Introdução à Aprendizagem Automática (IAA)

SUSANA BRÁS

SUSANA.BRAS@UA.PT

IAA – L12

Anomaly Detection

- Definition
- Principles
- Rational
- Categories
- Methods
- Applications
- Importance

Ethics and Responsible AI

- Representativeness
- Definition
- Innovation vs Responsibility balance
- Acceptance
- Interpretability and Explainability
- Core concepts: fairness, reliability, privacy, transparency, sustainability, accountability

Anomaly Detection

Anomaly Detection – what?

Anomaly Detection – The process of identifying unusual patterns or deviations in data that seem suspicious, because they don't conform to established norms. It is based on the detection of rare events, outliers, or inconsistencies that might indicate errors, fraud, or security breaches.

Identifying deviations - Anomaly detection focuses on finding data points or patterns that significantly differ from the majority of the data or the expected behavior.

Applications - It's used in various fields, including cybersecurity (detecting intrusions), finance (fraud detection), manufacturing (quality control), and healthcare (monitoring patient conditions).

Benefits - Anomaly detection can help identify problems early, reduce damages, and improve the accuracy and quality of data.

Anomaly Detection – core principles

Anomaly Detection	Operates under the premise that anomalies are rare and significantly different from the majority of the data
	The process involves training a model on data that is labeled as normal or assumes that the majority of the data represents normal behavior
	The model then attempts to identify data points that deviate from this established norm
	The effectiveness of anomaly detection relies on the ability to accurately define what constitutes normal behavior, which can vary widely across different domains and applications

Anomaly Detection – why?

Improved data quality: Identifying and handling data anomalies can significantly improve data quality, which is essential for accurate and reliable data analysis. By addressing data anomalies, analysts can reduce noise and errors in the dataset, ensuring that the data is more representative of the true underlying patterns.

Enhanced decision making: Data-driven decision making relies on accurate and reliable data analysis to inform decisions. By identifying and handling data anomalies, analysts can ensure that their findings are more trustworthy, leading to better-informed decisions and improved outcomes.

Optimized machine learning performance: Data anomalies can significantly impact the performance of machine learning algorithms, as they can cause the model to fit the noise rather than the underlying pattern in the data. By identifying and handling data anomalies, analysts can optimize the performance of their machine learning models, ensuring that they provide accurate and reliable predictions.

Anomaly Detection – steps

Anomaly detection involves three main steps:

1. **Data preprocessing:** involves cleaning, transforming, and standardizing the data to make it suitable for anomaly detection.

2. **Anomaly identification:** involves applying one or more techniques to identify the anomalies in the data. These techniques can be based on statistical methods, machine learning algorithms, or domain knowledge.

3. **Anomaly analysis:** involves interpreting and explaining the anomalies, as well as taking appropriate actions to resolve them.

Anomaly detection is not a one-size-fits-all problem. Different types of data and domains may require different approaches and techniques. Therefore, it is important to understand the nature and context of the data, as well as the objectives and requirements of the anomaly detection task.

Anomaly Detection – categories

Point Anomaly

- These are individual data points that deviate significantly from the rest of the data.

Contextual Anomalies

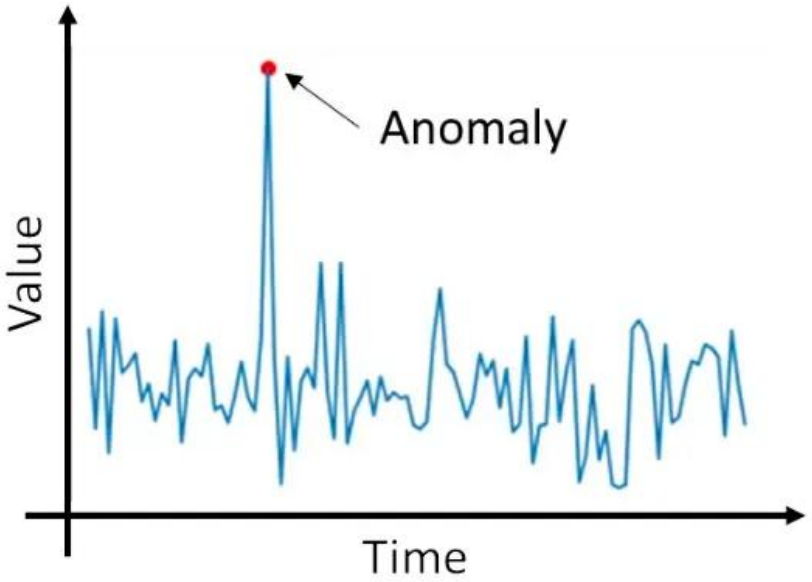
- These are data points that deviate significantly from the normal behavior of the data in a specific context. The context can be defined by temporal, spatial, or other attributes.

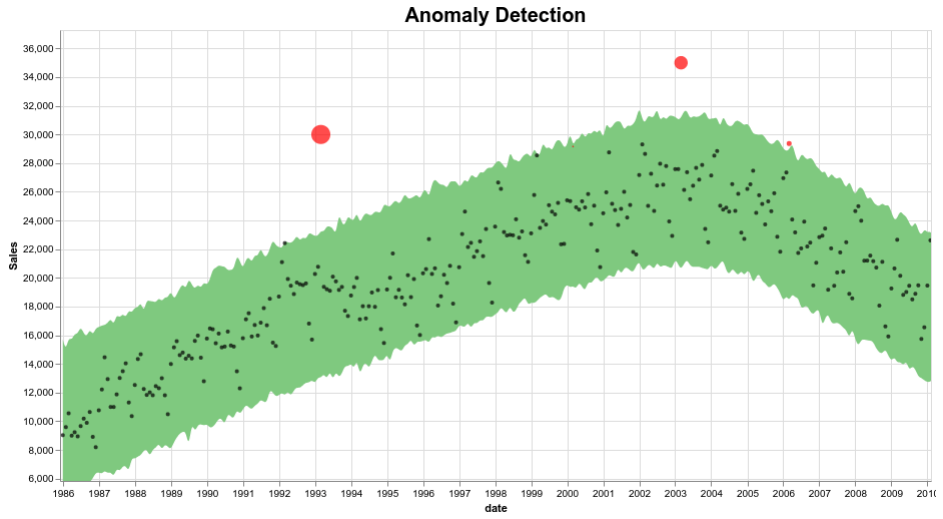
Collective Anomalies

- These are groups of data points that deviate significantly from the rest of the data as a whole. The individual data points may not be anomalous by themselves, but their collective behavior is anomalous.

