WL
White Label Consultancy

Funded by
the European Union

AI4Gov
Gender and Ethical
Management

Session no.3

*The Impact of AI on Gender
Discrimination*

We asked ChatGPT...

Prompt: Create an image that represents women advocating for AI and the issue of gender discrimination.

Prompt: Create an image showing a diverse group of women interacting with AI, highlighting diversity and challenging gender discrimination.

With additional prompt

Do you notice any issues with this?

By representing only young and beautiful women, the first image fails to capture the full spectrum of diversity within the female population. Furthermore, this type of representation perpetuates discrimination by marginalizing women who are older, have disabilities, or don't conform to societal norms regarding appearance.
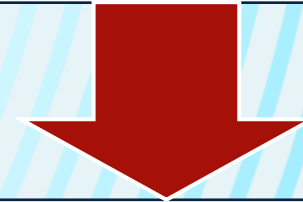
Through a second, more detailed prompt, the output offers a more inclusive and authentic representation of women.

The discriminatory output of the first prompt could only be identified and subsequently corrected by a **human observer**.

Bias vs Discrimination

Terminological clarification is necessary to precisely identify where discrimination originates, understand its mechanisms, and develop effective strategies to mitigate its harmful impacts and promote fairness in automated decision-making processes.

Bias and fairness	Discrimination and equality
✓ **Bias**: Systematic errors in algorithmic outcomes, encompassing various types of errors (statistical, cognitive, societal, etc.).	✓ **Discrimination**: Unequal treatment or disadvantage faced by specific groups, based on protected categories such as sex, race, disability, etc.
✓ **Fairness**: Ethical norms in algorithmic operations, including methods to prevent bias and ensure equitable outcomes.	✓ **Equality**: Legal principle guiding against discriminatory practices.

Algorithmic bias

Legal discrimination

Systematic errors in outcomes, potentially disadvantaging underprivileged groups, **not necessarily legally problematic**.

Adverse treatment based on protected categories, such as sex, race, disability, etc., with **legal implications under EU law**.

ALGORITHMIC BIAS

ALGORITHMIC DISCRIMINATION

✓ Algorithms may exhibit bias in different stages:

- Bias can originate from the **data used to train the algorithm**, especially if it under-represents a particular social group, or over-represents another.
- If the **metrics used to assess the performance of the algorithm** are not comprehensive or if they prioritize certain outcomes over others, it can lead to biased assessments of the algorithm's effectiveness.
- Once deployed, algorithms can continue to perpetuate biases if **feedback loops** are not carefully monitored and managed.
- *And so on…*

✓ Algorithmic discrimination refers to the **unjust treatment of individuals or groups based on the automated decisions made by algorithms**.

✓ This is the result of biases inherent in the algorithm, which, if not properly addressed, can lead to discriminatory outcomes, particularly affecting vulnerable groups.

ALGORITHMIC BIAS

Human oversight

ALGORITHMIC DISCRIMINATION

- Machine learning techniques interpret random occurrences as regular patterns, even when they might be unrelated. For this reason, **to mitigate the relationship between involuntary biases and their discriminatory effects on individuals, human involvement is crucial**.

- For instance, the 2019 Ethical Guidelines for Trustworthy AI underscore **human oversight** as one of the seven essential prerequisites for trustworthy AI.

- Despite its importance, relying solely on human oversight is not a universal fix. The "**tyranny of algorithms**" phenomenon highlights how humans may unquestioningly trust AI outcomes, leading to potential inaccuracies or unfairness. Thus, human oversight, when not combined with other measures, can exacerbate the negative impacts of AI systems.

So, is human intervention in automated decision-making processes beneficial or harmful?

Incorporating human oversight at every stage of the AI life cycle is not a one-size-fits-all solution; instead, it underscores an interesting paradox.

On the one hand, a human in the loop ensures that humans have oversight and control over AI systems, which can help prevent discrimination and bias.

On the other hand, it can reinforce existing human biases and further amplify discriminatory outcomes.

What other measures should be implemented?

