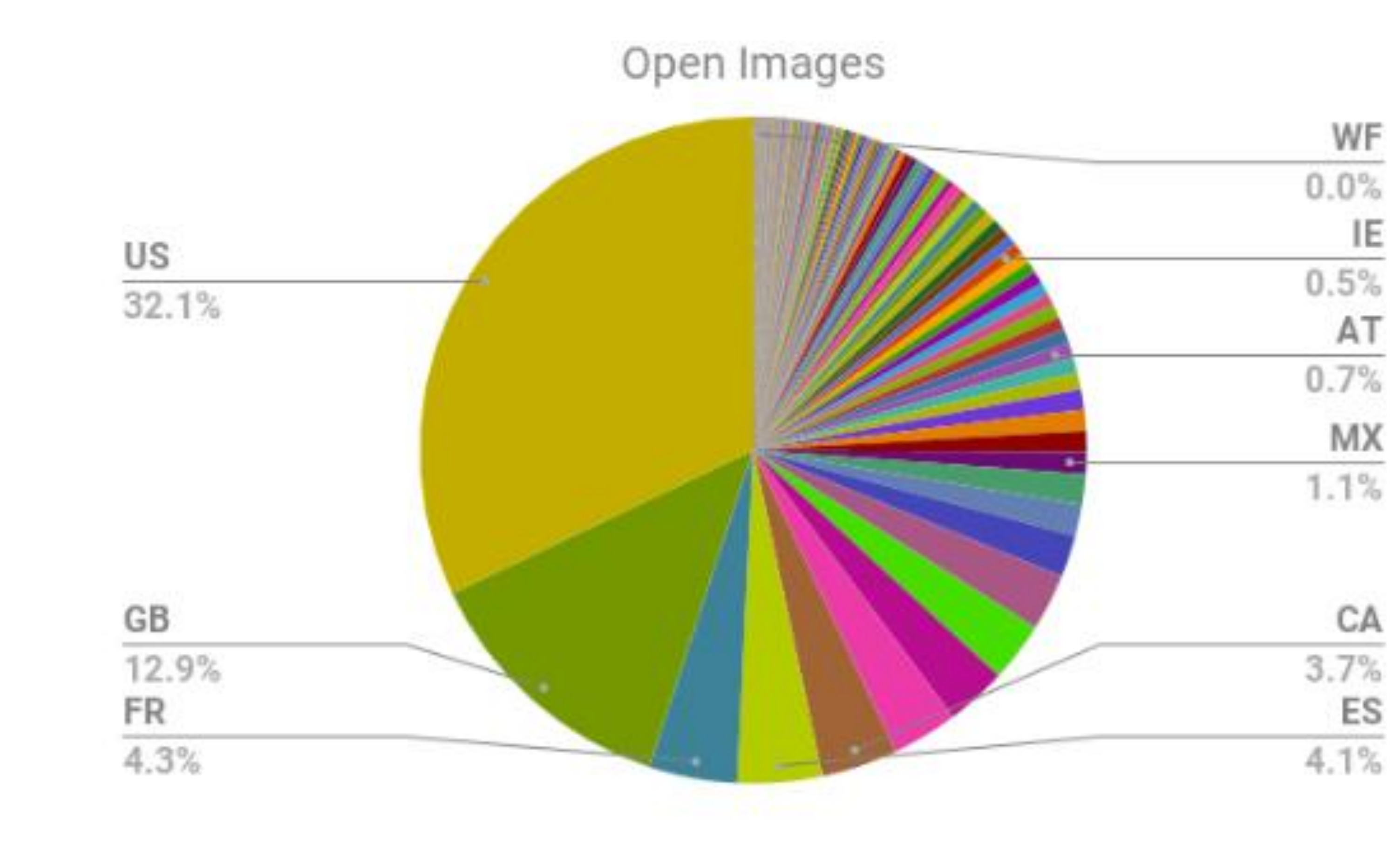
"Historical bias arises even if data is perfectly measured and sampled, if the world as it is or was leads to a model that produces harmful outcomes." (Suresh et. al. 2019)

