Responsible AI – Part 2

Milton de Sousa

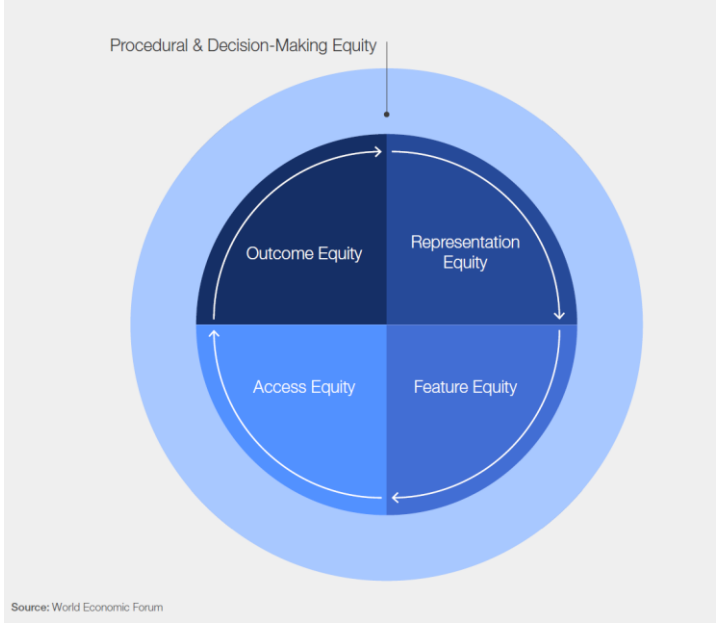
Managing AI responsibly

Ensuring Data Equity

Fairness / Equity

Outcome equity = impartiality and fairness in results. Maintain vigilance over unintended consequences that impact individuals or groups. Transparency, disclosure and shared responsibility are crucial to achieve fairness

Access equity = equitable accessibility of data and tools across varying levels of expertise. Address transparency and visibility issues related to model and data. Also encompasses disparities in terms of AI literacy and the digital divide.



Representation equity = seeks to enhance the visibility of historically marginalized groups within datasets while also accounting for data relevancy for the target populations.

Feature equity = seeks to ensure the accurate portrayal of individuals, groups and communities, necessitating the inclusion of attributes such as race, gender, location and income alongside other data.

Source: WEF (2023). Data Equity: Foundational Concepts for Generative AI. Briefing Paper.

Ensuring Data Equity – a multistakeholder perspective

Fairness / Equity

	Those responsible for driving and governing societal use of AI			Those using and impacted by AI systems		
	AI-creating organizations	AI-using organizations	Policy-makers and regulators	Civil society groups	Public	Communities
Focus areas	– Data collection and labelling – Data privacy and security – Transparency, traceability, and explainability – Mitigation strategies (incl. fairness and bias mitigation) – Continuous model evaluation – Inclusive model design	– Responsible AI practices – Data access and usage, incl. data privacy and security – Disclosure to impacted communities – Continuous monitoring – Mitigation strategies (incl. fairness and bias mitigation) – Context appropriate AI-human decision-making balance	– Develop ethical guidelines and standards[10] – Develop regulatory frameworks, including audits – Consideration of public interest – AI risk classifications – Clear delineation of rights of data subjects and contributors regarding AI – Raise public awareness	– Bridge gap between AI organizations and public by raising awareness through advocacy efforts – Promote ethical practices	– Increased awareness of AI – Understanding of AI ethics – Engagement with AI stakeholders	– Impact of AI on affected communities – Participation in AI decision-making discussions

Source: WEF (2023). Data Equity: Foundational Concepts for Generative AI. Briefing Paper.

Explainability

AI Explainability Dimensions

1. **Interpretability** = know how and why a model performed the way it did in a specific context and therefore the ability to understand the rationale behind its decision or behaviour. This sort of transparency is often referred to as '**opening the black box**' of AI.
2. **Transparency** = asks that the designers and developers of AI systems demonstrate that their processes and decision-making, in addition to system models and outputs, are sustainable, safe, fair, and driven by responsibly managed data.

Source: Adapted from Alan Touring Institute. AI Explainability in Practice

Types of Explanations

- **Outcome-based explanations** = include the components and reasoning behind model outputs while delineating contextual and relational factors.
- **Process-based explanations** = demonstrate that the AI project team has followed good governance processes and best practices throughout the AI project lifecycle.

