Cognitive biases in AI

Pierre Lauquin & Rafael Tiedra, May 2026

“ Does your expertise protect you from bias? “

A - Yes, to some extent

B - No, not really

Contents

Introduction

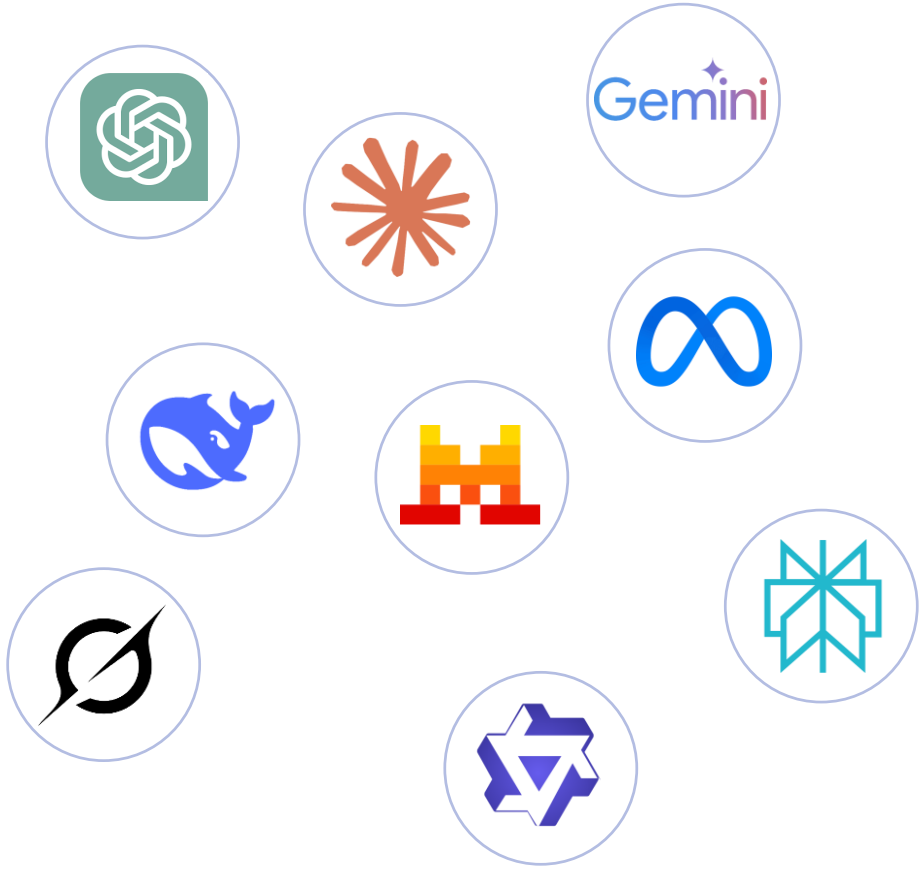
“ Define your terms, or we shall never understand one another. “

« Définissez les termes, vous dis-je, ou jamais nous ne nous entendrons. »

