AI Fairness-by-Design Multi-Stakeholder Methodology

-

A Comprehensive Framework for Fair AI Design and Development

AEQUITAS
unbias AI

ANNEX VI: Fair Outcomes Interpretation Methodology (FOIM)

Fair-by-Design Sub-methodology

Abbreviation	Meaning
AFF	Affectees
AIU	AI Users
DDM	Development Decisionmakers
DE	Domain Experts
EGTAI	Ethics Guidelines for Trustworthy AI
FbD	Fair-By-Design
FDCGM	Fair Data Collection, Governance, and Management
FMM	Fair Model Methodology
FOIM	Fair Output Interpretation Methodology
FRIA-F	Fundamental Rights Impact Assessment for Fairness
GDM	Governance Decisionmakers
SIM	Stakeholder Identification Methodology
TAIRA	Trustworthy AI Readiness Assessment
MAP	Multistakeholder Approach to AI Fairness-by-Design
ML	Machine learning
NLP	Natural Language Processing

AEQUITAS
unbias AI

Contents

Introduction

The Fair Interpretation Methodology aims to clarify how AI users interpret the outcomes of the system. The goal is to assist AI users/deployers in critically assessing the results of the system before deployment for a fair implementation.

It is paramount for users to assess the algorithmic outcomes before implementing them to keep oversight over the decision-making process, depending on the context where the system is implemented. Considering that AI is already developed in diverse fields and for multiple tasks, this methodology brings a holistic approach to assist users across all disciplines.

The interaction between humans and machines starts from the design of the AI system and should be maintained across its lifecycle stages to sustain human oversight at all levels. In the previous phases of the Fair by Design Methodology, namely, scoping, risk analysis, and development, WP6 provided guidelines and methodologies to mitigate technical, ethical, legal, and social constraints. Therefore, this methodology aims to focus on how humans interpret and use algorithmic outcomes.

The socio-ethical and legal necessity for Human Oversight

The level of autonomy of AI systems varies based on the way they are designed and the task they are serving. Regardless, **human oversight** needs to be maintained at all levels to control the system and intervene when needed. This close interaction between humans and machines enables the detection of **technical**, **ethical**, **legal**, and **social** constraints in the operations of the system and provides routes to redress such risks. **Human-in-the-Loop** (HITL) and **Human-on-the-Loop** (HOTL) are two approaches that describe how humans interact with AI systems at different stages of their operation. **Human-in-the-Loop** (HITL) refers to a scenario where humans are actively involved in the decision-making or operational process of an AI system. They provide input, intervene, or make judgments that directly influence the AI's outcomes. This approach is essential in the development stage of the AI lifecycle where human oversight ensures accuracy, fairness, and ethical considerations in the system's outputs.[1] **Human-on-the-Loop** (HOTL) builds upon HITL but involves humans in a more supervisory capacity.[2] It comes into play when an AI system has achieved a certain level of performance but still requires periodic human feedback and oversight to refine and improve its functionality. In this setup, humans monitor the system's performance, stepping in as needed to adjust or enhance its operations, ensuring that it continues to perform reliably and ethically. This is directly relevant to AI users who will be interacting with the system at its last lifecycle stages and making sure its deployment is legally, socially, and ethically sound.

Human oversight is one of the principles mentioned in the Ethics Guidelines for Trustworthy AI[3], the UNESCO Recommendation on the Ethics of Artificial Intelligence[4], and the Council of Europe Framework Convention on Artificial Intelligence and human rights, democracy and the rule of law[5]. The EU AI Act also emphasizes the importance of human oversight in high-risk AI systems to mitigate potential risks to health, safety, or fundamental rights. In Article 14, it outlines several key obligations for AI deployers and users. First, individuals responsible for oversight must **understand** the capabilities and limitations of the AI systems to monitor their operations effectively and address any anomalies. Second, Oversight personnel should remain cognizant of the risks associated with automatic reliance on AI outputs, especially

when those outputs inform critical decisions. It is also crucial for users to accurately interpret the AI system's results, utilizing available tools and methods to aid in understanding. Users should furthermore retain the right to **disregard or override AI outputs** based on the context of specific situations. Effective mechanisms, such as a **'stop' button** or other procedures, should be in place to safely halt the AI system if necessary. Overall, the Act reinforces the need for critical human involvement in managing AI technologies to ensure safety and accountability.

