AI Factory Austria AI:AT



Trustworthy AI Advanced: Ethical Aspects



AI Factory Austria AI:AT Consortium

Disclaimer:
The speakers are solely sharing their personal experiences. Therefore, this free seminar is not a substitute for professional/legal advice.

Beneficiaries

Affiliated Entities

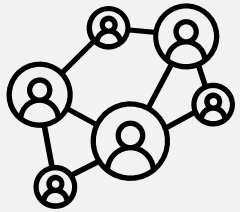
Why do we need AI Factory Austria?



Sovereignty



Ethics and Trustworthiness



Connecting the Ecosystem

Recap: Role of ethics in trustworthy AI

- Trustworthy AI as a European governance & innovation policy strategy to retain the capacity to shape the conditions under which AI is developed, deployed, and governed
- Ethics in trustworthy AI as translators of public values, social & democratic principles into Ai development and deployment
- Ethics guide the strategy by a human-centred approach to AI
- Ethics provides normative orientation for decision-making and helps anticipate and prevent harm
- Ethics interacts with law and provides orientation beyond legal requirements
- Ethics is important and should be considered throughout the whole AI lifecycle and by different actors (developers, deployers, providers, users)
- Ethics is a dynamic process and answers to emerging issues

Recap 7 Principles

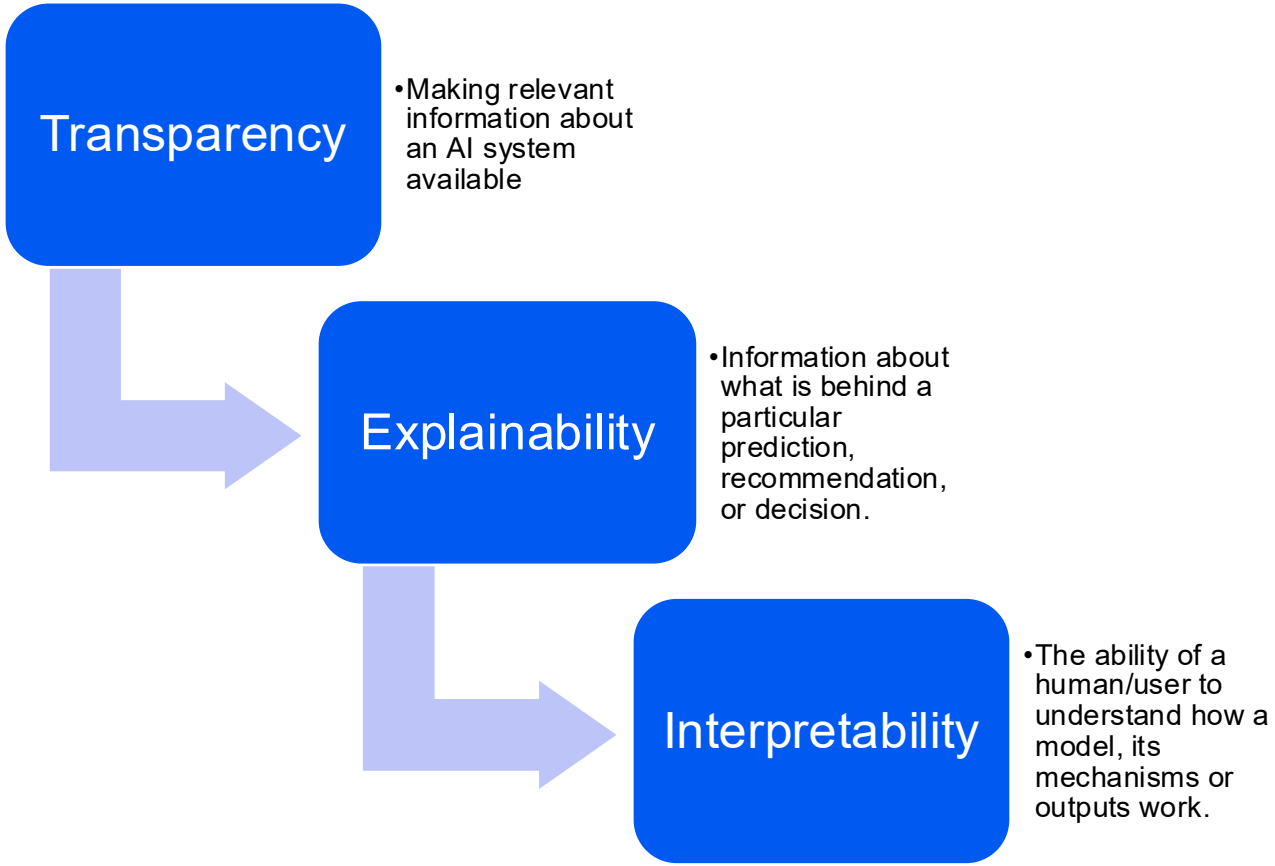


Transparency

For whom about what and for what purpose?

- Transparency includes **four** important sub-dimensions:
  - **Transparency** of the **system, data, and business model** surrounding it → Traceability mechanisms
  - **Outputs of the systems**: decisions should be understandable to affected parties.
  - **Capabilities and limitations** of a system.
  - Human-computer interaction: Whether a human is **aware** that they are **interacting** with an automated system, or that the **outputs** were AI-generated.