Voltaire - Dictionnaire Philosophique, 1764

Introduction

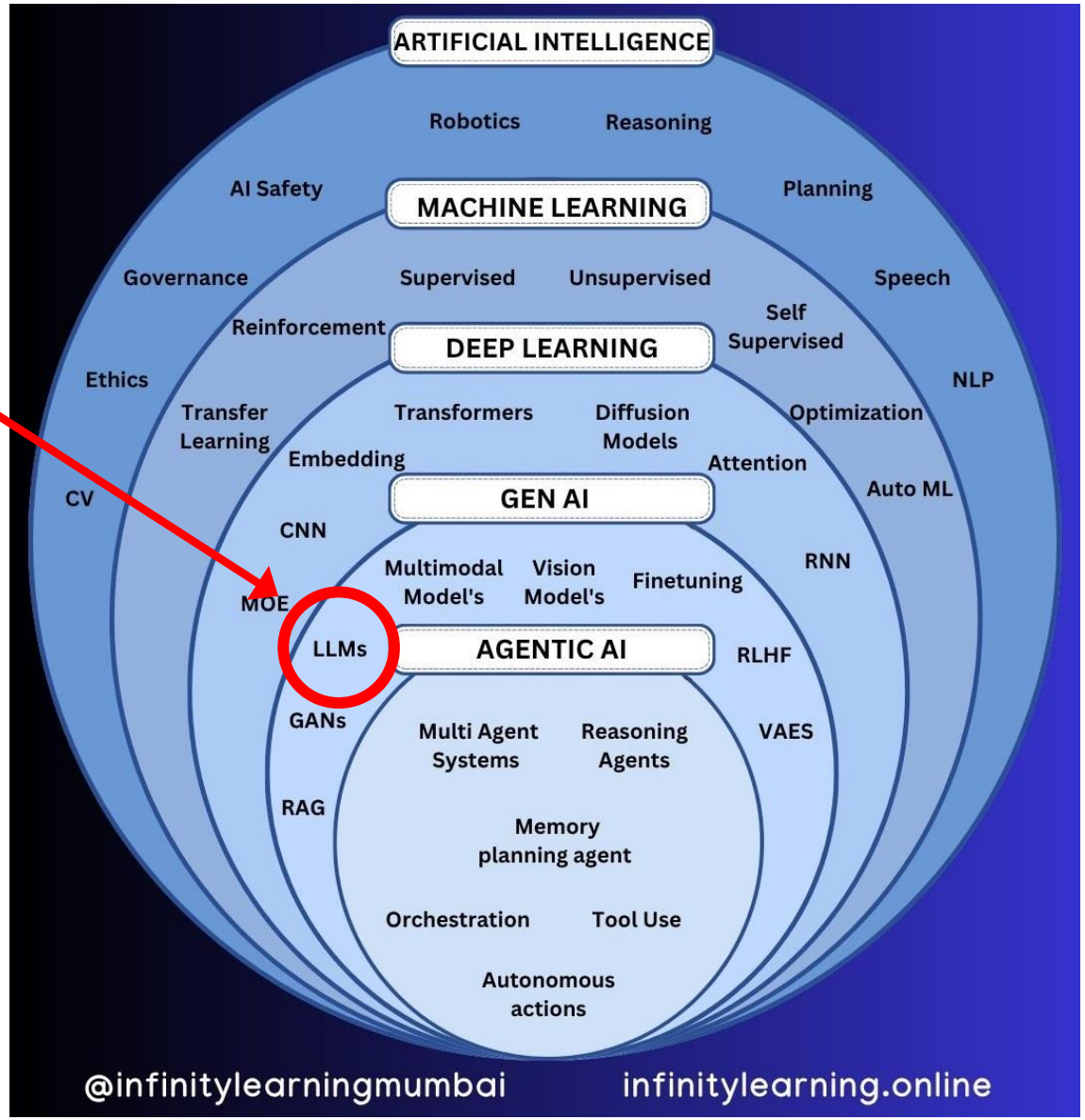
We talk a lot about LLMs, but...



Introduction

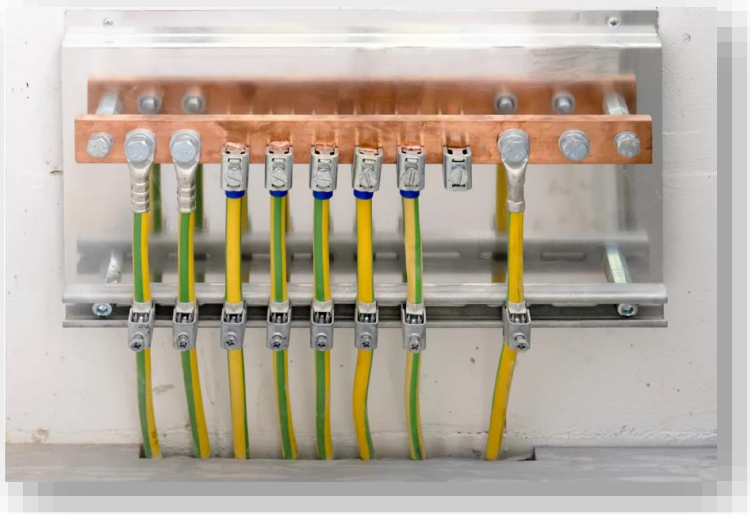
… LLMs are one type of AI systems.

There are numerous other types of AI systems !