✓ The 3rd Chapter of the Ethics Guidelines for Trustworthy AI features an Assessment List that evaluates AI systems based on seven key requirements:

- Human Agency and Oversight
- Technical Robustness and Safety
- Privacy and Data Governance
- Transparency
- Diversity, Non-discrimination, and Fairness
- Societal and environmental well-being
- Accountability

✓ The Human Agency and Oversight section aims to assess the organization's consideration of human involvement in the AI system. While crucial, this aspect must not be considered alone and must be complemented by adherence to all seven requirements for comprehensive protection against adverse AI effects.
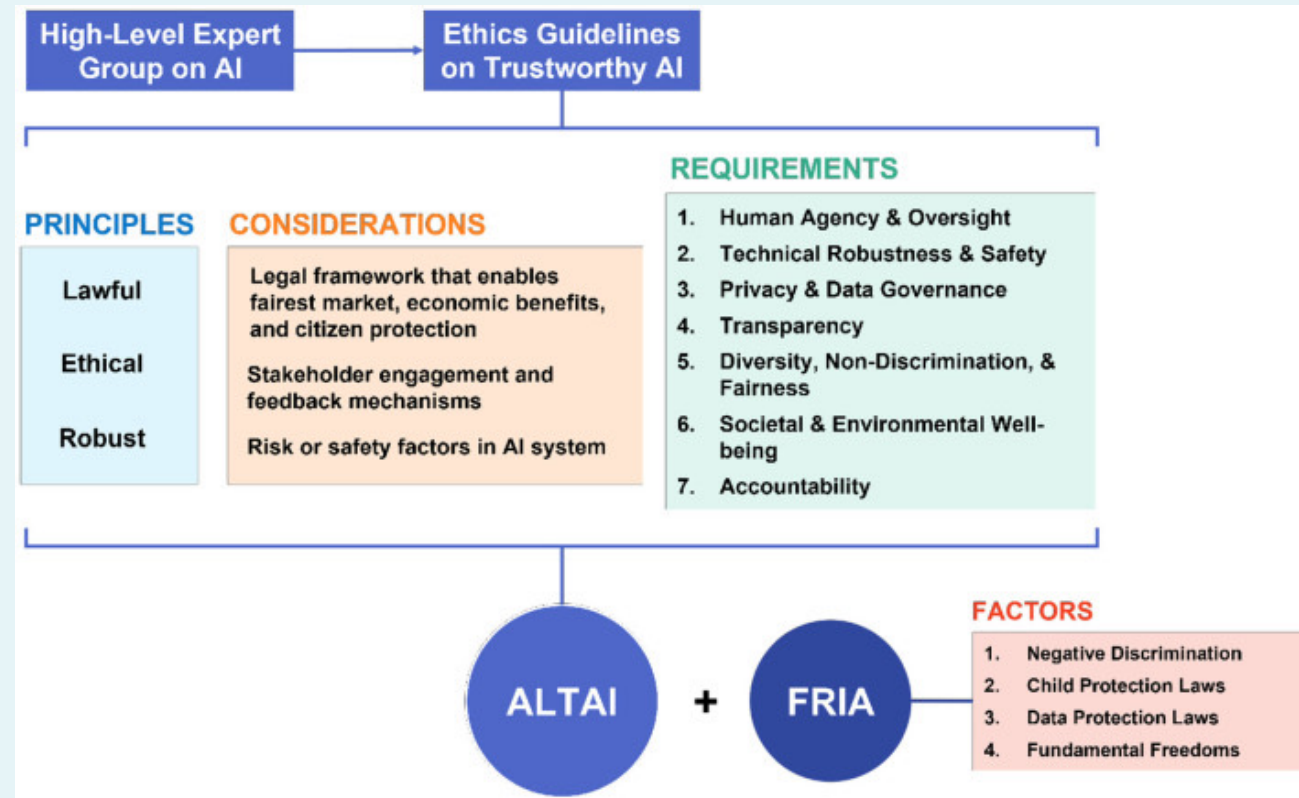
What will we discuss in the next and final session?

Step-by-Step: How to Perform a Fundamental Rights Impact Assessment	This session breaks down the step-by-step process of the FRIA, explaining how to assess the potential impact of AI technologies on fundamental rights, ensuring a comprehensive and systematic approach to ethical considerations and legal compliance.

Article 27 AI Act

Prior to deploying a high-risk AI system, **deployers shall perform an assessment of the impact on fundamental rights that the use of such system may produce**. This shall consist of consisting of:

a. A description of how the system will be used according to its intended purpose.
b. The period and frequency of the system's use.
c. The categories of people likely to be affected.
d. The specific risks of harm to these categories, based on information from the provider.
e. The implementation of human oversight measures.
f. Measures to address risks, including internal governance and complaint mechanisms.