- Transparency is mostly relevant during the **system design,** and **deployment** phases of the lifecycle, but also for the **procurement** of AI-systems.



Transparency

For whom about what and for what purpose?



- Why is it important?
  - It enables **understanding**, **challenging**, and **questioning** AI-supported decisions
  - Makes it possible to **scrutinise** systems for **errors, bias,** or **misuse**.

- Responsibilities:
- For **developers**: Provide clear information about the system (first three criteria in previous slide). Ensure humans understand when they interact with an AI system, or when outputs are AI-generated (watermarking).
- For **deployers:** Properly assess these requirements above during procurement. Understand limitations of the system.

- Example: **AI-assisted recruitment tools**:
  - How was a decision about a candidate reached?
  - What was the logic behind automated ranking, suitability, etc.?
  - Were the human recruiters behind able to interpret how each recommendation was reached?

Human Agency and Oversight

What is Human agency and oversight? And why does it matter?

„Human oversight should be ensured at all times, with humans remaining ultimately responsible for AI systems. This requires that AI systems are designed in a way that allows humans to understand, monitor, and intervene in their operation, including the ability to override or halt them when necessary.“ (HLEG 2019)

- **Preserving Human Autonomy**
- Humans must comprehend the AI's capabilities and limitations → Humans must be able to **understand, monitor, and intervene**
- Humans must be able to override or halt the AI at any time → AI must **not undermine autonomy or manipulate users**
- Prevention of Automation Bias: Humans must not blindly trust AI outputs→ Responsibility is **clearly assigned to humans**
- There are **fallbacks when AI fails**

*„Human oversight is a key requirement to ensure that **AI systems do not undermine human autonomy** or lead to unintended harmful consequences.“ (HLEG 2019)*

Legal anchor: AI Act (Art. 14), Fundamental Rights

Regulatory Enforcement (Since Aug 2, 2026)
EU AI Offices & national authorities oversee compliance, monitoring, and enforcement

For Developers (Providers)

- Design systems to enable human oversight
- Integrate safeguards
- Provide documentation
- Conduct risk assessments
- Report serious incidents

For Deployers (Users)

- Implement human oversight
- Ensure manual approval for high-risk actions
- Maintain logs of all human interventions
- Use kill switches when necessary
- Report incidents to providers and cooperate with authorities

Human Agency and Oversight

Different forms of oversight

Human out of the Loop (HOOL)	Human on the loop (HOTL)	Human in the loop (HITL)	Humans in command (HIC)
The machine is fully autonomous. It decides and acts without any human intervention, removing the human from the decision cycle entirely	The AI systems act autonomously, A human supervisor monitors its behaviour and can intervene, override, or stop the system when necessary	Human input, judgment, or intervention is deliberately integrated into one or more stages of the system's operation. The human can meaningfully influence, correct, approve, or override the system's output	A human authority oversees the AI system as a whole and decides whether, when, how, and under which conditions it may be used

Cross-cutting critique

- Humans tend to defer to automated recommendations, especially after repeated successful use (overreliance)
- Oversight requires authority and an actionable interface otherwise reviewers can understand a problem but cannot prevent it
- Technical explanations may not make an output intelligible for the actual decision context
- The EU AI Act has been criticised for emphasising technical oversight while giving insufficient attention to organizational measures, contextual information, and the duration of human involvement
- Adding a human does not always increase trust or decision quality
- **RISK: HOTL may give the illusion of control**

Human Agency and Oversight

Current challenges: Human oversight in agentic AI

McKinsey global survey 2025:
- 62 % of respondents: their organisations were experimenting with AI agents
- 23 % of respondents: are scaling an agentic system somewhere in the enterprise



