Diversity, non-discrimination, and fairness

The fifth of the seven requirements on Ethics Guidelines for Trustworthy AI refers to “Diversity, Non-Discrimination, and Fairness”.

Unfair bias must be avoided, as it could have multiple negative implications from the marginalization of vulnerable groups to the exacerbation of prejudice and **discrimination**. Fostering **diversity**, AI systems should be **accessible to all, regardless of any disability**, and involve relevant stakeholders throughout their entire life cycle.

Figure 1.Diversity, non-discrimination, and fairness

This key requirement emphasizes the need for AI systems to **promote inclusivity, prevent bias, and ensure equitable outcomes**. Assessing this requirement involves multiple dimensions to ensure that AI systems are designed and operated in a manner that upholds these principles.

Bias Detection and Mitigation

One of the primary sources of discrimination in AI systems is biased data. Assessing this involves examining the datasets used for training AI models to **ensure they are representative of diverse populations.** In addition, algorithms themselves can introduce bias. Tools and methodologies for bias detection should be employed.

Regular audits and evaluations are necessary to ensure ongoing fairness as the system evolves.

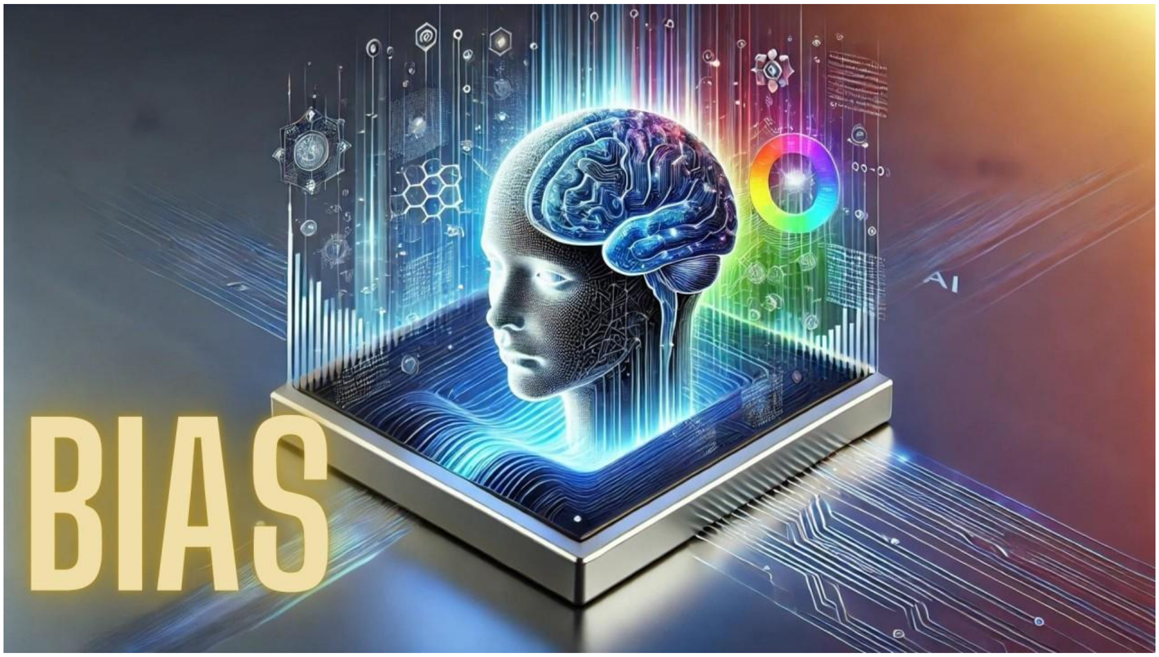

Figure 2: Bias Detection and Mitigation

Inclusive Design

Involving a **diverse group of stakeholders** (e.g., developers, end-users, experts, regulators, etc.) in the design and development process is crucial. This ensures that various perspectives are considered, and potential biases can be identified and addressed early.

On the other hand, **AI systems should also be designed to be accessible to people with different abilities.** This involves ensuring that interfaces are usable by people with disabilities and that the AI system's outputs are understandable by all users, regardless of their background. Standards like the Web Content Accessibility Guidelines (WCAG) provide a framework for assessing accessibility.

Fairness in Outcomes

Fairness is quite a challenging and dynamic dimension. Thus, it is important to monitor the outcomes of the decisions of an AI system to ensure that they do not affect any particular group disproportionately. To achieve this, **continuous monitoring and analysis of the decisions and outcomes produced by the AI**

system can be crucial to address unfair results. Any detected disparities should be investigated, and corrective measures should be implemented.