Source: Adapted from Alan Touring Institute. AI Explainability in Practice

6 key AI explanations

Explainability

Outcome-based

- **Rationale Explanation**: the reasons that led to a decision outcome.
- **Data Explanation**: what data was used in a particular AI decision, as well as the data used to train and test the AI model.
- **Impact Explanation**: the considerations taken about the effects that the AI may have on an individual and society.

Process-based

- **Responsibility Explanation**: who is involved in developing and managing the model, and who to contact for a review of a decision.
- **Fairness Explanation**: the steps taken to ensure AI decisions are generally unbiased and equitable
- **Safety Explanation**: the measures and steps taken to maximise performance, reliability, security, and robustness of the AI outcomes, and the justification for the chosen AI system.

Source: Adapted from Alan Touring Institute. AI Explainability in Practice

Basic principles for sustainable AI

Energy Efficiency

- Minimize energy consumption in AI model training and deployment (small language models?)
- Use energy-efficient hardware and infrastructure
- Optimize resource usage through targeted, domain-specific models
- Employ specialized hardware and efficient cooling systems for data centers

Resource Management

- Reduce computational waste through model reuse and optimization
- Implement efficient cooling systems for data centers
- Consider water consumption impacts in AI operations
- Utilize renewable energy sources for AI infrastructure

Quick exercise: check what NVIDIA and AMD do to reduce processor power consumption!

Sustainability

CLIMATE / ENERGY / SCIENCE

Microsoft is going nuclear to power its AI ambitions

Satya Nadella, CEO of Microsoft, speaks during an interview in Redmond, Washington, on Wednesday, March 15th, 2023. Image: Chona Kasinger / Bloomberg via Getty Images

/ Microsoft is looking at next-generation nuclear reactors to power its data centers and AI, according to a new job listing for someone to lead the way.

By Justine Calma, a senior science reporter covering climate change, clean energy, and environmental justice with more than a decade of experience. She is also the host of Hell or High Water: When Disaster Hits Home, a podcast from Vox Media and Audible Originals.

Sep 26, 2023, 3:32 PM GMT+1

35 Comments (35 New)

Basic principles to ensure privacy (1)

Lawfulness, Fairness, and Transparency

- Organizations must process personal data legally and ethically. All data collection and processing activities must be clearly communicated to individuals using plain language. This includes having a valid legal basis for data processing, such as user consent (see GDPR standards in Europe).

Purpose Limitation

- Data should only be collected for specific, explicit, and legitimate purposes. Organizations cannot process data in ways incompatible with these original purposes, though exceptions exist for scientific research and statistical analysis.

Basic principles to ensure privacy (2)

Data Minimization

- Organizations must limit data collection to what is strictly necessary for their stated purposes. This means gathering only essential information rather than collecting additional "nice to have" data.

Accuracy

- Personal data must be kept accurate and up-to-date. Organisations must take reasonable steps to promptly correct or erase inaccurate information.