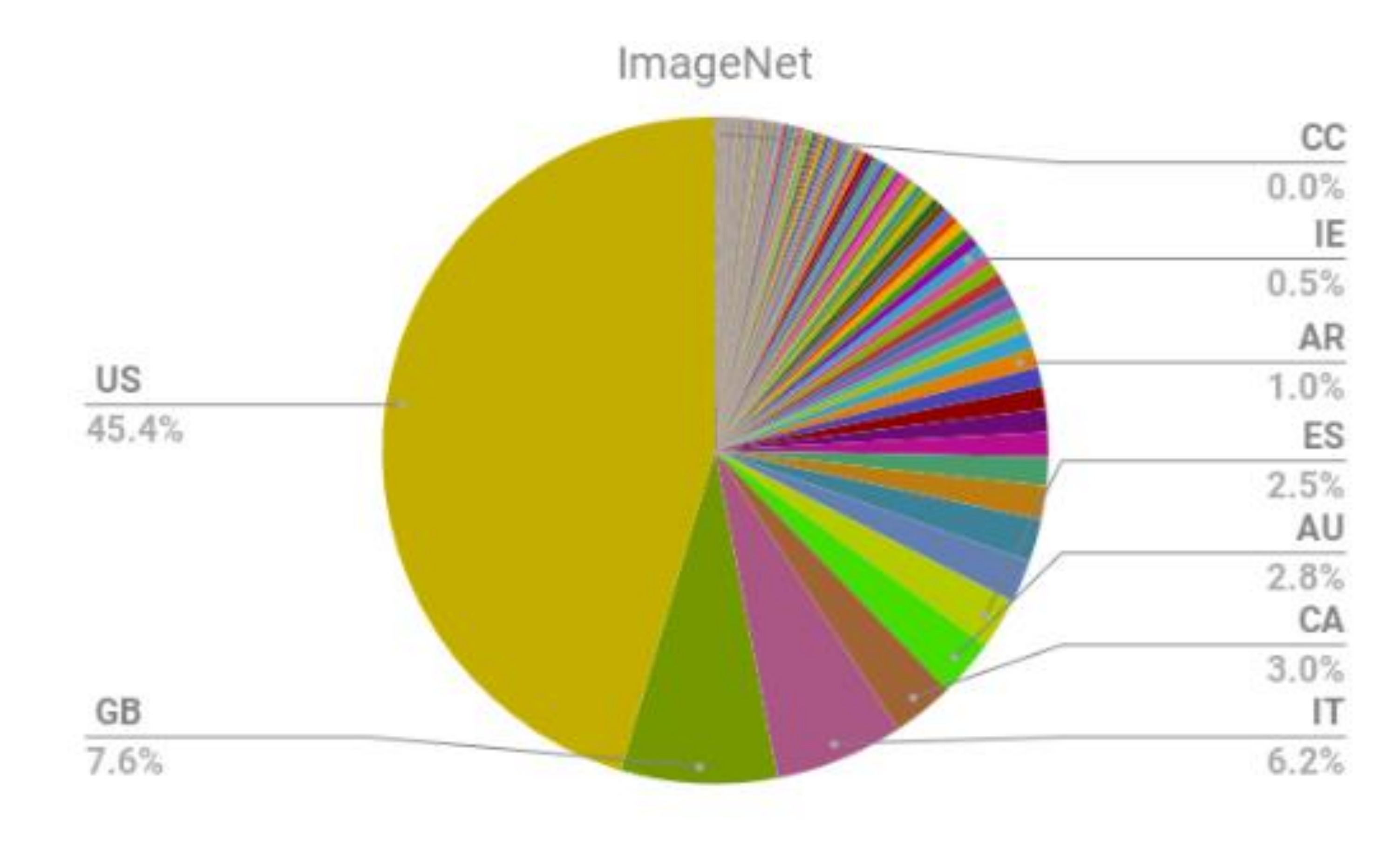


Source: Humans Are Biased. Generative AI Is Even Worse

Representation bias

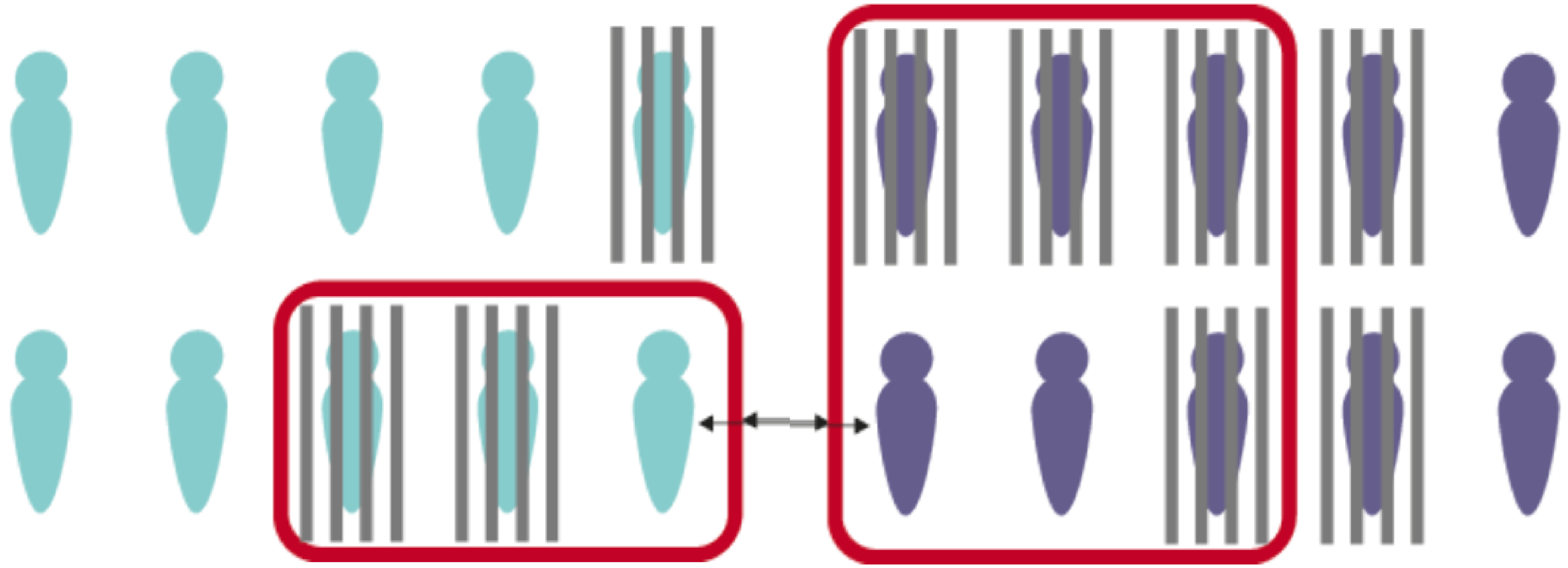
Representation bias occurs