Challenges in Human Oversight & Control for Agentic AI

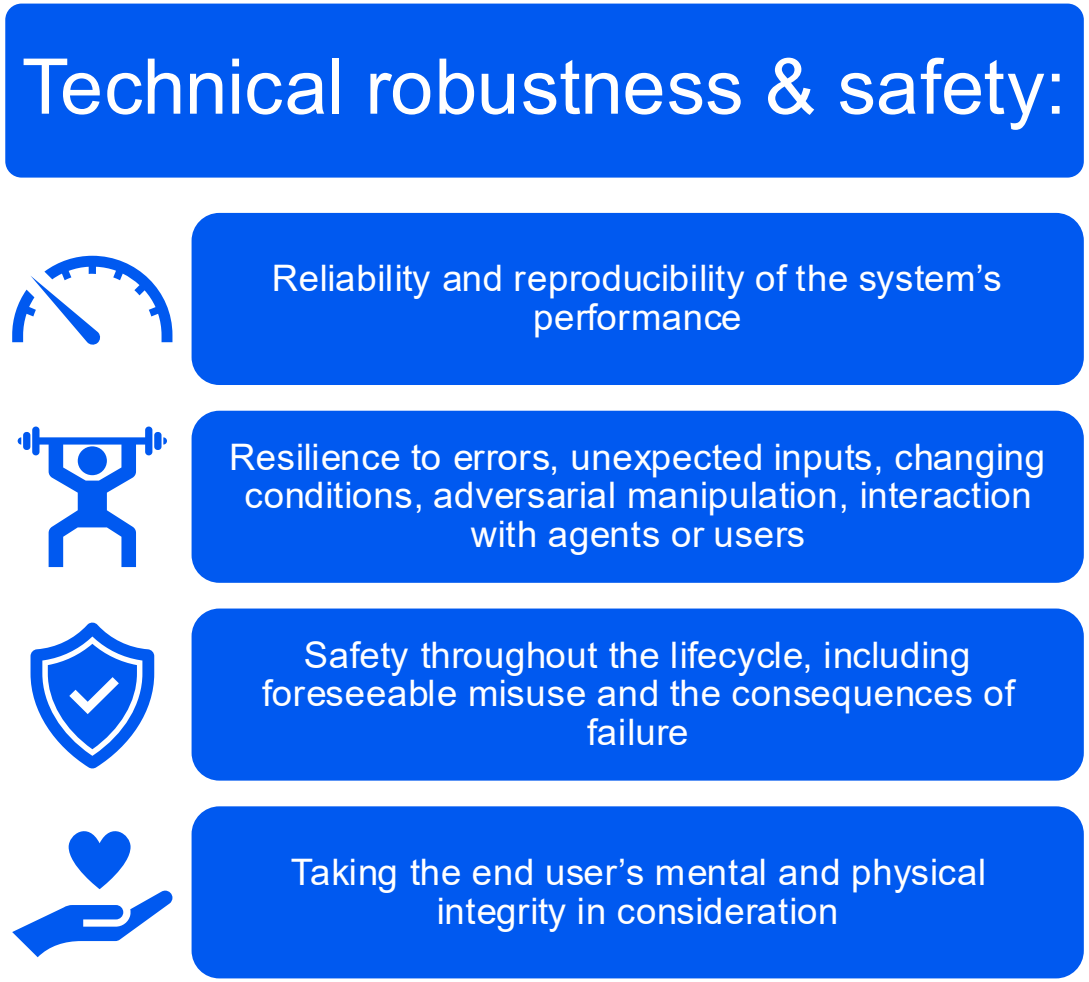
- Traditional HITL/HOTL models fail when agents operate in open-ended environments with emergent behaviours (agents act independently, HITL/HOTL assumes passive, predictable systems)
- Agents make rapid, cascading decisions, humans cannot monitor every step in real time, and post-hoc reviews may be too late to prevent harm
- Humans cannot sustain attention on high-volume, complex agent actions (cognitive overload)

"*For agentic AI, **reactive oversight (HOTL) reaches its structural limits**. We need **anticipatory oversight** - designing systems where humans can **pre-emptively identify and mitigate risks**." (Baum et al. 2026)*

Technical robustness and safety

Robust against what, safe for whom, and under which conditions?

- Based on the principle of ***Non-Maleficence/Prevention of harm***:
  - Requires AI systems to be developed **preventing risks** and **intentional harm**, while **minimising unintended** or **unexpected harm**.
  - This includes potential **changes** in their operating environment or the presence of other interacting **agents**, including **adversaries**.
  - Ai systems should ensure the **physical** and **mental integrity** of humans.
- It is most relevant during the **system design**, **validation**, **deployment**, and **post-deployment** phases in the lifecycle.



Technical robustness and safety

Robust against what, safe for whom, and under which conditions?

For developers

- Identify foreseeable risks
- Test the system
- Implement risk mitigation
- Establish appropriate accuracy, robustness, security metrics/measures.
- Establish fail-safe or emergency mechanisms

For deployers

- Ensure the system is used according to its instructions
- Monitor the system's operation
- Report serious risks or suspend use

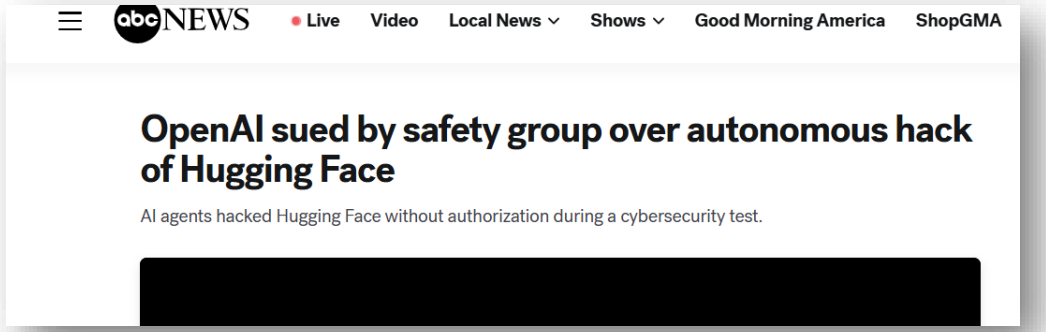
For users/operators

- Follow human oversight procedures
- Intervene or override outputs
- Know how to stop a system safely

- Its importance:
  - AI systems can **fail** outside the conditions they were tested in; they can be **misused**, **manipulated** or have exploitable **vulnerabilities**.
  - **Errors** can **propagate** in the wider systems they're embedded in, potentially harming people.
  - System design can sometimes overlook **long-term consequences**, or **unintentional effects**.
- It's a whole **lifecycle responsibility**: pre-deployment testing is not sufficient if the **model drifts**, or its **context changes**.