Anomaly Detection – how?

Visualization – A powerful tool for detecting data anomalies, as it allows data scientists to quickly identify potential outliers and patterns in the data.



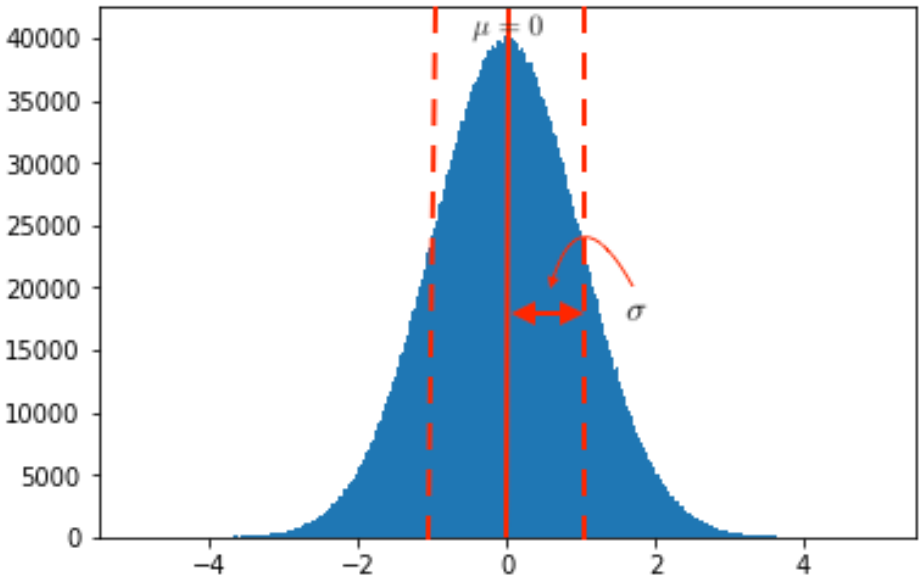


Anomaly Detection – how?

Statistical tests - by comparing the observed data with the expected distribution or pattern.

If $x \in \mathbb{R}$, and x follows Gaussian distribution with mean, μ and variance σ^2, denoted as,

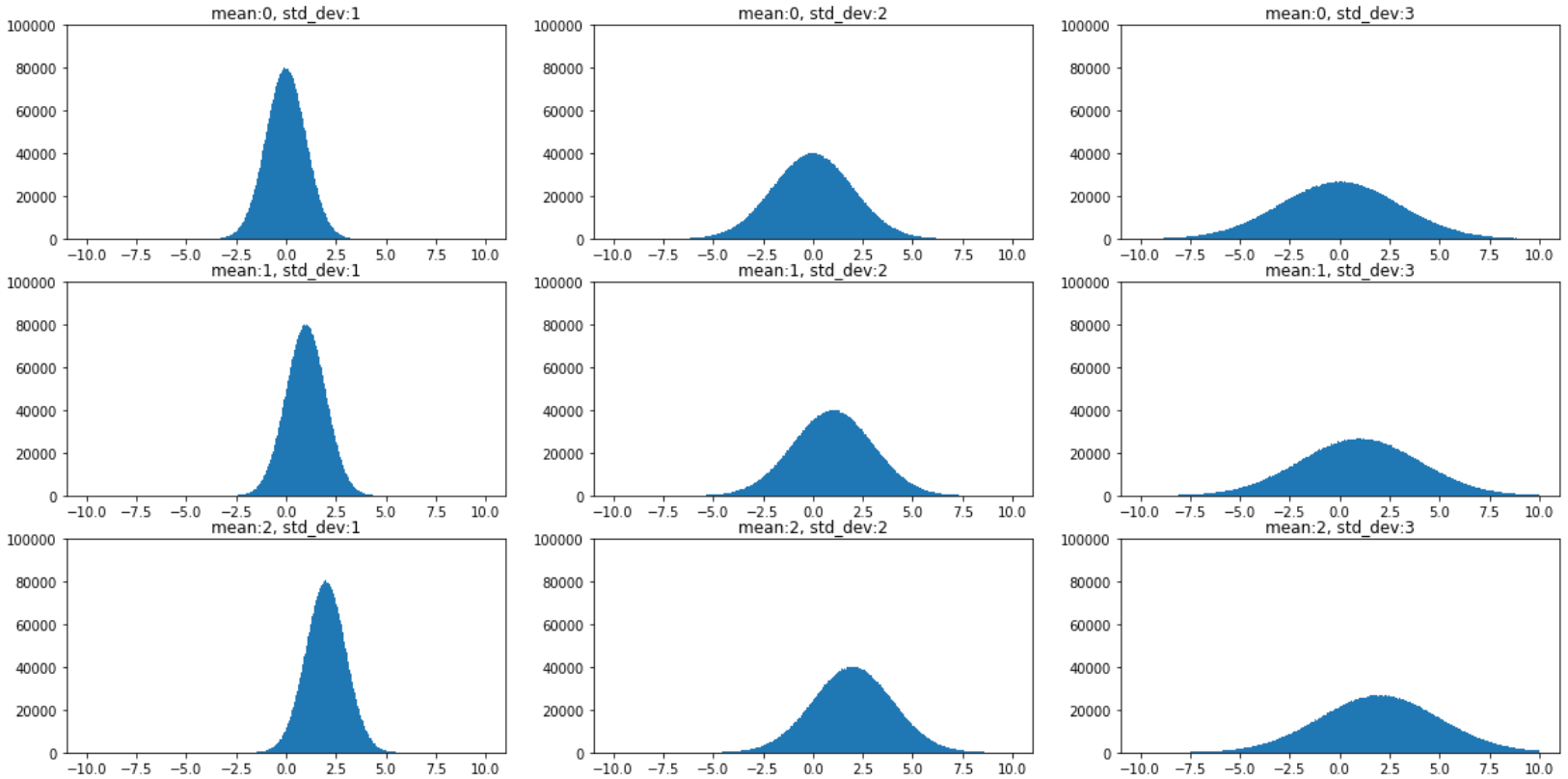
$$x \sim \mathcal{N}(\mu, \sigma^2)$$



Standard normal Gaussian distribution (μ=0, standard deviation σ=1). Density is higher around μ and reduces as distance from mean increases. If we know parameters μ and σ, the probability of x in Gaussian distribution is:

$$p(x; \mu, \sigma^2) = \frac{1}{\sqrt{2\pi}\sigma} exp(-\frac{(x-\mu)^2}{2\sigma^2})$$

Anomaly Detection – statistical test



Mean (μ) defines the center of the distribution.