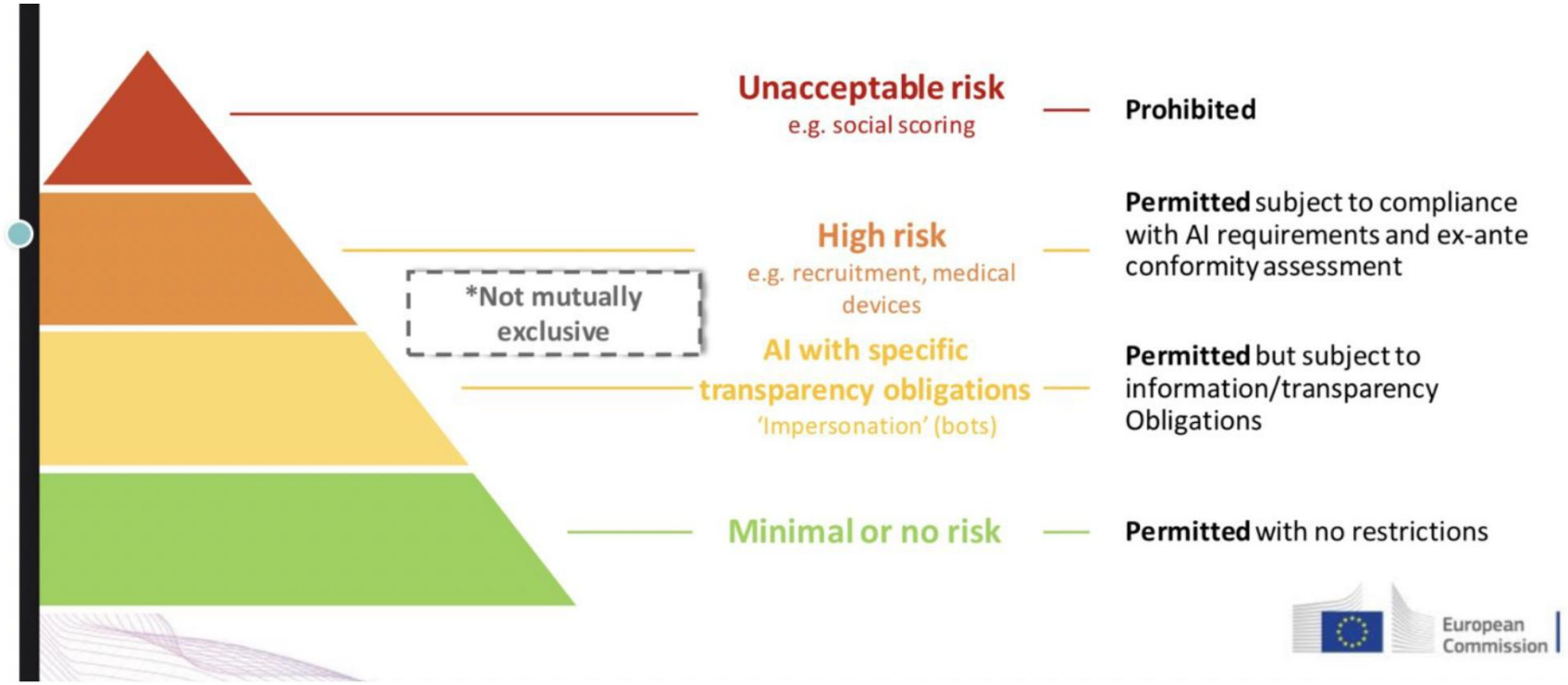
Storage Limitation

- Personal data should only be retained for as long as necessary to fulfil its intended purpose. Organisations must establish clear timelines for data deletion (in many countries mandatory) or review.

The EU AI Act

- On 12 July 2024, the final text of the EU AI Act was published in the Official Journal of the European Union.
- The final text combines a human-centric approach with a product-safety approach and is designed to establish a harmonized framework for AI regulation across the EU.
- The AI Act is a world first, setting a global precedent for AI regulation through its risk-based approach.

The EU AI Act


Unacceptable risk
e.g. social scoring
Prohibited
High risk
e.g. recruitment, medical devices
Permitted subject to compliance with AI requirements and ex-ante conformity assessment
*Not mutually exclusive
AI with specific transparency obligations
'Impersonation' (bots)
Permitted but subject to information/transparency Obligations
Minimal or no risk
Permitted with no restrictions
European Commission

Privacy

The four risk categories

A. Unacceptable Risk (Prohibited)

AI systems that:

- Manipulate human behavior
- Enable social scoring by governments
- Use real-time biometric identification in public spaces
- Exploit vulnerabilities of specific groups

B. High-Risk AI Systems (defined in Annex III of the Act)

Sectors include:

- Critical infrastructure
- Education and vocational training
- Employment and worker management
- Access to essential private and public services
- Law enforcement
- Migration and border control
- Administration of justice and democratic processes

C. AI with specific transparency obligations

- Require transparency
- Must disclose AI system usage
- Examples: Chatbots, AI-generated content

D. Minimal Risk AI Systems

- Most AI systems fall into this category
- Fewer regulatory requirements
- Encouraged to follow voluntary codes of conduct

Exercise

- Imagine you are an AI Auditor
- Select one large company = developer or heavy user of AI
- Explore their responsible AI practices based on the four dimensions
- Give a score from 1 to 5 on clarity and comprehensiveness per dimension
- First think for yourself, then use Chat GPT!

Some Generative AI trends

- From LLMs to Multimodal models => expected acceleration of productivity gains
- Explicability (e.g. GPT o1) = ensuring a better understanding of outcomes
- Agentic AI = enlarging automation potential
- Increasing regulation (EU AI Act)
- Small Language Models