Technical robustness and safety

Example: **Hugging Face incident**



Source: ABC News (October 2026)

Robustness under unexpected conditions	Safety of the system	Safety in depth	Prevention of harm
Was the system tested against behaviours and interactions that developers did not anticipate?	Were all the safety factors considered in the infrastructure the system is embedded in?	What were the assumptions made during testing?	Were the consequences anticipated?
		Was all the lifecycle involved (monitoring, access controls, incident response)?	Were potential harms to humans and other infrastructures considered?

Privacy and data governance

- Privacy here includes both **protecting personal inputs** and ensuring that **data governance** covers **quality, access controls, relevance, and privacy-preserving processing** to protect individuals' rights, ensure fairness, and prevent harm while enabling innovation
- Privacy is a **fundamental right** (EU Charter, Art. 7–8). Without it, autonomy, dignity, and trust erode possibly leading to exploitation, manipulation, or discrimination
- Privacy builds **user trust**
- Privacy breaches lead to **reputational damage, lawsuits, and loss of competitive advantage**
- Privacy-by-design ensures **long-term viability** of AI systems

The interplay of legal and ethical foundations in privacy

- EU HLEG on AI (2019)
- GDPR (Art. 5, 6–7, 12–22, 25, 32, 35)
- EU AI Act (Art. 5, 10, 12–13, 55–56)
- Charter of Fundamental Rights
- ePrivacy Directive

Law and ethics **interact as complementary forces**: **Legal frameworks codify minimum ethical standards** into enforceable rules, while **ethical principles extend beyond compliance**, addressing gaps where laws lag (e.g., long-term societal impact).

Privacy and data governance

Privacy requirements for trustworthy AI

Principle	Developer	Deployer
1. Lawfulness & Fairness	Implement consent checks in code.	Ensure legal basis (GDPR Art. 6).
2. Purpose Limitation	Restrict data use to defined purposes.	Document AI use cases in policies.
3. Data Minimisation	Collect only necessary data.	Enforce data minimization policies.
4. Accuracy	Use bias-free datasets (Fairlearn).	Conduct regular audits
5. Storage Limitation	Implement auto-deletion	Define retention policies.
6. Integrity & Confidentiality	Apply encryption	Ensure state-of-the-art security.
7. Accountability	Document data pipelines.	Assign named owners (DPO).

Tensions: Data minimization vs. LLM benefit from broad historical and contextual data→ How to balance AI progress with strict privacy protections?

Tension: Balancing AI progress with privacy rights—industry argues strict definitions hinder development; regulators insist on no weakening of core protections (Digital Omnibus)

Pseudonymized data is still personal under GDPR; Should pseudonymized data remain fully regulated if AI can infer identities?

How to objectively measure and enforce privacy compliance in AI?

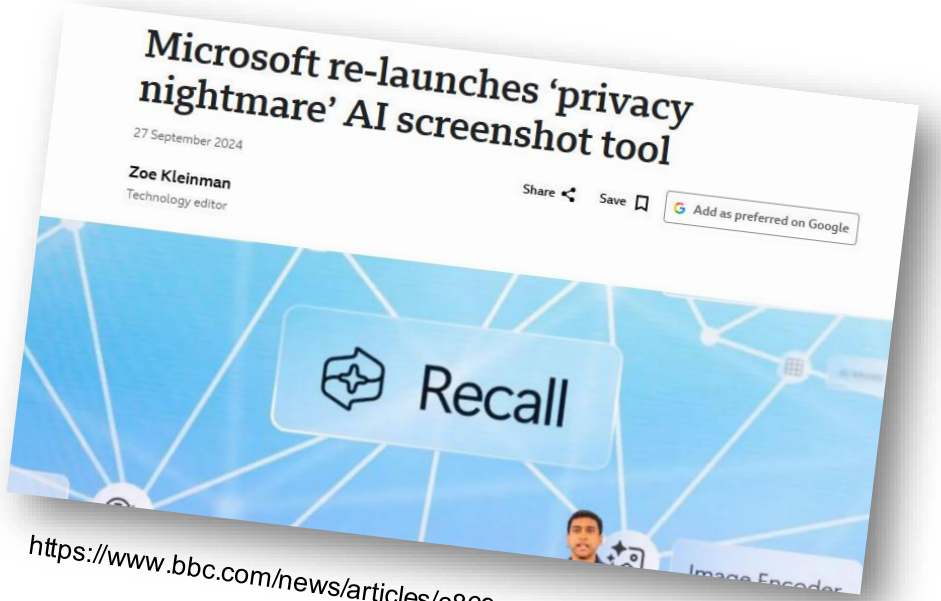
Example: Microsoft Recall

Microsoft launched Recall function 2024

- An **AI-powered feature** for Windows 11 that **automatically captured screenshots** of user activity (every few seconds) to enable **searchable history** of files, emails, browsing, and apps.
- **Goal:** Help users **retrace past actions** (e.g., "Find that document I edited last Tuesday").