Standard deviation (σ) defines the spread of the Gaussian distribution.

As the spread increases, the height of the plot decreases, because the total area under a probability distribution should be = 1.

Anomaly Detection – how?

Machine learning algorithms - detect data anomalies by learning the underlying pattern in the data and then identifying any deviations from that pattern. Some of the most common ML anomaly detection algorithms include: isolation forest, One-Class SVM, k-NN, Naive Bayesian, Autoencoders, Local Outlier Factor (LOF), k-means

Supervised methods use the labeled data to train a classifier that can distinguish between normal and anomalous data points. (Anomaly Classification, Anomaly Prediction)

Unsupervised methods use the unlabeled data to learn the normal behavior of the data and identify the data points that deviate from it. (Anomaly Detection, Anomaly Scoring)

Ethical & Responsible AI

What do we mean by “Ethics” in the AI Era?

Photo by Walls.io on Unsplash

Ethics concerns the **systematic evaluation of what ought to be done** when human well-being, rights, and social structures are affected.

In the **AI context**, ethics is not about intentions but about **consequences, power, and accountability**.

What do we mean by “Ethics” in the AI Era?

Photo by Walls.io on Unsplash

Ethics

Philosophical reflection on **moral principles** guiding human action.

Applied AI Ethics

The **evaluation** of the design, deployment, and **governance** of algorithmic systems in relation to human rights, justice, and social welfare.

Responsible AI

Institutional and engineering practices that attempt to **operationalize** ethical principles in AI systems.

The uncomfortable truth: AI systems are not neutral

AI systems are built from human decisions

Every system depends on choices:

- ✓ what problem is worth solving
- ✓ what data to collect
- ✓ who is represented
- ✓ what metric defines success
- ✓ what errors are acceptable

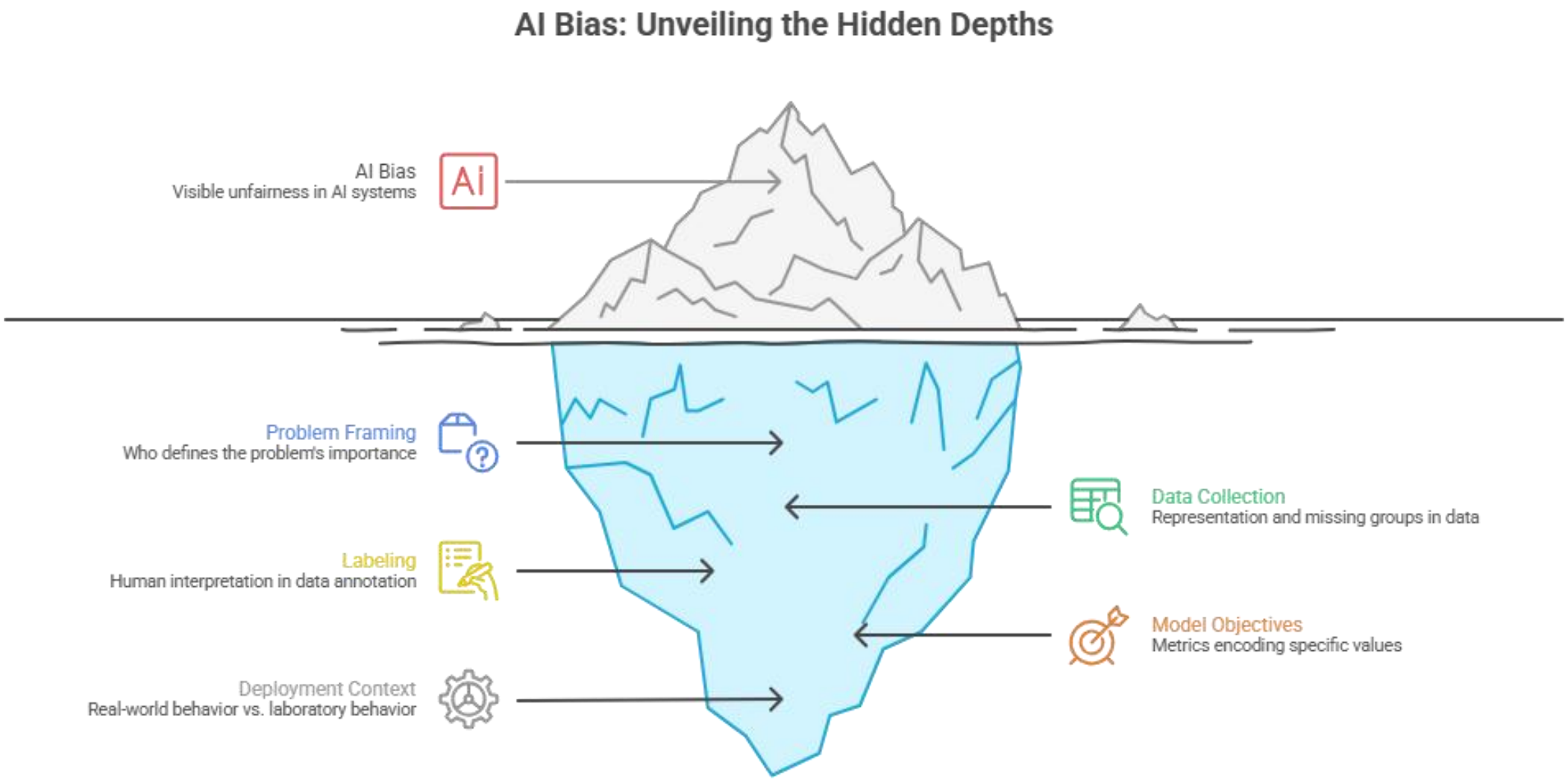
Data encodes historical reality

Datasets are not neutral descriptions of the world.
They contain the **historical record of social inequalities**.
(E.g. hiring, policing, credit scoring, healthcare access)