when the development sample underrepresents some part of the population.





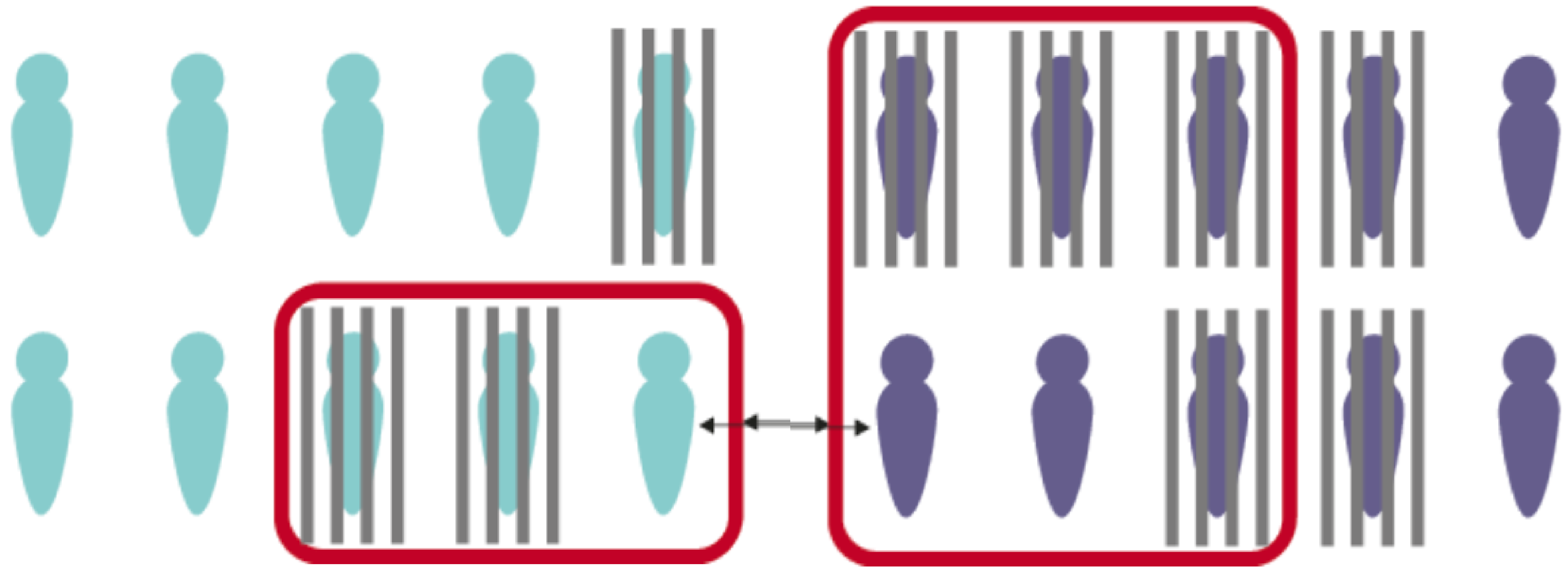
Evaluation bias

"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:

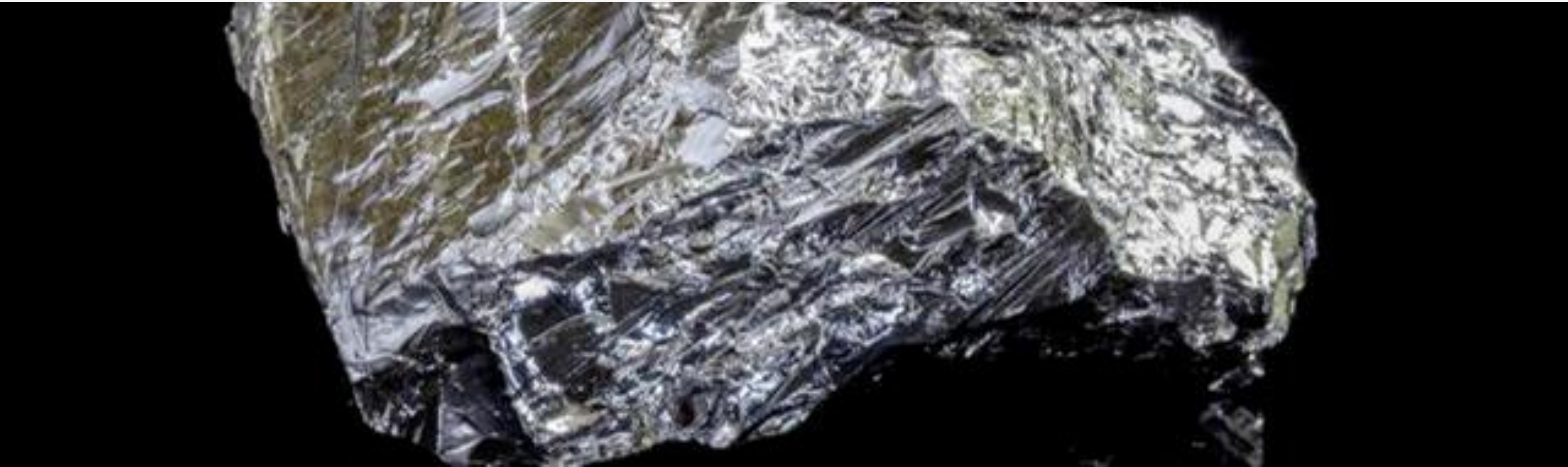
- measure the accuracy in the groups separately: a facial recognition model can have an accuracy of 80% on average, but 60% for black women and 90% for white men.
- another way is to assess disproportionate impacts, that is, to assess the balance between false positives for each group;