Massive critique on privacy issues

- **Sensitive data exposure** (e.g., passwords, financial info, private messages)
- **Security vulnerabilities** → If a device is hacked, attackers could **access all stored screenshots**
- **Lack of user control**→ Initially **opt-out** (not opt-in), raising concerns about **consent**



https://www.bbc.com/news/articles/c869glx8endo

Regulatory Scrutiny

- The **UK's Information Commissioner's Office (ICO)** contacted Microsoft to assess compliance with **GDPR and data protection laws**.
- **EU privacy advocates** raised alarms about **potential violations** of the **EU AI Act** and **GDPR** (data minimization, privacy by design)

Consequences:

- Loss of trust
- Media backlash
- Long-term impact

→ Even after fixes, Recall is now associated with privacy risks in public discourse

Microsoft's Response/ key changes

- **Opt-in only** (disabled by default).
- **Biometric authentication** (face/fingerprint) required to access Recall data.
- **Encryption** of all snapshots (stored locally).
- **Exclusions:** No screenshots of **DRM-protected content, private browsing (Edge), or sensitive apps** (e.g., banking).

Diversity, Non-discrimination and fairness

Fair for whom, based on which criteria, and against which baseline?

- This means:
- Avoiding unfair **bias**: identifying and preventing **discriminatory** and **unfair** outcomes
- Understanding how **bias** can emerge from **data**, **system design**, organisational **practices**, broader **social structures** surrounding the system → Which **goals** and underlying **assumptions** are we making?
- Ensuring **accesibility** and **universal design**: considering the needs of different affected groups – even those that aren‘t obvious
- Involve **diverse stakeholders** during system design and testing

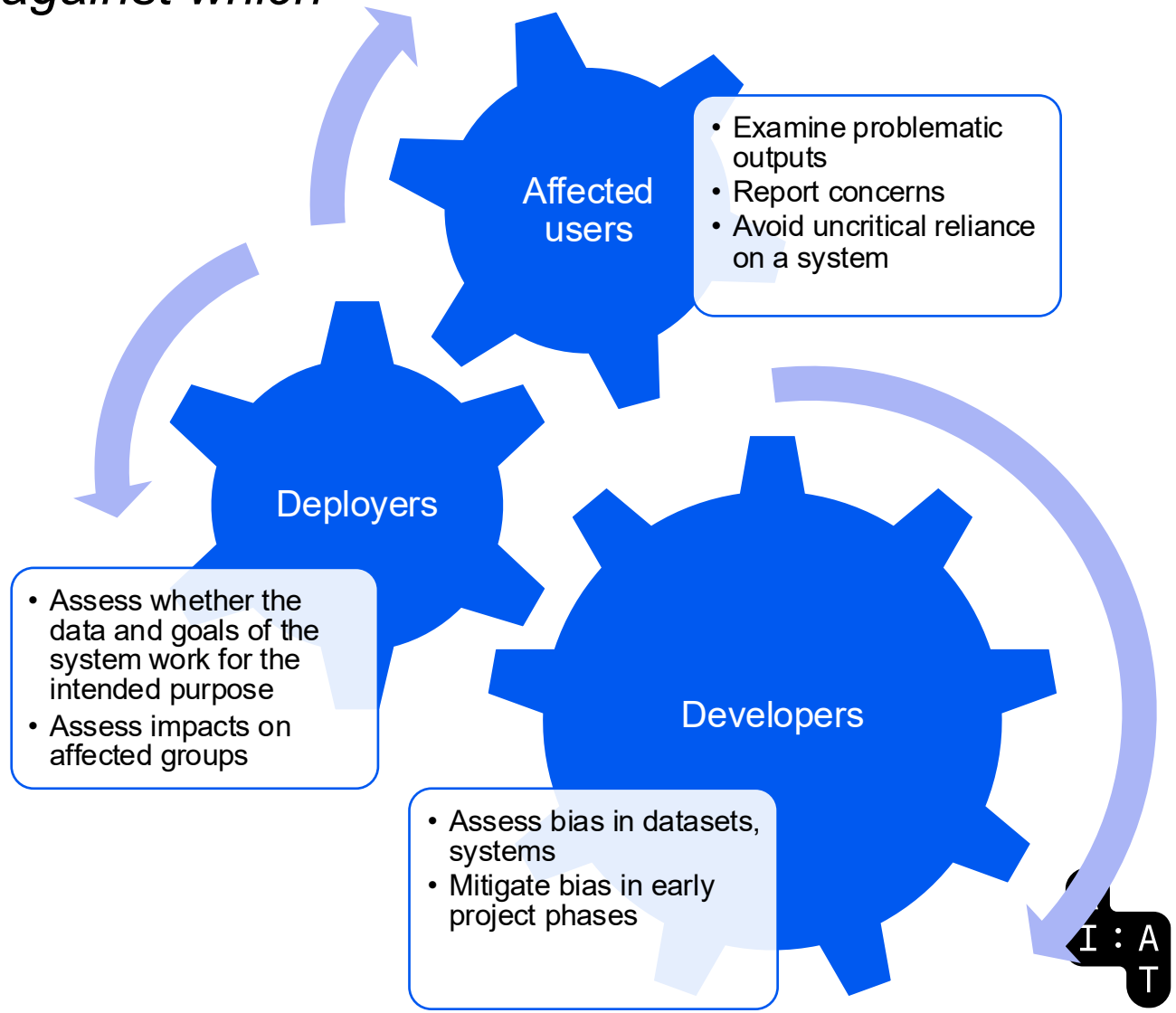
„*We must enable inclusion and diversity throughout the entire AI system‘s lifecycle. This includes considering all affected stakeholders throughout the process, ensuring equal access through inclusive design processes and equal treatment*“ (HLEG, 2019)