Furthermore, it is quite important to allow **end-users to provide feedback on the decisions and outcomes of the AI system**. Implementing feedback mechanisms can help identify instances of unfair treatment. This feedback should also be included in the investigation and analysis**, aligning corrective measures implemented towards fairer outcomes.**

On the other hand, **optimising AI models requires a grasp of the trade-off between improving generalisation without compromising fairness**, to ensure that AI systems are both equitable and effective.

Additional Transparency and Accountability Measures

Efforts to ensure diversity, non-discrimination, and fairness in AI systems should be **transparent to build trust and allow for actual and meaningful improvements**. This includes publishing reports on the fairness audits and assessments conducted, measures implemented to mitigate bias, and the outcomes of these efforts.

On the other hand, clear accountability structures should be in place to oversee the fairness of AI systems. This includes having designated roles or teams responsible for monitoring and ensuring compliance with diversity and fairness guidelines and processes. As this is most often an evolving process, processes should also be established for addressing grievances related to unfair treatment by AI systems.

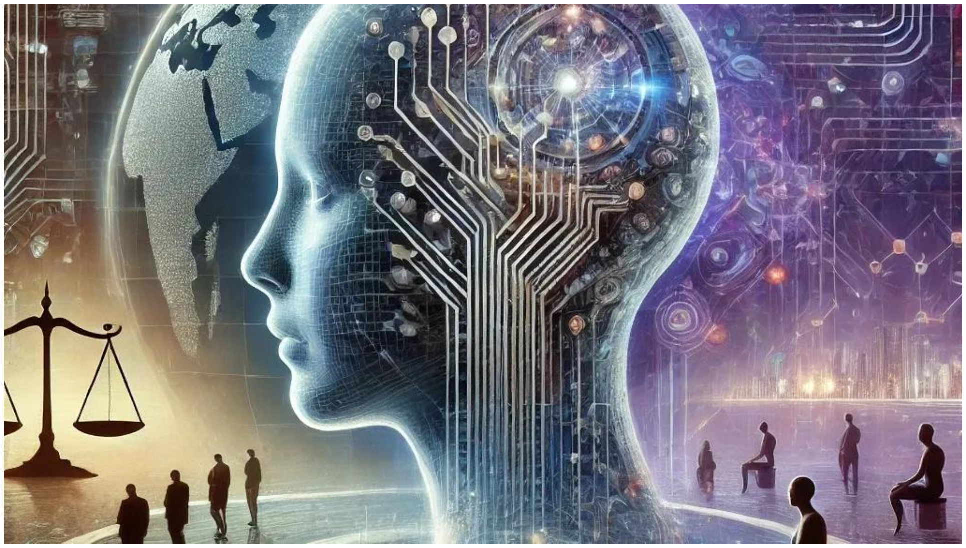

Figure 3: Transparency and Accountability Measures

Diversity, Non-Discrimination and Fairness in MANOLO

MANOLO addresses this key requirement early in the project's lifetime, by adopting **co-design and co-definition approaches for defining its benchmarking framework** (led by ATOS IT) **and its system architecture** (led by NUIDUCD-CeADAR). Both of these activities are fed by a landscaping exercise that involves desk research, interviews, and an online survey that engages several stakeholders, from developers and experts to regulators, policy makers, and citizens. All MANOLO partners are actively involved in these activities.

Focusing more on the technical dimension, **bias in data is addressed by the Data Quality Estimation methods** that will be developed (led by "NCSR Demokritos") that will allow **identifying biased and/or maliciously manipulated data**. This investigation will also be expanded to the models, which will be assessed by the MANOLO benchmarking framework (also led by "NCSR Demokritos").

In parallel, the adoption of the Z-Inspection® process will allow the analysis of the achieved results, in parallel with the design, development, and deployment activities, under the guidance of the Arcada.

Wrap Up

Assessing the requirement of "Diversity, Non-discrimination, and Fairness" in AI systems involves a comprehensive approach that includes bias detection and mitigation, inclusive design practices, internal and external monitoring of outcomes, compliance with standards/guidelines, and ensuring transparency and accountability.

By systematically addressing these areas, it is possible to ensure that AI systems promote fairness and equity, thereby fostering trust and inclusivity in their deployment and use.

In MANOLO, these principles are integrated from the start. The project uses co-design approaches for defining frameworks and system architecture, involving diverse stakeholders through desk research, interviews, and surveys. **Data and model bias are addressed throughout the project's lifecycle, ensuring that MANOLO's AI systems are fair and inclusive.**