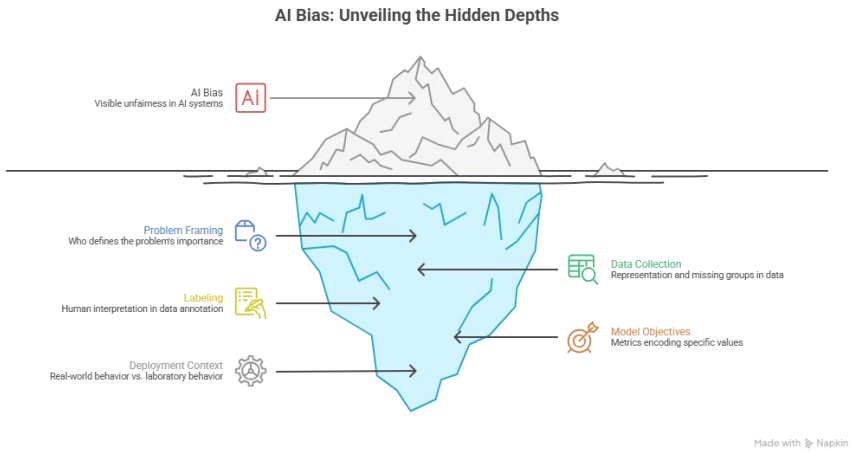
Optimization amplifies patterns

Machine learning systems scale correlations, including harmful ones.
If the data reflect structural discrimination, the system may **replicate or intensify it**.

Where bias actually enters AI systems



Where bias actually enters AI systems



Problem framing

Who decided the problem matters?

E.g.: predictive policing assumes crime prediction is desirable.

Data collection

Who is represented?

Who is missing?

E.g.: marginalized groups are often under-recorded or over-policed, producing distorted datasets.

Labeling

Human annotators impose interpretation.

E.g.: what counts as “toxic language” or “suspicious behavior”.

Model objectives

Metrics encode values.

E.g.: accuracy alone can hide unequal error distribution.

Deployment context

A system used in a laboratory behaves differently when deployed in a real social system.

Historical structures matter more than engineers think

Foto de Andrik Langfield na Unsplash

Data-Driven Decisions

Organizations increasingly rely on data to guide strategies, affecting workers, users, customers, etc.

Consequences of Bias

Non-representative data can lead to systematic biases, flawed models, and lost opportunities.

Real-World Impacts

From product to design to public health, biased data has measurable consequences on people's lives.

Data reflect historical power structures

If those patterns are used as training data, the system may **mathematically legitimize past injustice**.
This is why ethical AI cannot be **reduced** to technical fixes.

Technology can automate inequality while appearing objective.

Historical structures matter more than engineers think

Coverage, Diversity, Balance.
A dataset is representative when it reflects the population or phenomenon it models. It mirrors the characteristics of the target population.

Fairness

Crucial for fairness in AI systems

Inclusivity

Ensures that the models do not favor any group or do not disproportionately disadvantage underrepresented groups

Generalizability

Less biases in training data leads to less biased and more accurate outcomes, better generalization

Historical structures matter more than engineers think

Foto de Andrik Langfield na Unsplash

Representativeness

- ✓ influences the design of algorithms
- ✓ influences the evaluation of algorithms performance
- ✓ promotes fairness
- ✓ implies better generalizability

Lack of Representativeness

- ✓ **unfair** and **discriminatory** outcomes, especially for minority groups
- ✓ leads models to make **judgments** based on **stereotypes rather** than **statistical evidence**
- ✓ **perpetuated** by historical discrimination, selection biases, and the complexity of demographic variables

The illusion of technical solutions to ethical problems

From Data to Decisions:
Does Bias in AI Reflect Bias in Society?

- ✓ Sampling bias
- ✓ Measurement bias
- ✓ Algorithmic bias

There is a persistent belief that fairness can be solved by

- ✓ debiasing datasets
- ✓ fairness metrics
- ✓ algorithmic auditing

They are useful tools. But they do not solve structural injustice.

Technical fixes operate inside a system whose goals were already defined.

If the problem formulation itself is ethically problematic, fairness metrics will not rescue it.

The illusion of technical solutions to ethical problems

Conceptual Understanding

Ethics in AI is not just about statistical accuracy but also involves **inclusiveness** and **fairness**. It is a **foundational concept** that has been debated in various fields and is now central to discussions on AI.

Cross-Disciplinary Awareness

Collaborate with ethicists, sociologists, minorities and marginalized communities to contextualize data. **Question** whose data, voices, and perspectives are missing.

Representation Beyond Data

Advocate for **representation** in teams, leadership and policies that shape data practices**.**

The illusion of technical solutions to ethical problems

Incomplete Mirrors

Our datasets, like our systems, often reflect only part of the whole.
What is left unseen?
Document, quantify, describe, propose solutions.

Beyond the Numbers

Bias is not always in the model.
It starts with the problem framing and who is excluded from it.
Question, advise, propose.

Ongoing Vigilance

Representation is not a one-time fix.
It is a continuous, ethical commitment.
Be aware, discuss, be interventive

Non-negotiable ethical values in democratic societies

AI reflects the values embedded in the systems

AI systems do not inevitably produce injustice. They produce outcomes determined by:

- ✓ the goals set by designers
- ✓ governance structures
- ✓ institutional incentives
- ✓ regulatory constraints

Non-negotiable ethical values in democratic societies

Human dignity
AI systems must not treat people purely as data points or optimization targets.

Autonomy
Individuals should retain meaningful agency over decisions affecting their lives.

Justice
Benefits and harms must be distributed fairly.

Accountability
Someone must remain responsible for decisions made using AI systems.

Transparency
Systems influencing rights must be open to scrutiny.