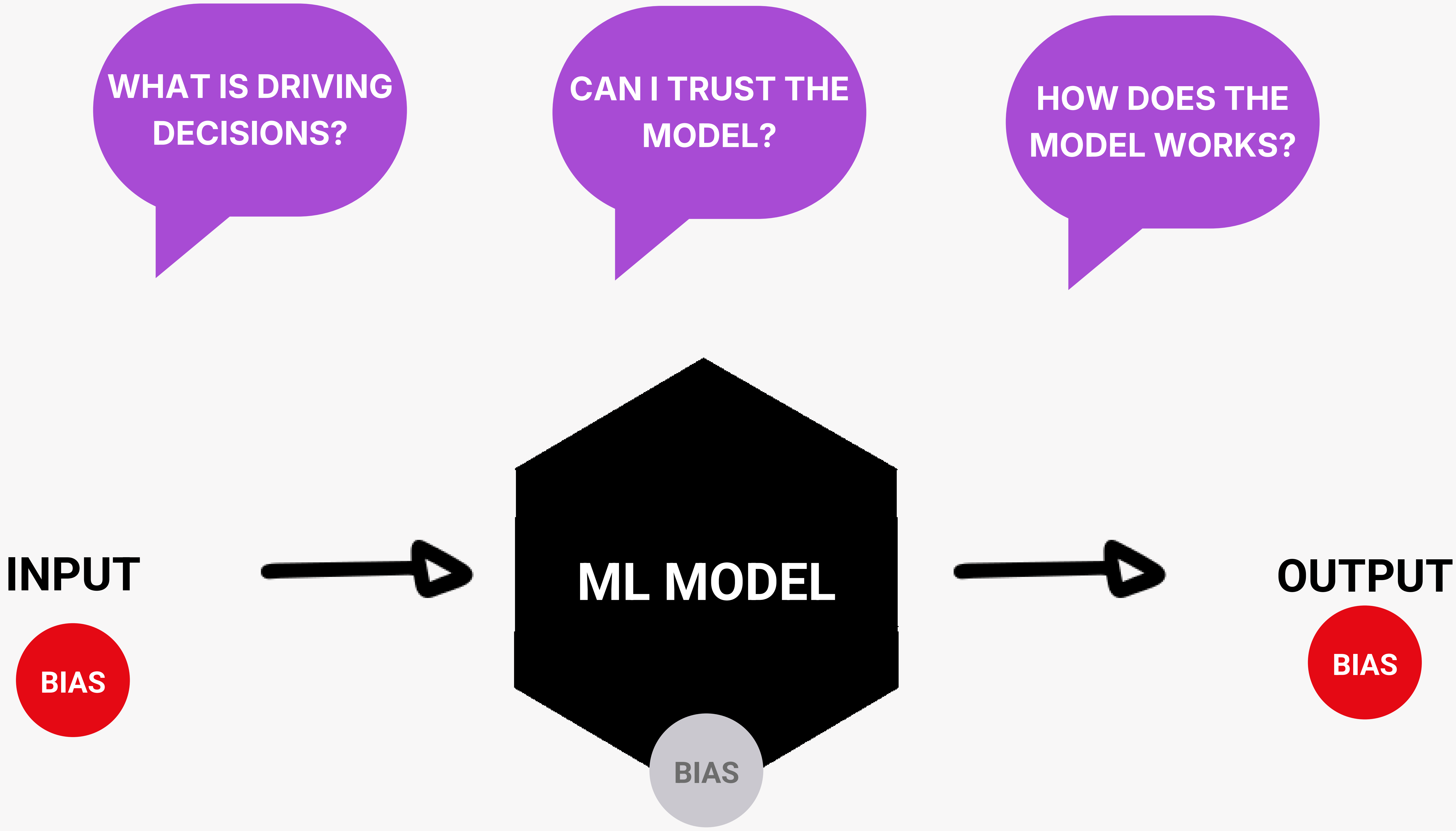
Source: Better Images of AI project

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?

Machine intelligence makes human morals more important.

"We cannot outsource our responsibilities to machines."

(Zeynep Tufekci)

Product *Thinking Human-centered Approach*

Thinking of AI as a product...

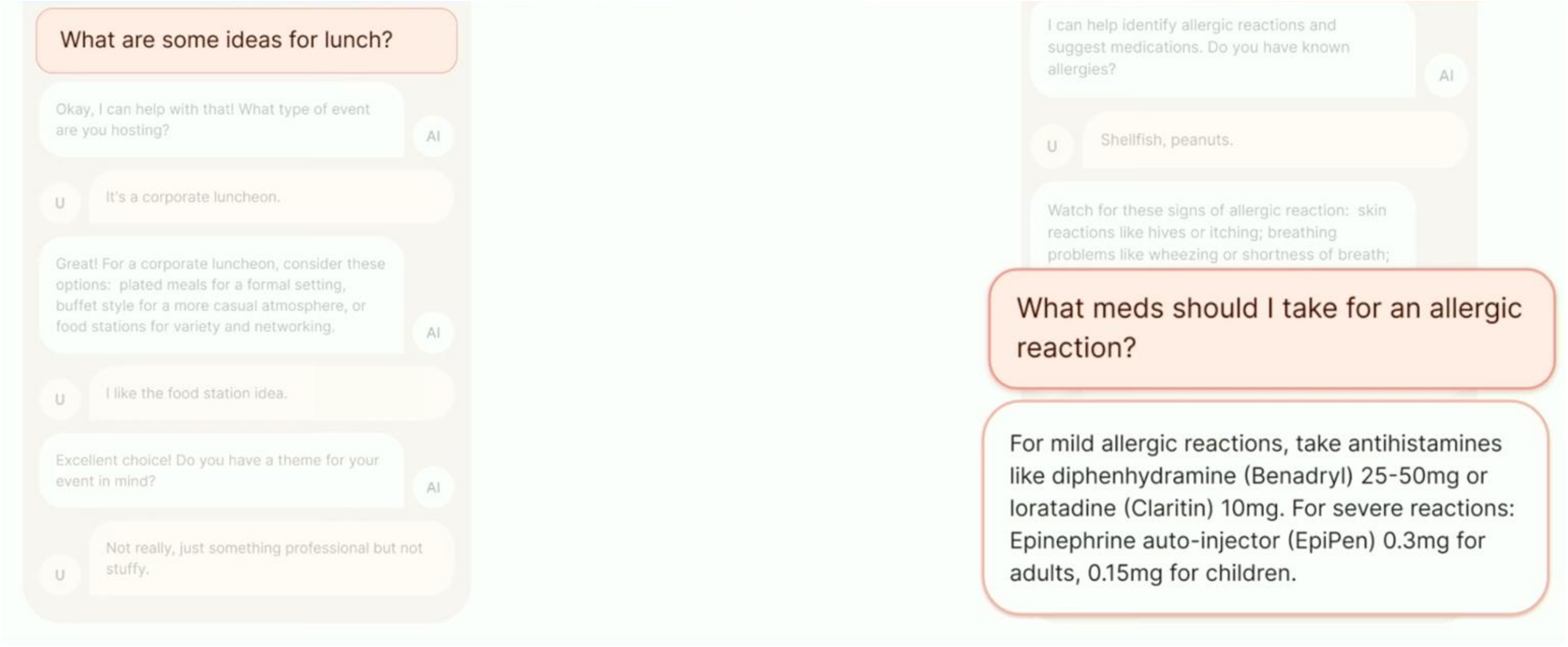
Balance people-first and technology-first thinking

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.