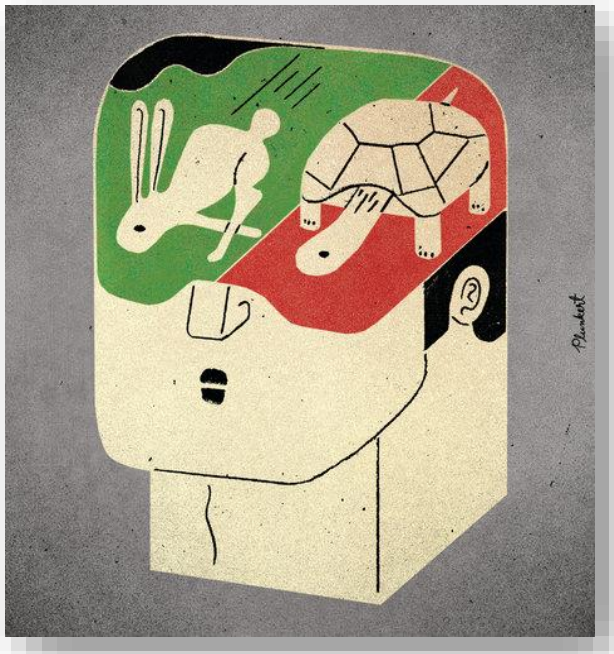


Introduction

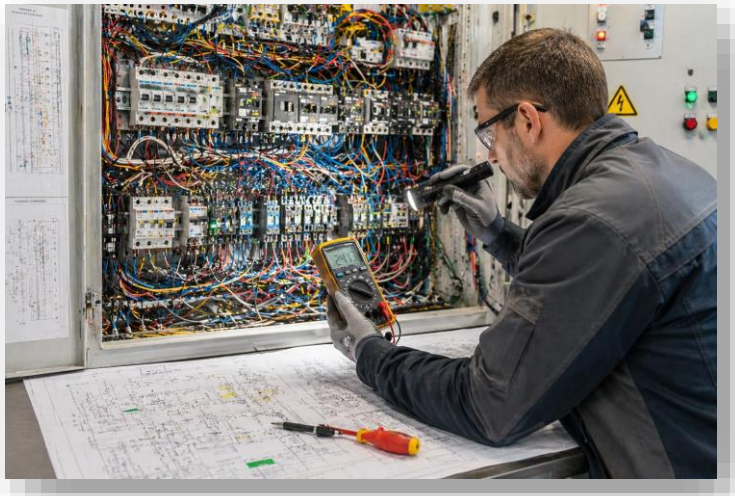
System 1



Unconscious
Fast
No energy

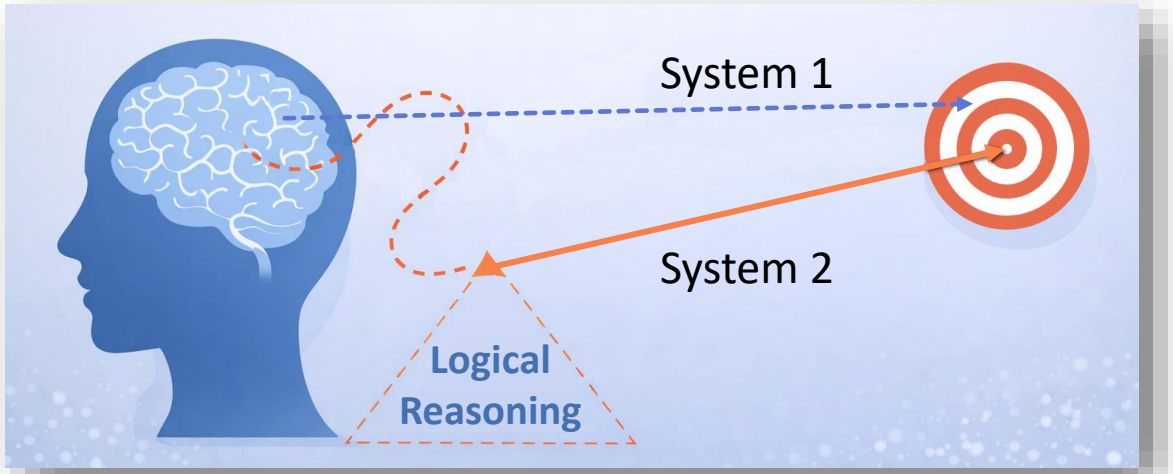


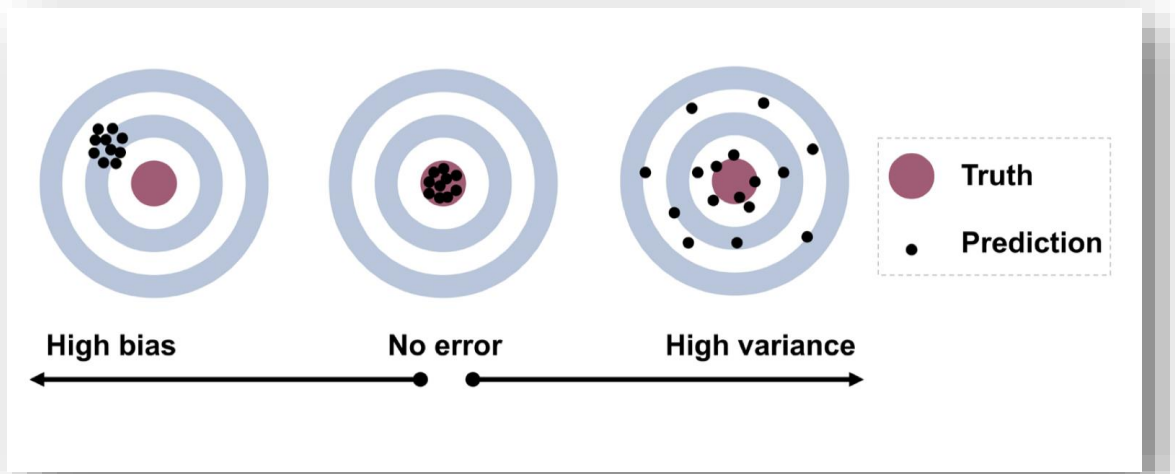
System 2



Conscious
Slow
Lot of energy

Introduction





Cognitive bias:
Systematic and predictable deviation of judgement or reasoning from a logical, statistical or probabilistic norm

Statistical bias:
Systematic discrepancy between reality and the model's estimates, caused by the way in which the data is selected, processed or used