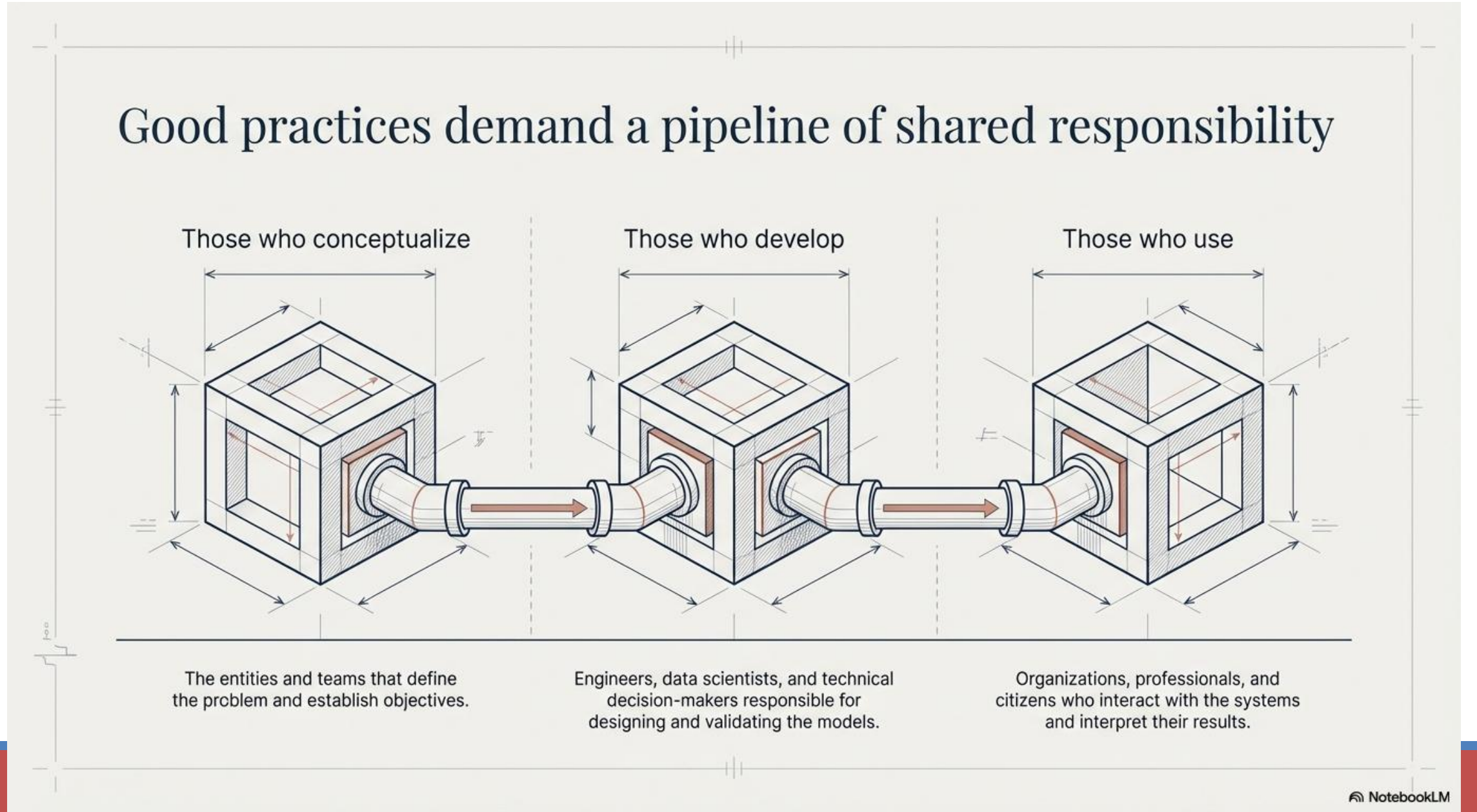
Non-negotiable ethical values in democratic societies

The ethical risks of AI are not primarily technical failures.
They are failures of governance, responsibility, and moral imagination.

Developing AI systems is not just engineering
It is participation in social decision-making.
And that responsibility cannot be outsourced to the algorithm.

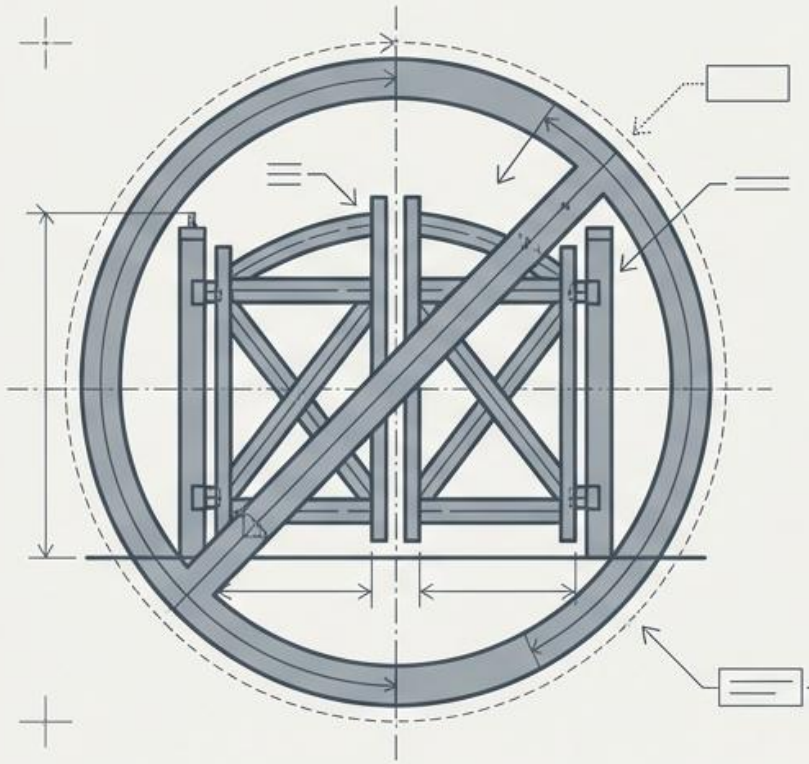
AI systems do not simply reflect society; they shape the future distribution of power.