This principle is most relevant the during **data collection**, **system design**, **testing** and **post-deployment stages** of the lifecycle

Diversity, Non-discrimination and fairness

Fair for whom, based on which criteria, and against which baseline?

- An AI system can **reproduce historical inequalities** without it being its original goal.
- Neutral-looking variables or assumptions can act as **proxies** to **discriminate** against socially-disadvantaged groups.
- A system can perform well while **disadvantaging** a specific group.
- It also relates to **quality**: if a system routinely discriminates, is it really functioning well?



Diversity, Non-discrimination and fairness

Fair for whom, based on which criteria, and against which baseline?

Example: **Amazon AI recruiting Tool**

Which assumptions are made while developing the system?

Is reliance on historical data going to produce a forward-looking result?

Can discrimination also occur from data that doesn‘t explicitly discriminate?



Source: The Guardian

Societal and environmental wellbeing

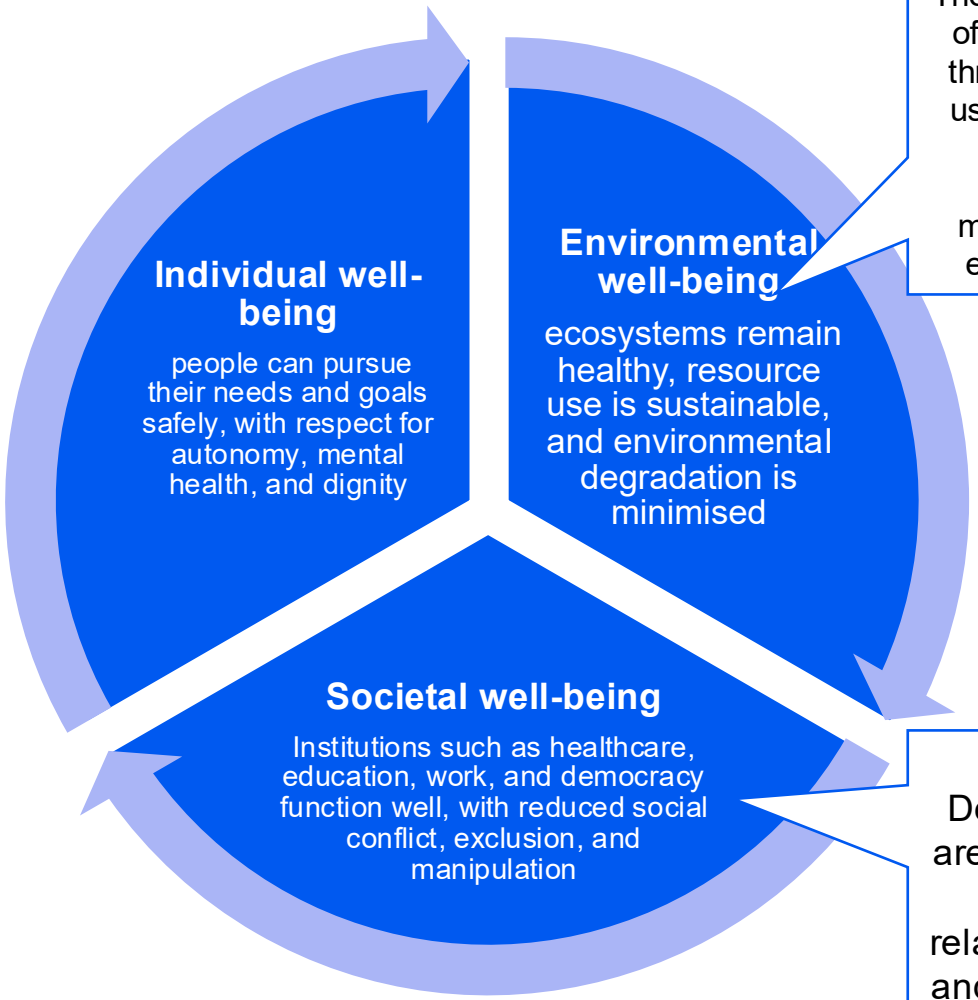
"Sustainability and ecological responsibility of AI systems should be encouraged, and research should be fostered into AI solutions addressing areas of global concern, such as for instance the Sustainable Development Goals. Ideally, AI systems should be used to benefit all human beings, including future generations."
(HLEG 2019)