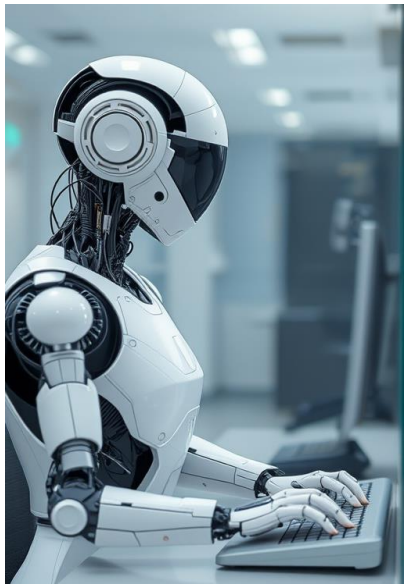
Introduction

Confirmation bias

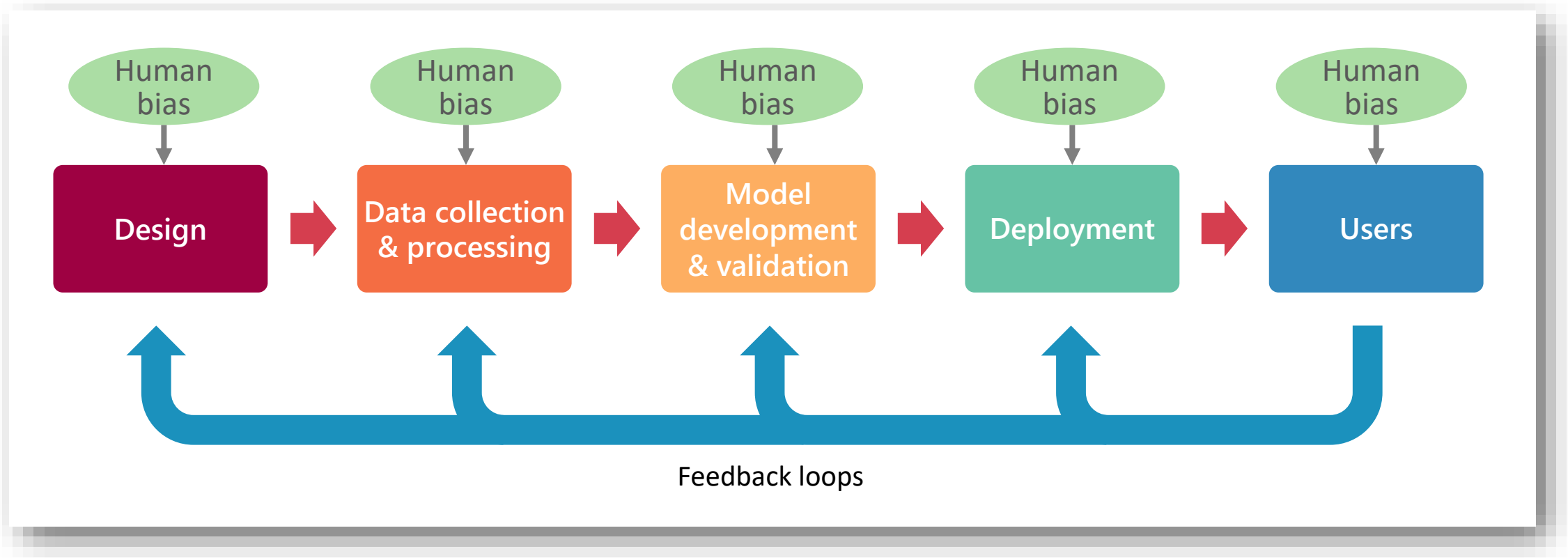
What about biases in IA ?

AI system life cycle

0. The project's founding vision
1. Definition of what constitutes a “good answer”
2. Compilation of the raw training data
3. Filtering, cleaning and curating data
4. Choice of architecture and inductive biases
5. Initial training objective
6. Alignment via human feedback (RLHF or equivalent)
7. Definition and integration of safety rules
8. Internal evaluation of the model
9. Building the user interface
10. Public deployment and indirect learning
11. Correction loops and updates



AI system life cycle



Koçak et al. (March 2025)

AI system life cycle

Fundamental biases are introduced even before data is collected.

Phase where the AI system is defined, along with its goals and success criteria, before any data or models are considered.

Examples of cognitive biases during the design: Confirmation bias, framing bias, simplification bias, …

E.g., during the design of an AI system that should predict which job applicants will be “high performers”, designers believe that candidates with degrees from top universities are inherently better performers (see Fabris et al.)

AI system life cycle

The model learns about the visible world, not the real world.

The data collection and processing phase consists in gathering the data the AI system needs and transforming it into a clean and structured form.