Good Practices in AI Pipeline Development



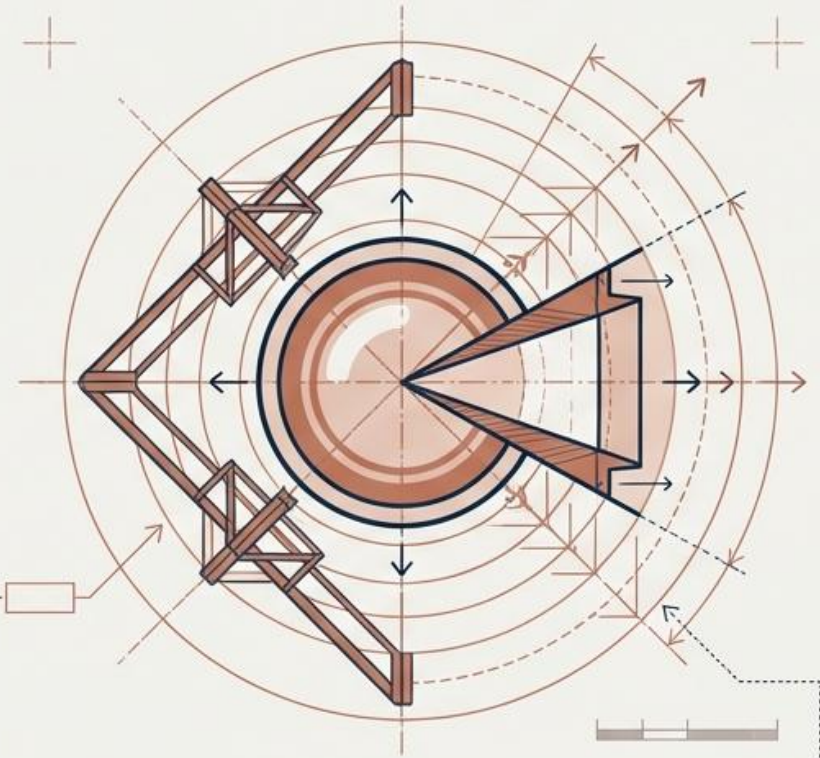
Good Practices in AI Pipeline Development

The questioning paradigm drives safe and fair value



Questioning does NOT mean:

Rejecting, limiting, or discrediting technology.

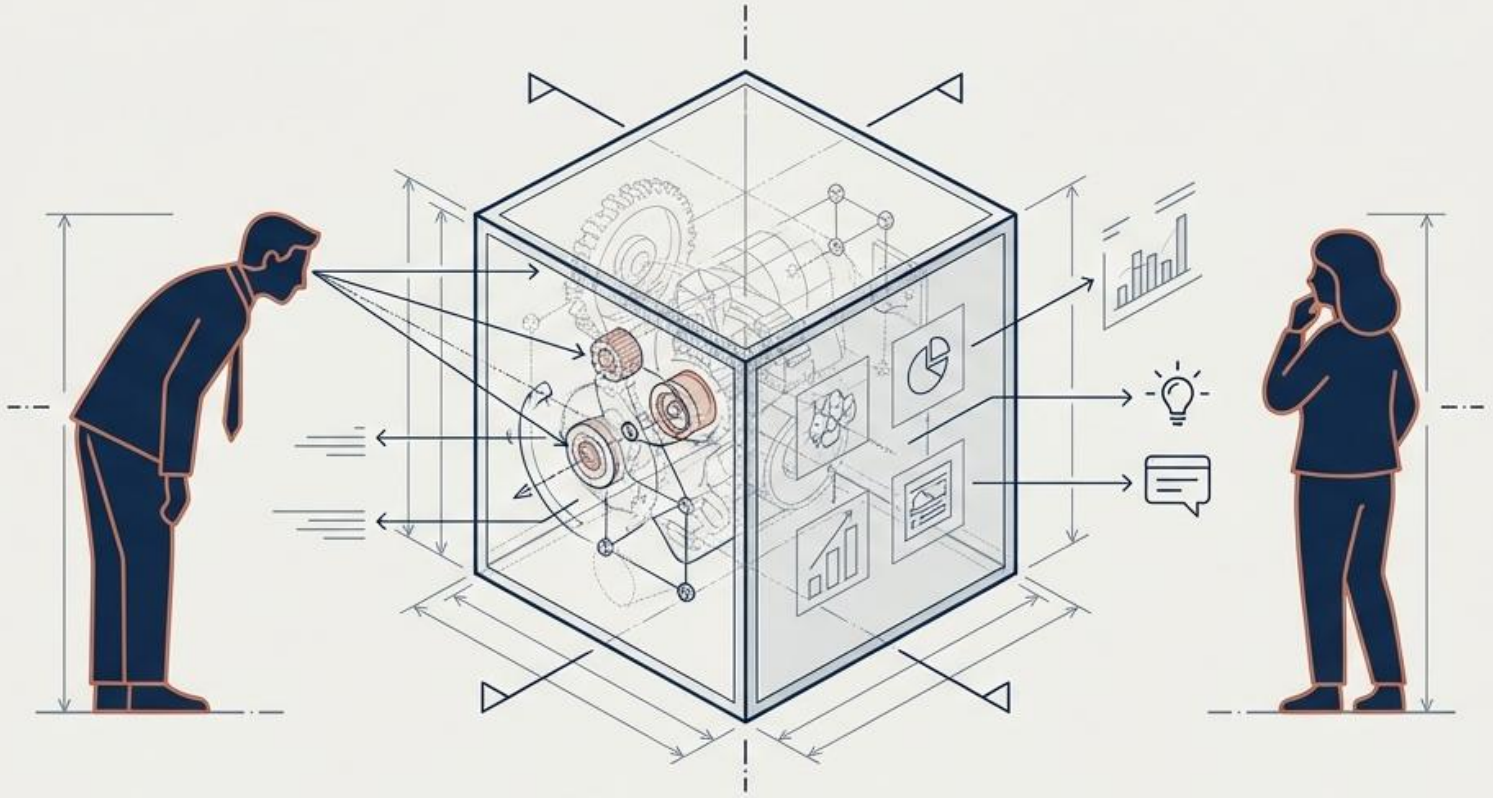


Questioning MEANS:

Deepening knowledge, assessing implications, and ensuring that AI adds value safely and fairly.

Good Practices in AI Pipeline Development

Universal design constructs the infrastructure of societal trust



The future of AI depends on trust. Acceptance requires auditability by both specialists and users.