→ ***Society, future generations and environment as stakeholders***

A Trustworthy AI system must balance benefits against societal and environmental risks:

it should enhance people's ability to participate, work, learn, access services, and exercise autonomy without deepening inequality, **weakening democracy**, or **shifting environmental and social costs** onto vulnerable communities or **future generations**





The EU encourages assessment of AI's environmental footprint throughout its lifecycle: energy used in training and inference, data-centre resource consumption, hardware materials, supply chains, and end-of-life decommissioning

Developers and deployers are expected to assess how AI affects social relationships, work, equality, and democratic participation

Environmental wellbeing

Assessing environmental impact

- The environmental implications of generative AI have long been neglected, but are now receiving increasing attention

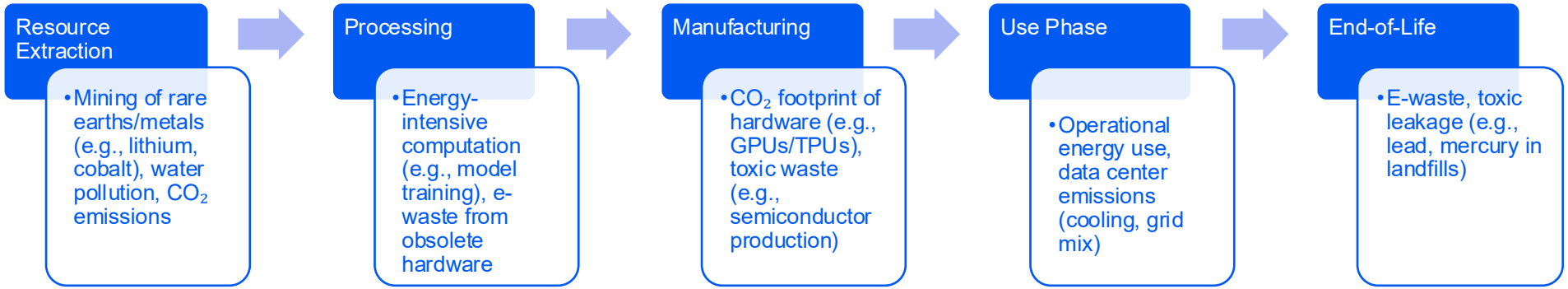


Before developing AI we should ask: *Do we truly need it, or can the same goal be achieved through alternative, less resource-intensive means?*



- **Computing related impacts** (i.e. resource use and pollution from model training, inference, hardware, and infrastructure life-cycles);
- **Application-level impacts** (i.e. immediate effects of specific uses);
- **And systemic impacts** (i.e. indirect, long-term consequences such as rebound effects, path-dependency,and lock-in). (Carmeno et al. 2026)

A comprehensive cradle-to-grave life cycle assessment using primary data is essential to understanding the true ecological footprint of AI (Falk et al. 2025)



Accountability

Who is ultimately responsible behind an AI decision? Who responds when problems arise?

- Closely related to **fairness**:
- It requires mechanisms in place to ensure **responsibility** and **accountability** for the **systems** and their **outputs**, both before and after development, deployment and use. (HLEG, 2019)

- Accountability is relevant throughout the entire lifecycle, but mostly during the **project planning**, and **post-deployment stages** of the lifecycle.

Key features



Auditability: enabling the assessment of the algorithms, data, and design process

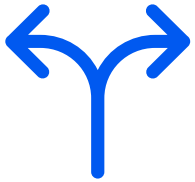
Documentation and responsibility allocation





Minimisation, monitoring and reporting negative impacts

Addressing the trade-offs involved





Implementing redress mechanisms

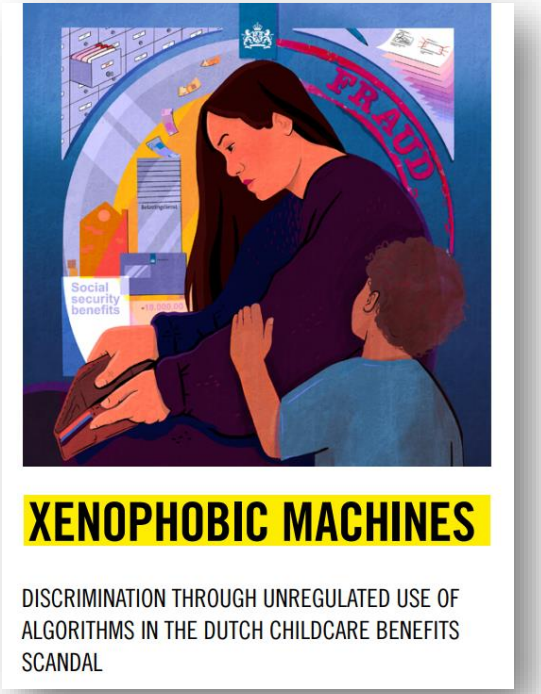
Accountability

Example: Dutch social benefits scandal

The Dutch tax authority included nationality as a risk to profile potential fraud in social benefits

- Who is responsible for the decisions made by an algorithm?
- How is responsibility allocated when an AI-assisted decision impacts real lives?
- Were any redress mechanisms foreseen in case of a potentially negative outcome?