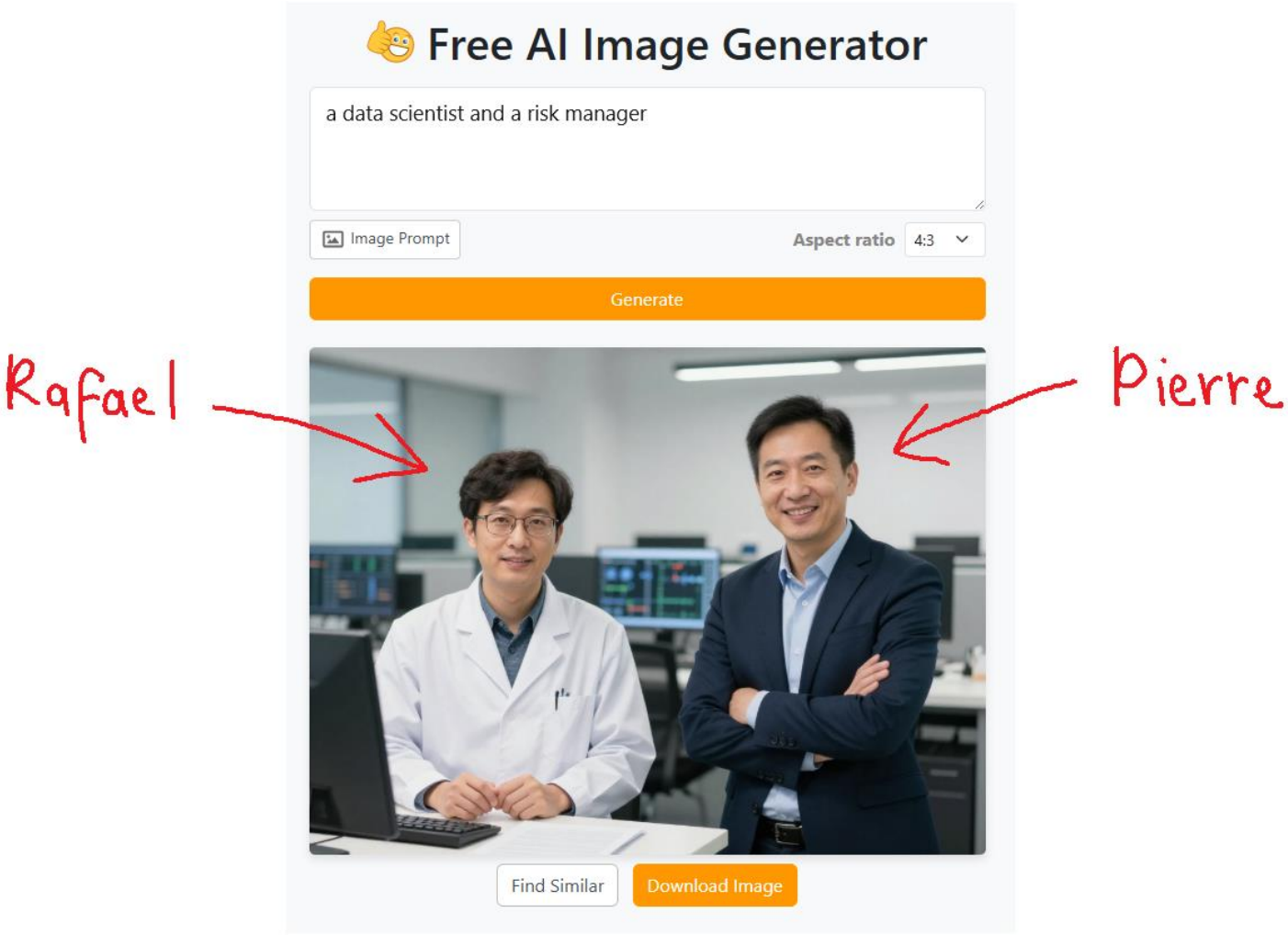
Examples of (statistical biases that can be caused by cognitive biases) during data collection:
Selection bias, historical bias, cleaning bias, …

E.g., a team collects most of its training images from dermatology clinics in wealthy, predominantly light-skinned neighbourhoods. As a result, the dataset contains mostly light-skin images (see Koçak et al.)

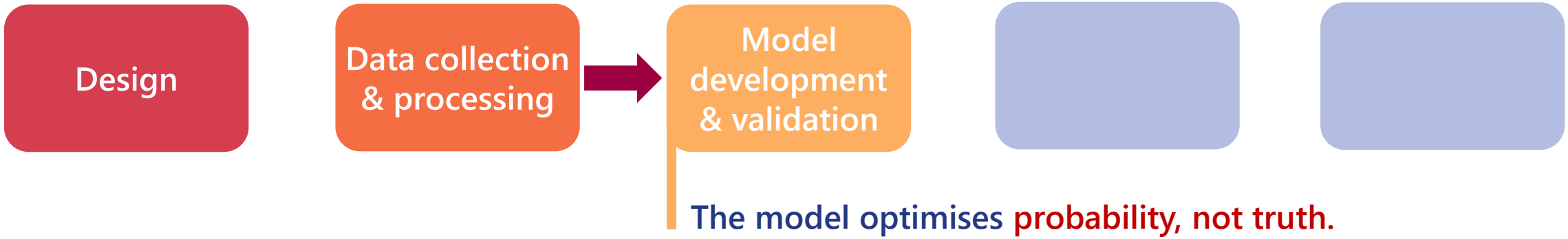
AI system life cycle

20+ trials, with the same prompt " a data scientist and a risk manager ", always produces a similar image on this AI generator.

A case of imbalanced training dataset ?



AI system life cycle



Model development is the process of building an AI system by choosing algorithms and training them on collected data to perform tasks.

Example of cognitive bias during model development: **Availability bias** (developers may rely more on familiar algorithms or validation techniques rather than those best suited to the problem).