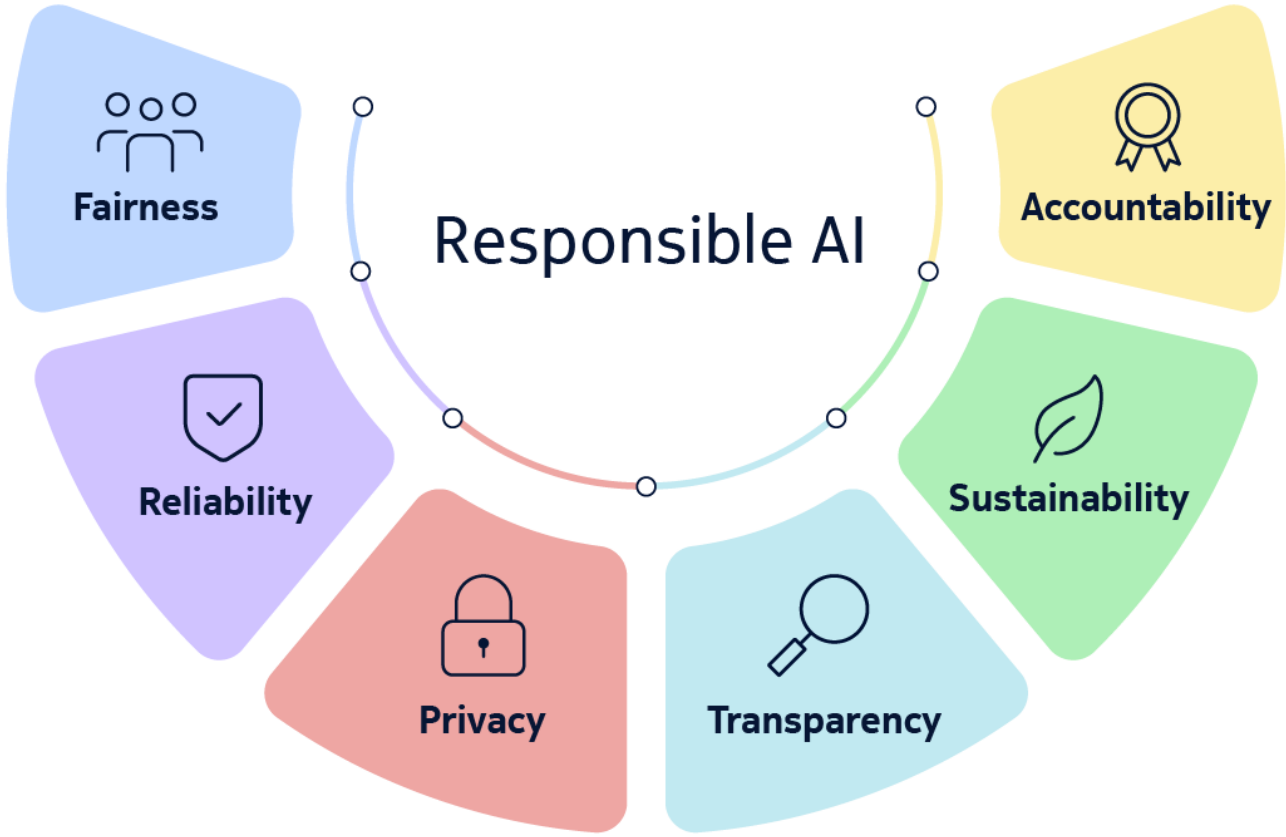
Systems must be conceived to minimize risk, enable scrutiny, and operate with sufficient transparency.

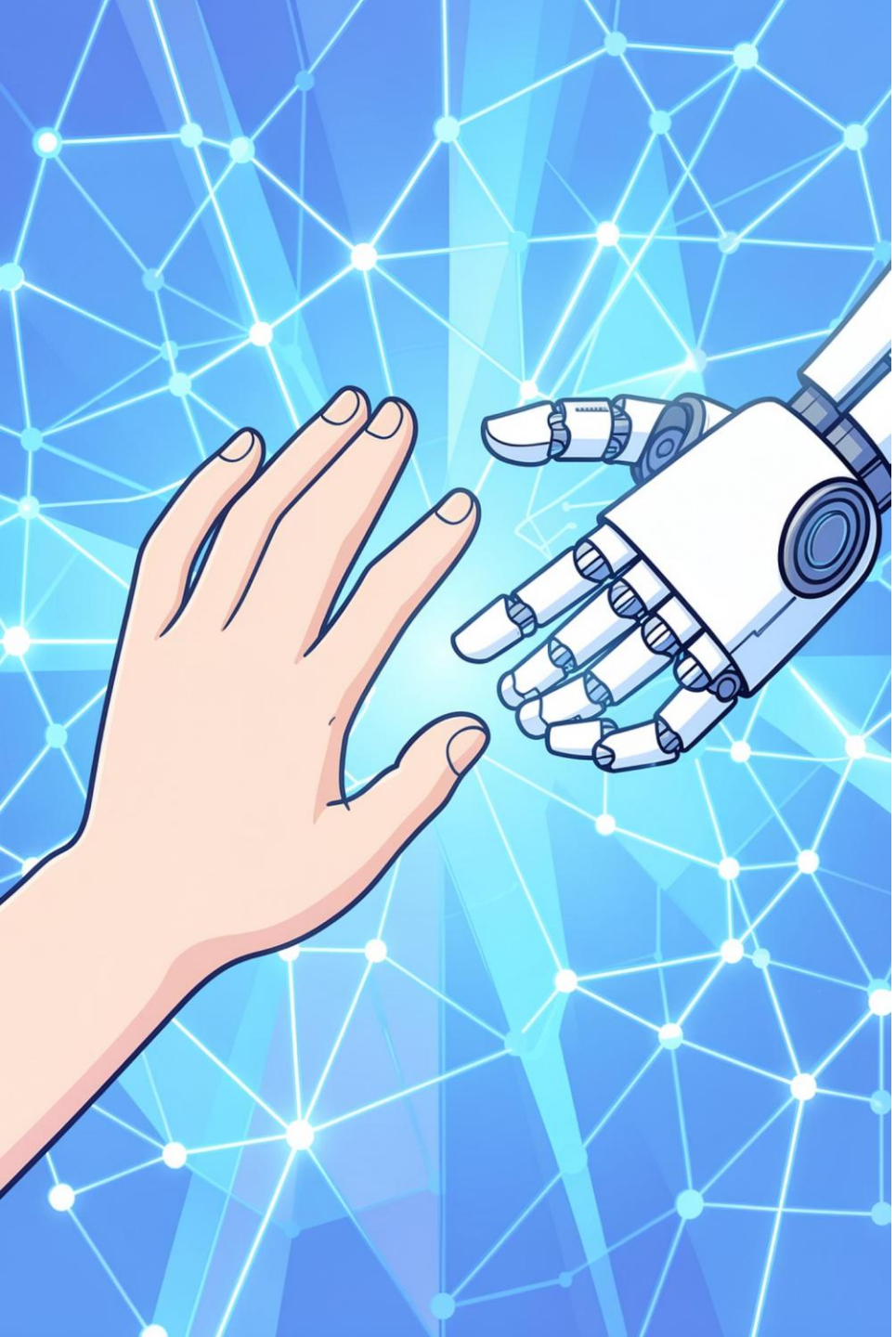
We must all be able to understand their limitations, uncertainties, and margins of error.