Self-identifying interpretation biases

Overview of the system

Before implementing the system, AI users (AIU) should understand and decrypt the results for known and foreseeable biases. To do so, WP6 designed the following questions that AI deployers can follow for a fair interpretation process.

Based on the Overarching methodology, AIU will have access to the Trustworthy AI and Technical Leaflet to retract this information accordingly.

A. **Purpose Identification:**		
Question 1	Is the system general-purpose or single purpose?	
Answer	☐ Single Purpose	☐ General Purpose
Implications		
Question 2	What specific task or range of tasks is it intended to perform?	
Answer		
Question3, If single purpose	Will you use the system to describe, predict, or prescribe solutions?	
Answer	☐ Descriptive purpose ☐ Prescriptive purpose ☐ Predictive purpose	
Question 4	Are you using the system in alignment with its intended purpose?	
Answer	☐ Yes	☐ No
Implications		Reconsider the way you are using the system and the inaccuracy that may arise when deployed in another context.
B. **Data Source and Representation:**		
Question 1	Is the training data representative of the target population or/and use case?	
Answer	☐ Yes	☐ No
Implications		
Question 2	Does the data account for temporal shifts, geographic diversity, or demographic variance?	
Answer	☐ Yes	☐ No
Implications	Explain which are accounted for.	Data bias happens when data doesn't represent all the groups or scenarios the system might encounter. Keep these shortcuts in mind when you consider the outcomes of the

		system to not discriminate against any group.
C. **Outcome Evaluation**		
Question 1	How is success measured in the system's outcomes?	
Answer		
Implications	Explain if this success matrix is comprehensive to the context and domain where the system is used.	
Question 2	Have historical biases, or others, been identified and accounted for in the design?	
Answer	☐ Yes	☐ No
Implications	Make sure all historical biases were addressed depending on the context and country where the system is used.	Past injustices or errors in data can be carried over into AI systems, causing unfair outcomes. List the perceived historical biases in the context, domain, and county where the system is used to dodge a social reproduction of unfairness.
Question 3	Are you able to provide direct feedback to the AI system?	
Answer	☐ Yes	☐ No
Implications		Contact the AI provider to discuss channels for feedback incorporation.
Question 4	Are there mechanisms to control the system outcomes including the 'stop' button?	
Answer	☐ Yes	☐ No
Implications		Contact the AI provider to discuss control mechanisms.

Questionnaire to self-assess bias:

Based on the literature review, AI users could incorporate cognitive biases leading to overreliance on the AI system, misapplying explanations, or under-reliance. These elements can lead to a variety of issues. For instance, if the AI system is deployed to calculate the solvability of customers requesting loans, bank agents must keep a critical eye when decrypting the system's results before deciding to allocate the loan. For instance, if they over-rely on algorithmic results, they may be biased against customers from ethnicities or genders not comprehensively represented in the training data, leading to unfair treatment.

Once AI users get an overview of the system, its purpose, and mitigated bias risks, they will dive deeper into the biases that may occur during the deployment of the system. To assist AIUs in this process, WP6 designed a set of questions that can be followed. Each question has room for a binary answer then a section to detail the implications of each. The questions go as follows:

A. Consumer Bias	This bias happens when users approach a system with pre-existing beliefs (e.g., expecting certain outcomes based on stereotypes).
Question 1	Am I assuming the algorithm's output confirms my expectations or beliefs without verifying its validity?