Other biases: **Mean bias, correlation vs causality bias, …**

For instance, to predict customer churn for a subscription service, data scientists immediately choose a logistic regression algorithm as the baseline because they've used it successfully in past churn-prediction projects, without challenging it.

AI system life cycle

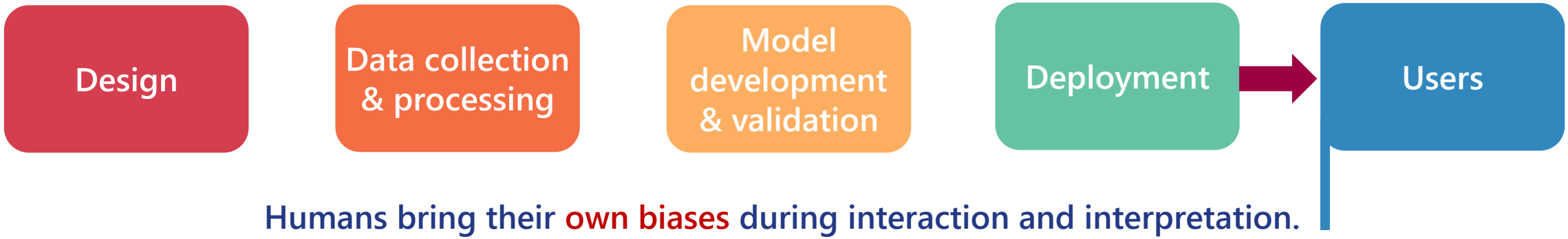
The deployment phase is where the AI system is integrated into real infrastructure so it can operate with live data and real users.

Example of cognitive bias during deployment: **Overconfidence bias** (assuming the model will perform well in real-world conditions without sufficient monitoring or testing).

Other biases: **Confirmation bias, anchoring bias, sunk cost fallacy, ...**

For instance, a hospital develops an AI system to predict patient readmission risk. During testing, the model performs well on historical data from the hospital's records, leading teams to assume that the model will perform just as well in real-world conditions (see Cross et al.).

AI system life cycle



User interaction is the phase where people use the AI system in real-world conditions, providing inputs and receiving outputs.

Example of cognitive bias during user interaction: **Automation bias** (users may over-trust the AI's outputs and rely on them without critical thinking).

Other biases: **Fluency bias, authority bias, halo effect, …**

E.g., a loan officer relies on an AI credit-scoring tool that mistakenly labels a reliable customer as high risk due to a data-entry error. Instead of double-checking the customer's information, the officer accepts the AI's recommendation and denies the loan (see Parasuraman et al.)