Lower risk

Higher risk



Is it ethical to proceed? 🤔

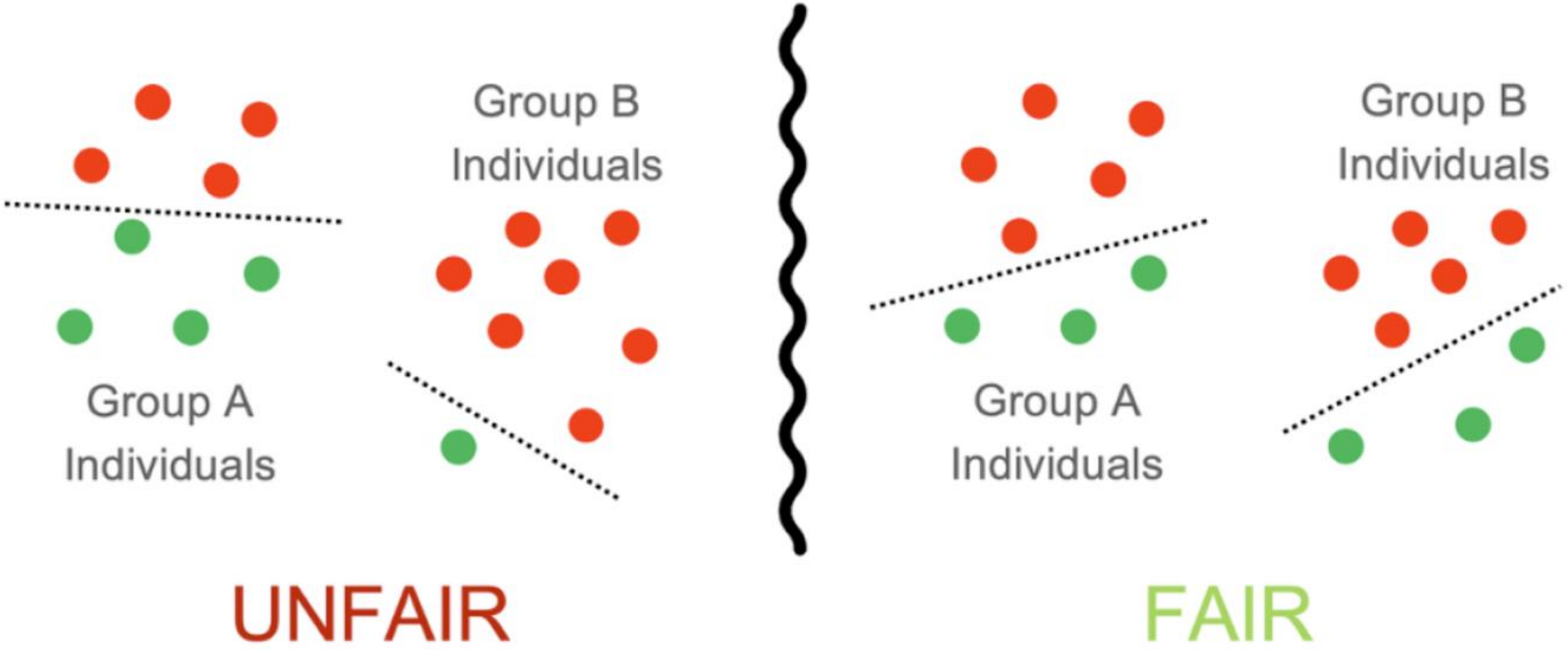
Just because you can do something, doesn't mean you should.

Think about the humans your creation impacts!

Who benefits and who might be harmed?

What defines high-quality, representative data for your product?