Answer	☐ Yes	☐ No
Implications	*Which ones and how will they be managed*	
B. Mode Confusion Bias	It's essential to understand how the system operates in its different modes.	
Question 1	Am I clear on whether this outcome was generated by a manual process, an automated process, or a mix of both?	
Answer	☐ Yes	☐ No
Implications	*Precise which one.*	*Check the Technical leaflet or other shared documentation by the AI developers/ providers.*
Question 2	Could my misunderstanding of the system's mode influence how I evaluate this result?	
Answer	☐ Yes	☐ No
Implications	*Precise how and keep track of these points when implementing the system.*	
C. Cognitive Bias	Mental shortcuts may lead to snap judgments about outcomes without deeper analysis. Here is a list of 50 cognitive biases to explore for a more representative and comprehensive decision-making process.	
Question 1	Am I jumping to conclusions about the outcome based on how familiar or simple it seems? Have I taken the time to evaluate the reasoning behind this result?	
Answer	☐ Yes	☐ No
Implications	*List the identified assumptions and how they will be managed.*	
D. Dunning-Kruger Effect	Overestimating your knowledge can lead to misinterpretation of results.	
Question 1	Am I confident in my understanding of the system's limitations, or might I be overestimating my expertise?	
Answer	☐ Yes	☐ No
Implications		
Question 2	Have I consulted additional resources or experts if the outcome is unclear?	
Answer	☐ Yes	☐ No
Implications	Clarify which experts were contacted, what was not clear in the outcomes, and how did they help.	Consult experts regarding the unclear scope to make sure the interpretation is accurate.
E. Loss of Situational Awareness	Overreliance on algorithms can make you overlook important real-world context.	
Question 1	Am I balancing this algorithmic output with my understanding of the broader situation? Could I be missing critical context that affects the outcome's relevance or accuracy?	
Answer	☐ Yes	☐ No
Implications	List and redress the identified unbalances.	Double-check if the use case and field where the AI system is implemented are aligned with the outcomes of the system.
F. Rashomon Effect	Different perspectives can lead to varying interpretations of the same outcome.	
Question 1	Am I considering how someone else might interpret this outcome differently?	
Answer	☐ Yes	☐ No

Implications	Explain how this was considered and if/how did it shift the outcome interpretation.	Consider putting yourself in another person's shoes to explore if different interpretations are possible.
G. Selective Adherence	It's tempting to focus only on parts of the result that confirm your beliefs.	
Question 1	Have I critically evaluated the entire output, even if some parts contradict my beliefs?	
Answer	☐ Yes	☐ No
Implications	Explain which ones and how this altered the outcome interpretation.	Consider critically assessing your acceptance of the outputs and if they reinforce/align with your beliefs.
H. Presentation Bias	How results are displayed (e.g., rankings, graphs) can unconsciously influence decisions.	
Question 1	Am I focusing on the substance of the result rather than how it is presented?	
Answer	☐ Yes	☐ No
Implications		
Question 2	Am I aware of how the format or layout of this outcome might be affecting my interpretation?	
Answer	☐ Yes	☐ No
Implications		

Impact Assessment before Deployment:

A. AI Risk Level	Assess which risk level following the EU AIA compliance checker	
Answer	☐ Unacceptable Risk ☐ High-Risk ☐ Limited-Risk ☐ Minimal-Risk	
Implications	If Unacceptable Risk, don't deploy the system.	If High, Limited, or Minimal-Risk, follow the next questions.
B. Impact of the system	Does the AI system have a **direct negative impact** on:	
1. Impact on Health	☐ Health	
Implication	If yes, explain how and on which group	
Question	If yes, can you mitigate this risk?	
Answer	☐ Yes	☐ No
Implications	Detail the mitigation tools and explain if the residual risk is acceptable.	Stop deployment, contact authorities, and reach out to the provider.
2. Impact on Safety	☐ Safety	
Implication	If yes, explain how and on which group	
Question	If yes, can you mitigate this risk?	
Answer	☐ Yes	☐ No
Implications	Detail the mitigation tools and explain if the residual risk is acceptable.	Stop deployment, contact authorities, and reach out to the provider.
Impact on Fundamental Rights		
3. Impact on dignity[i]	☐ Dignity	