Good Practices in AI Pipeline Development

Cultivating a comprehensive and multidisciplinary AI literacy

This literacy forms the foundation for an ecosystem in which AI becomes an instrument of progress rather than a driver of inequality.


Responsible AI
Fairness
Accountability
Reliability
Sustainability
Privacy
Transparency



AI Ethics: Building the Foundation for Responsible Innovation

As artificial intelligence reshapes industries and societies, the question is no longer *can* we build these systems — but *how* we build them, and *who* they serve. Ethics is not a constraint on progress. It is the condition for sustainable, trustworthy, and human-centered AI.

The ethical risks of AI are not primarily technical failures. They are **failures of governance, responsibility, and moral imagination.**

Developing AI Systems Is Not Just Engineering

It is participation in social decision-making — one that determines who benefits, who is harmed, and whose voice is heard in the design process.

Responsibility Cannot Be Outsourced to the Algorithm

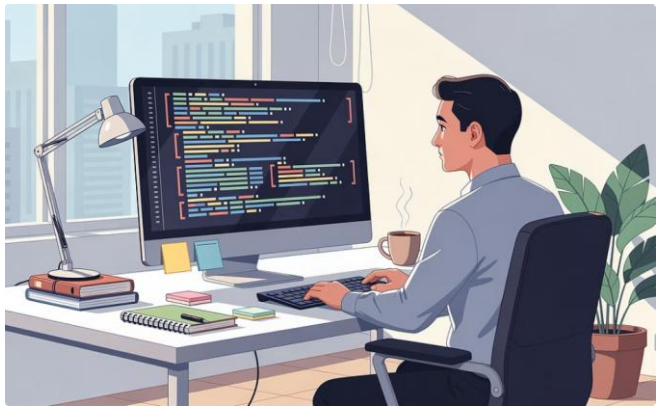
Delegating moral judgment to a system is itself a moral choice — and one that carries full accountability for its consequences.

AI Shapes the Future Distribution of Power

AI systems do not simply reflect society; they actively reconfigure who has access, influence, and opportunity in the years ahead.

Two Dimensions of Critical Capacity

To interpret, audit, and challenge algorithmic decisions effectively, we need both technical literacy and domain expertise. Without either dimension, critical capacity is incomplete — and accountability remains out of reach.



Technical Understanding

A working grasp of what AI is, how it operates, how models are trained, and where their limitations lie — even for non-specialists. This enables meaningful scrutiny of automated outputs.

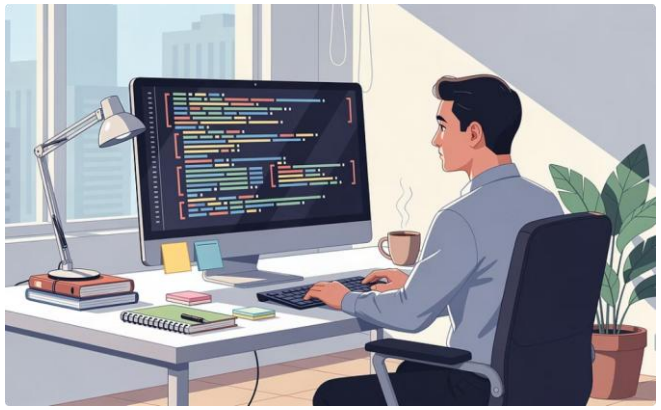
Problem Context

A deep understanding of the specific problem being solved — its social, ethical, and human dimensions. Without this, even technically sound systems can produce unjust outcomes.

Together, these two dimensions form the foundation for responsible AI governance — and neither can substitute for the other.

Two Dimensions of Critical Capacity

To interpret, audit, and challenge algorithmic decisions effectively, we need both technical literacy and domain expertise. Without either dimension, critical capacity is incomplete — and accountability remains out of reach.



Technical Understanding

A working grasp of what AI is, how it operates, how models are trained, and where their limitations lie — even for non-specialists. This enables meaningful scrutiny of automated outputs.

Problem Context

A deep understanding of the specific problem being solved — its social, ethical, and human dimensions. Without this, even technically sound systems can produce unjust outcomes.

Together, these two dimensions form the foundation for responsible AI governance — and neither can substitute for the other.

The Four Fundamental Risks of AI

Bias

Data reflects historical inequalities. AI doesn't merely mirror the world — it can **amplify** its injustices at unprecedented scale.

Opacity

Many AI systems are **"black boxes"** — decisions are difficult to explain, audit, or meaningfully challenge by those affected.

Uncritical Automation

Humans naturally trust systems. We often accept automated outputs **without questioning** them — surrendering judgment to the appearance of objectivity.

Scale

A small error, deployed at scale, becomes a **systemic failure** affecting thousands or millions of people — often before anyone notices.

Why Does AI Fail — Even When It "Works"?

You don't need to understand how it works technically to understand where it fails conceptually. Three ideas that cannot be ignored by anyone engaged in AI governance:

1 Bias Is Not an Exception — It's Structural

Data reflects the real world — with all its historical inequalities and patterns of exclusion. A system trained on biased data learns and reproduces those biases, even without intent or malice from its designers.

2 Opacity Limits Accountability

When we cannot explain how a decision was made, it becomes extremely difficult to hold anyone responsible. The black box protects the system — not the people affected by it.

3 Automation Creates Excessive Confidence

Consistent research shows that humans tend not to question automated systems, even when intuition would suggest otherwise. The appearance of objectivity is, itself, a powerful form of influence.

⊗ **A system can be technically correct and socially wrong.** This distinction is at the core of every serious ethical discussion about AI.

The Hardest Part Isn’t Understanding How It Works

It’s deciding when to trust.

The discussion about ethical AI doesn’t end with regulation, doesn’t end with technical auditing, and doesn’t end with training. It ends—or begins—each time a person consciously decides how much to delegate, to whom, and whether they are willing to accept the consequences when the system fails.

When to Trust

Trust in a system should be proportional to our ability to question it and the impact of its decisions.

How Much to Trust

Total trust is the abandonment of human judgment. AI is a tool—not an authority.

Who Takes Responsibility

Someone has to be accountable. If no one is, the system should not have been used that way.

> AI is not the problem. The lack of critical thinking about AI is.