"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


Dataset 1
Dataset 2
All swans are white

Attacking discrimination with smarter machine learning

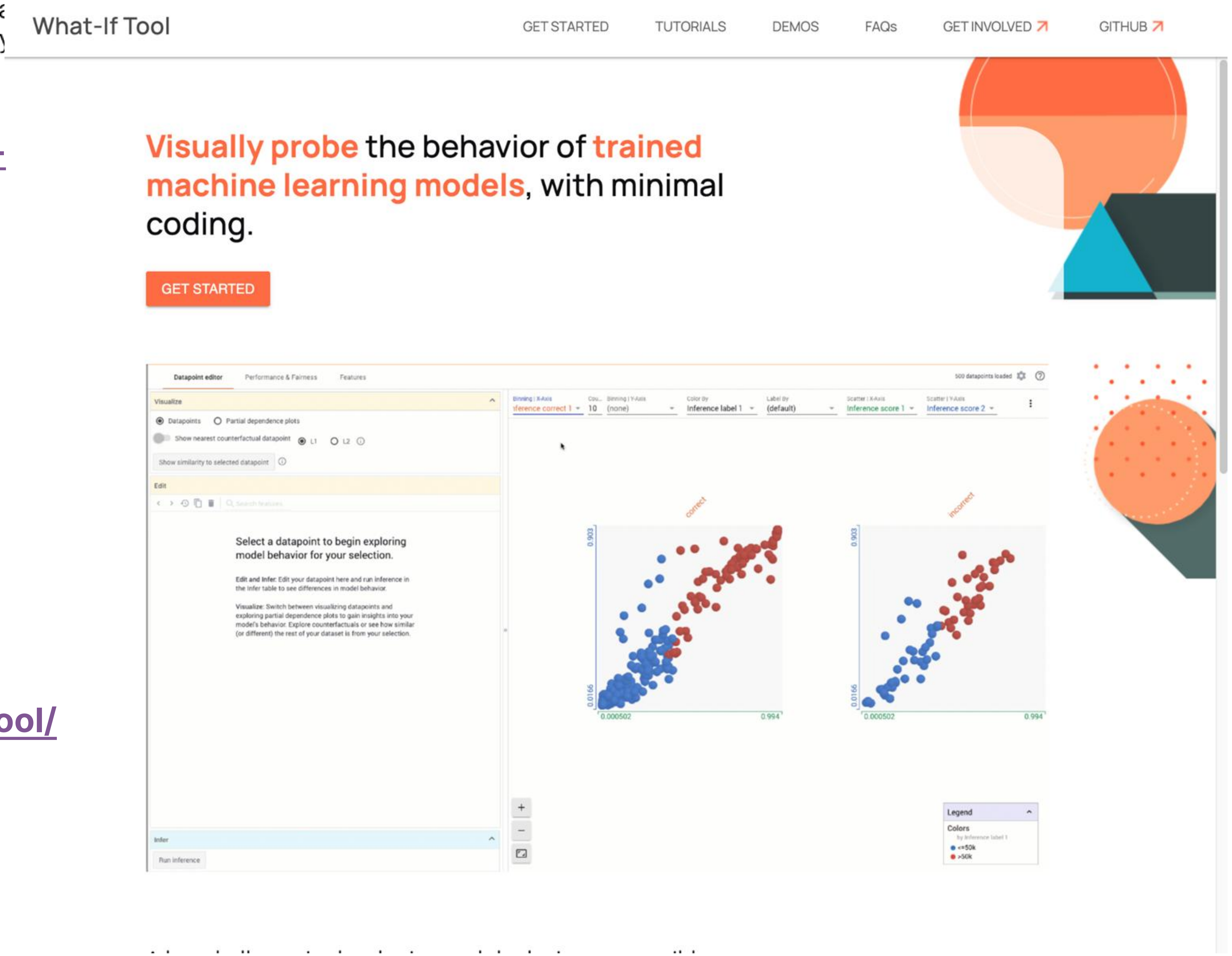
As machine learning is increasingly used to make important decisions across core social domains, the work of ensuring that these decisions aren't discriminatory becomes crucial.

Here we discuss "threshold classifiers," a part of some machine learning systems that is critical to issues of discrimination. A threshold classifier essentially makes a yes/no decision, putting things in one category or another. We look at how these classifiers work, ways they can potentially be unfair, and how you might turn an unfair classifier into a fairer one. As a illustrative example, we focus on loan granting scenarios where a bank may grant or deny loan based on a single, automatically computed number such as a credit score.

By Martin Wattenberg, Fernanda Viégas, and Moritz Hardt.

This page is a companion to a recent paper by Hardt, Price, Srebro, which discusses ways to

research.google.com/bigpicture/attacking-discrimination-in-ml/



pair-code.github.io/what-if-tool/

When should the system provide explanations?

Trust and explainability are inherently linked

Explanations are different depending on the **stakeholder....**

Developers

Users

Regulators

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

Challenges

- 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?

Diversity of perspective matters!

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

Workshops

Exercises for your team to put the Guidebook's design patterns into practice in your product.

Workshop slides

Use this set of slides to guide you as you:

- Define user problems
- Make your system explainable
- Develop feedback and control mechanisms
- Mitigate system errors
- Manage data responsibly

Download pdf, 2.2mb

Download ppx, 2.7mb

5 mins

Scenario 4

Apply back to your product:

How might someone use your system in an unintended or adversarial way? How might you prevent this?