AI system life cycle

Example of automation bias

Driver: “ This road looks like it ends in a lake. “

GPS AI: “ In 50 meters, continue straight. “

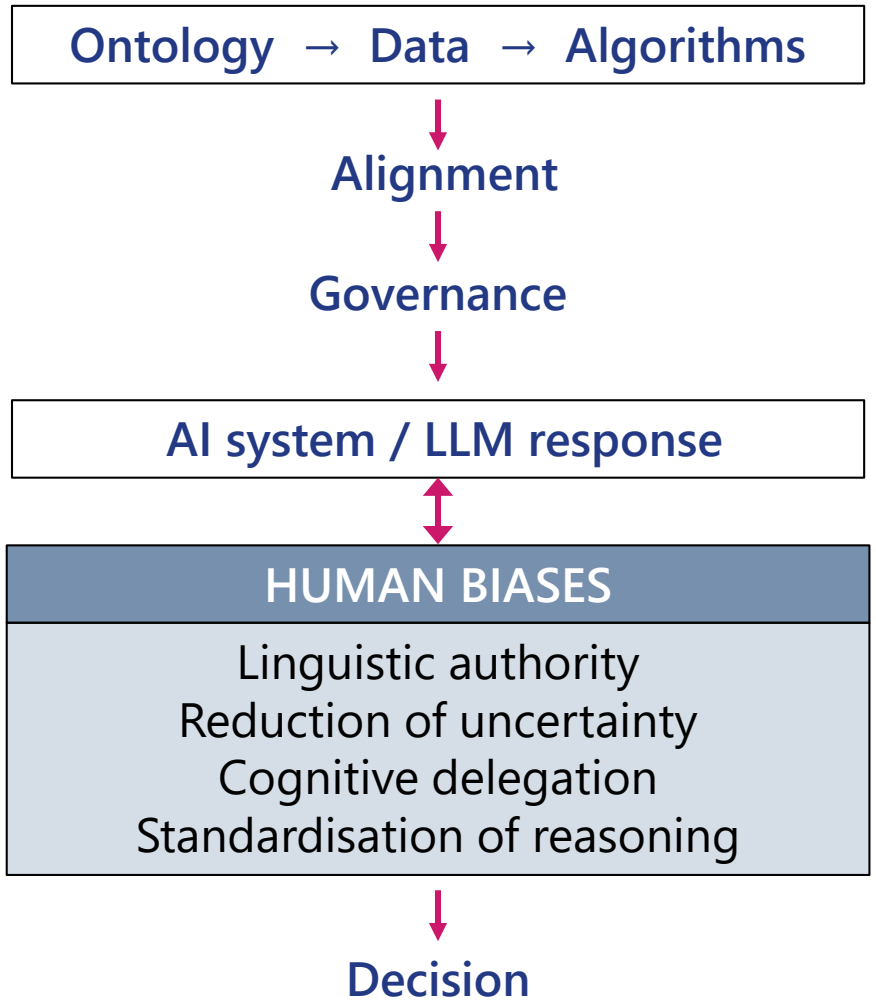
Passenger: “ There are ducks swimming there. “

Driver: “ The satellite probably knows something we don’t. “

(The car slowly rolls into the water.)

AI system life cycle

First conclusions



" An AI system is a medium that transmits bias, from model biases... to human biases. "

An AI system is not biased “ by accident “.

It is biased as the product of thousands of human, cultural, technical and political decisions, made under uncertainty.

Some AI models biases mitigation

Clarify the scope of use before building the model

E.g.: For a fraud detection model in the insurance sector, specify whether it is used to prioritise checks or to automatically reject a claim. In the first case, the risk is manageable; in the second, bias may lead to unfair or unlawful decisions.

Audit the training data … before training

E.g.: An HR model trained on an organisation's past recruitment data may learn to replicate that organisation's historical preferences, even if no explicit variables such as gender, age or ethnic background are included.

Test performance by subgroup, not just as an average

E.g.: A speech recognition model may show an average accuracy of 95%, but this figure can drop significantly for certain accents, second languages or older people. The average then becomes misleading.

Incorporating an independent human review

E.g.: A data scientist can validate the statistical performance, whilst a legal expert identifies a risk of indirect discrimination and a business specialist spots an operational inconsistency.

Some humans biases mitigation

- **Do not confuse mathematical precision with truth**

  E.g.: A customer risk score of 87% may seem very precise. But if the historical data reflects past controls that were already biased, the score is essentially just a reflection of the organisation's history.

- **Be wary of automation and fluency bias**

  E.g.: Before acting on a rejection recommendation, ask a colleague: " What would be the best reason not to follow the model? "

- **Replace rigid categories with uncertainty intervals**

  E.g.: Instead of saying " high risk ", opt for: " estimated probability between 20% and 35%, with high uncertainty due to a lack of recent data "

- **Train users on the most common biases**

  E.g.: Present two model outputs with different wording, then demonstrate how participants give more credence to the one that is better written, even if it is less accurate.

- **Be wary of the anchoring effect of the first response**

  E.g.: When defining a data governance policy, request several possible architectures before choosing a plan. Otherwise, the first plan quickly becomes the implicit benchmark.

At the end...

" Does your expertise protect you from bias? "

A - Yes, to some extent

B - No, not really

" Expertise is valuable, but it can also reinforce anchoring, overconfidence or hindsight bias "

Thank you !

References

https://pmc.ncbi.nlm.nih.gov/articles/PMC11542778 (Cross et al.)

https://dl.acm.org/doi/epdf/10.1145/3696457 (Fabris et al.)

https://dirjournal.org/articles/doi/dir.2024.242854 (Koçak et al.)

https://journals.sagepub.com/doi/10.1518/001872097778543886 (Parasuraman et al.)

https://chatgpt.com (text & image generation)

https://claude.ai (text generation)

https://freegen.app (image generation)