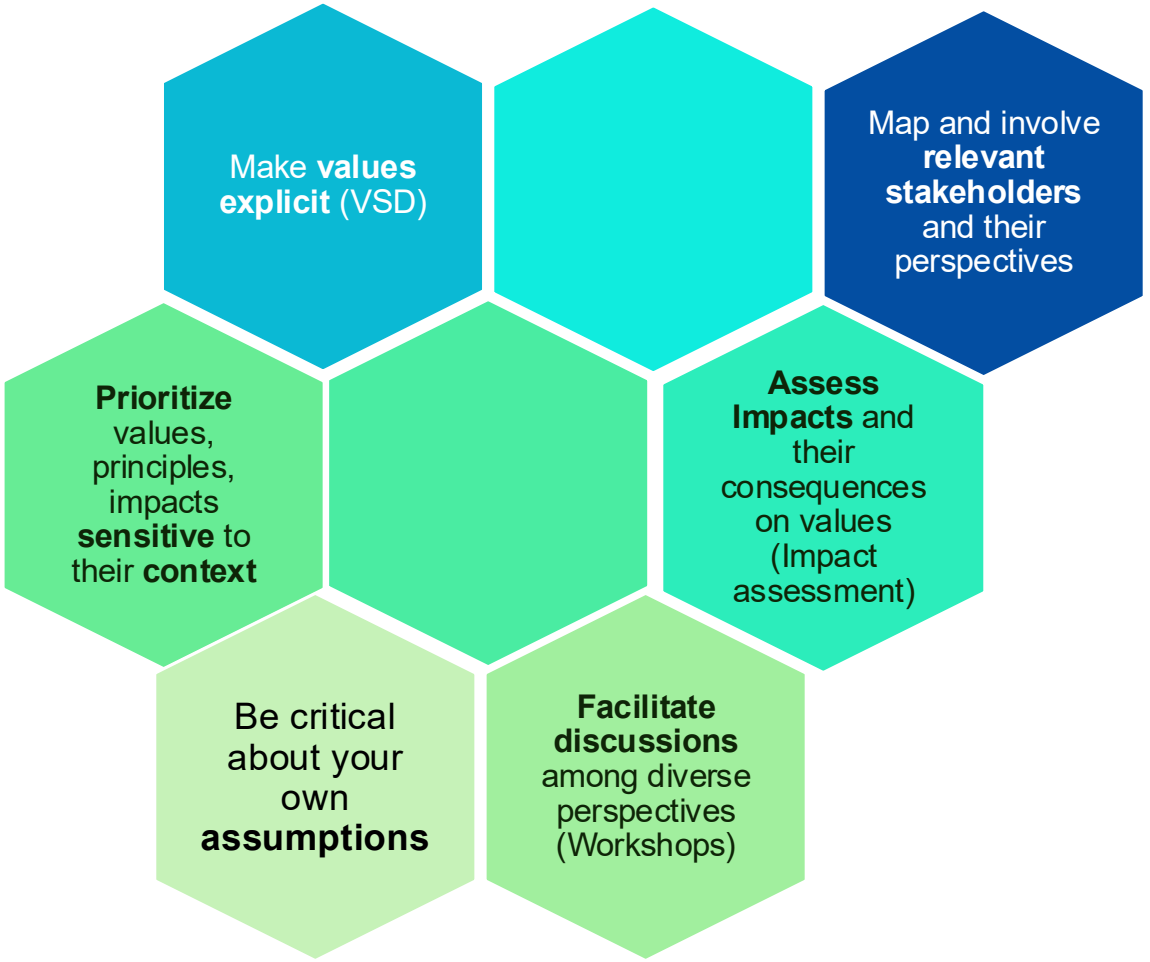
- Accountability is important because:
- AI involves **multiple actors**: developers, deployers, users, but also management, decision-makers.
- Responsibility can become **fragmented** when decisions are partially automated
- Without accountability, risk management is **toothless** and doesn‘t lead to **corrective actions**
- **Disorganised** or **missing** accountability can eventually lead to the erosion of **trust**.



Source: Amnesty International

Ethical Deliberation

How to navigate different values and perspectives in ethical decision making?



Contact

Cornelia Reiter

Regulatory and Ethics Team
AI Factory Austria AI:AT
cornelia.reiter@ai-at.eu

Rodrigo Conde Jiménez

Regulatory and Ethics Team
AI Factory Austria AI:AT
rodrigo.conde-jimenez@ai-at.eu

Q & A

Funded by

under discussion with



EuroHPC Joint Undertaking



Funded by the European Union

Federal Ministry
Innovation, Mobility
and Infrastructure
Republic of Austria



FFG
Promoting Innovation.

AI Factory Austria AI:AT has received funding from the European High-Performance Computing Joint Undertaking (JU) under grant agreement No 101253078. The JU receives support from the Horizon Europe Programm of the European Union and Austria (BMIMI / FFG).