What obligation do you have to your users when this happens?

Your answer here

Google

People + AI Guidebook

Day 1 agenda

Morning		Afternoon	
9:00	Arrival & review opportunity statements	1:00	Controls audit
9:15	Errors audit	1:30	Feedback audit
10:00	**Break**	2:00	**Break**
10:10	Explainability audit	2:10	Dataset checklist
10:50	AI Onboarding	2:30	Review questions
11:30	Break/Stretch	3:00	Share with group and next steps
11:35	Trust calibration	3:30	**End of workshop**
12:00	**Lunch break**		

Facilitator note

Feel free to modify this to suit the needs of your workshop.

Delete this note

Google

Peo

https://pair.withgoogle.com/guidebook/workshops

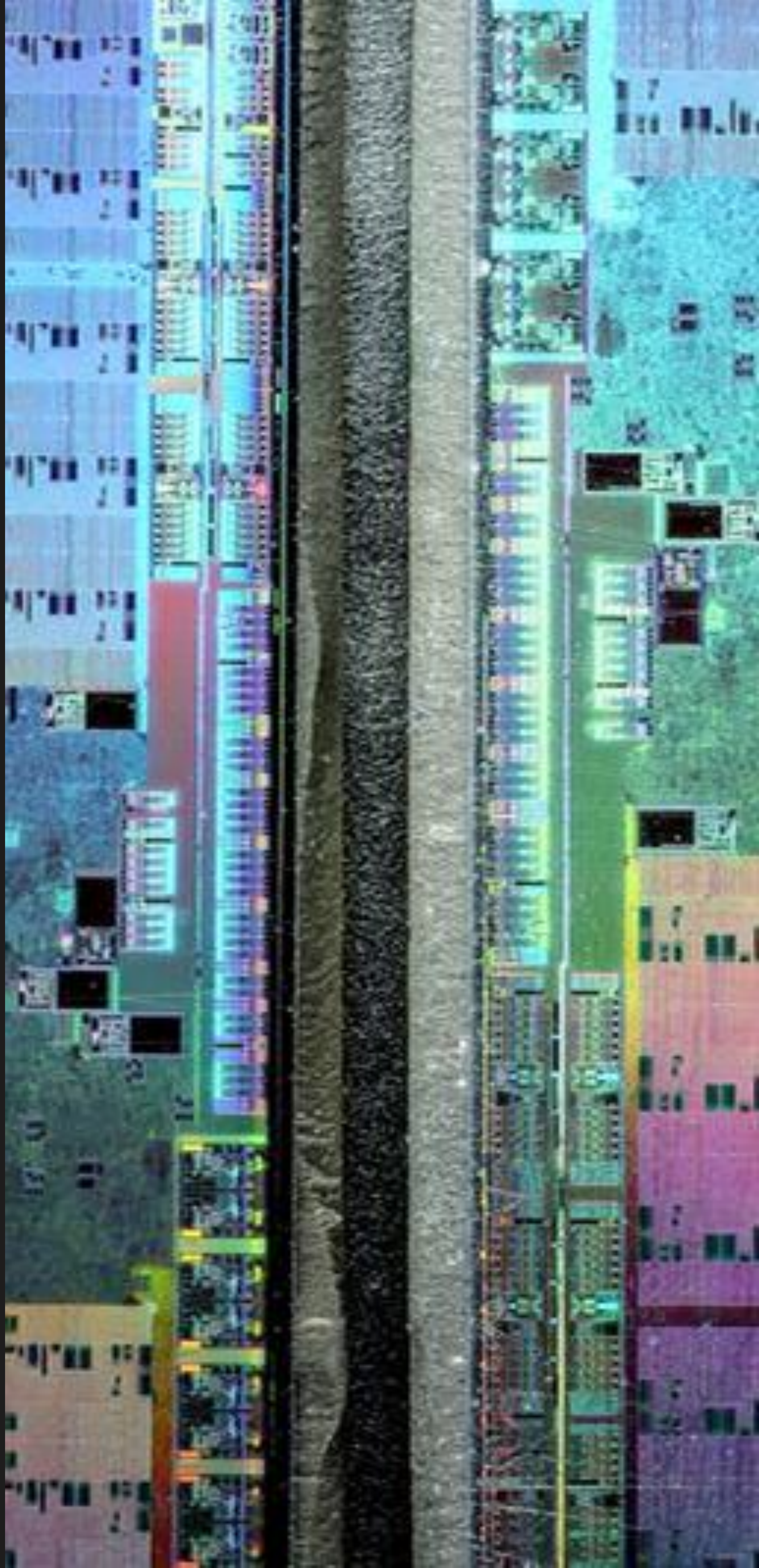
Summary

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