Implication	If yes, explain how and on which group	
Question	If yes, can you mitigate this risk?	
Answer	☐ Yes	☐ No
Implications	Detail the mitigation tools and explain if the residual risk is acceptable.	Stop deployment, contact authorities, and reach out to the provider.
4. Impact on Freedoms[ii]	☐ Freedoms	
Implication	If yes, explain how and on which group	
Question	If yes, can you mitigate this risk?	
Answer	☐ Yes	☐ No
Implications	Detail the mitigation tools and explain if the residual risk is acceptable.	Stop deployment, contact authorities, and reach out to the provider.
5. Impact on Equality[iii]	☐ Equality	
Implication	If yes, explain how and on which group	
Question	If yes, can you mitigate this risk?	
Answer	☐ Yes	☐ No
Implications	Detail the mitigation tools and explain if the residual risk is acceptable.	Stop deployment, contact authorities, and reach out to the provider.
6. Impact on Solidarity[iv]	☐ Solidarity	
Implication	If yes, explain how and on which group	
Question	If yes, can you mitigate this risk?	
Answer	☐ Yes	☐ No
Implications	Detail the mitigation tools and explain if the residual risk is acceptable.	Stop deployment, contact authorities, and reach out to the provider.
7. Impact on Citizens' Rights[v]	☐ Citizens' rights	
Implication	If yes, explain how and on which group	
Question	If yes, can you mitigate this risk?	
Answer	☐ Yes	☐ No
Implications	Detail the mitigation tools and explain if the residual risk is acceptable.	Stop deployment, contact authorities, and reach out to the provider.
8. Impact on justice[vi]	☐ Justice	
Implication	If yes, explain how and on which group	
Question	If yes, can you mitigate this risk?	
Answer	☐ Yes	☐ No
Implications	Detail the mitigation tools and explain if the residual risk is acceptable.	Stop deployment, contact authorities, and reach out to the provider.

Conclusion

In summary, the Fair Interpretation Methodology serves as a vital framework for ensuring that AI users engage with and understand the outcomes generated by AI systems. As AI systems continue to influence a broad array of sectors, it is essential for users to critically evaluate the results before implementation to uphold fairness and ethical practices. The methodology outlined emphasizes the importance of maintaining human oversight throughout the AI lifecycle, including during deployment.

By recognizing the diverse levels of autonomy within AI systems, users can better assess the implications of these technologies on decision-making processes. The integration of ethical guidelines and legal frameworks further reinforces the necessity of human involvement to mitigate risks associated with AI outputs. Moreover, fostering an awareness of interpretation biases through structured questionnaires allows users to enhance their ability to understand and appropriately respond to the findings of AI systems.

Ultimately, the commitment to critical human interpretation and oversight not only promotes accountability and safety but also helps to align AI implementations with socio-ethical values, legal frameworks, and technical robustness. As the landscape of AI continues to evolve, ongoing education and adaptation of these methodologies will be paramount for responsible AI deployment.

[i] Right to life; Right to the physical & mental integrity of the person; Prohibition of torture and inhuman or degrading treatment or punishment; Prohibition of slavery and forced labour. (Charter of Fundamental Rights of the EU)
[ii] Right to liberty and security; Respect for private and family life; Protection of personal data; Right to marry and right to found a family; Freedom of thought, conscience and religion; Freedom of expression

and information; Freedom of assembly and of association; Freedom of the arts and sciences; Right to education; Freedom to choose an occupation and right to engage in work; Freedom to conduct a business; Right to property; Right to asylum; Protection in the event of removal, expulsion or extradition. (Charter of Fundamental Rights of the EU)

[iii] Non-discrimination; Cultural, religious and linguistic diversity; Equality between women and men; The rights of the child; The rights of the elderly; Integration of persons with disabilities. (Charter of Fundamental Rights of the EU)

[iv] Workers' right to information and consultation within the undertaking; Right of collective bargaining and action; Right of access to placement services; Protection in the event of unjustified dismissal; Fair and just working conditions; Prohibition of child labour and protection of young people at work; Family and professional life; Social security and social assistance; Health care; Access to services of general economic interest; Environmental protection; Consumer Protection (Charter of Fundamental Rights of the EU)

[v] Right to vote and to stand as a candidate at elections to the European Parliament; Right to vote and to stand as a candidate at municipal elections; Right to good administration; Right of access to documents; Right to petition; Freedom of movement and of residence; Diplomatic and consular protection. (Charter of Fundamental Rights of the EU)

[vi] Right to an effective remedy and to a fair trial; Presumption of innocence and right of defence; Principles of legality and proportionality of criminal offences and penalties; Right not to be tried or punished twice in criminal proceedings for the same criminal offence. (Charter of Fundamental Rights of the EU)