

Appendix B: Use case S2 detailed findings

B.1 Trustworthy AI Readiness Assessment

The Trustworthy AI Readiness Assessment serves as the initial component of a comprehensive Multi-stakeholder Assessment Methodology for AI Fairness, developed as part of the AEQUITAS project's comprehensive approach to ensuring fairness in AI systems. The Trustworthy AI Readiness Assessment Framework addresses the critical "question zero" in artificial intelligence (AI) applications: determining both the necessity and readiness of AI solutions within a rigorous ethical, sociotechnical and legal lens. Accordingly, this assessment is conducted during the scoping phase of the AI lifecycle, so organizations can establish a strong foundation for subsequent stakeholder engagement and risk management processes.

The TAIRA methodology is a 7-step process designed to evaluate and enhance an AI systems' readiness in terms of trustworthiness:

- Stating the problem
- Defining the solution
- Identification of Relevant Stakeholders and Factors
- Assessment of the AI Solution against the Ethics Guidelines for Trustworthy AI (EGTAI)
- Testing the AI solution
- Choosing the (AI) solution
- Reflecting & Documenting

The assessment was tested and validated against one of the AEQUITAS use cases, in a multi-disciplinary workshop of AEQUITAS partners. This report presents the findings of this workshop.

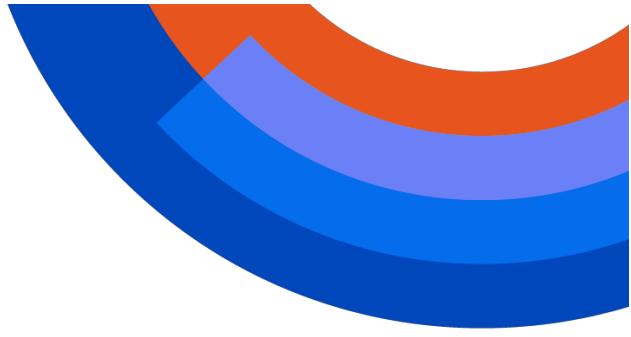
B.1.1 Attendees

- Technical Expert from UNIBO (University of Bologna)
- Socio-technical Expert from UMU (Umeå University)
- Technical Expert from ITI (Technological Center for Research, Development and Innovation in Information and Communications Technologies)
- Policy and Labor Expert from Eurocadres
- Technical Expert from Akkodis
- Ethical and Legal Expert from ALLAI
- Moderator from ALLAI

B.1.2 Use Case Summary

TAIRA was tested and validated against the AEQUITAS use case "AI-assisted identification of disadvantaged students". This AI-system is aimed to identify students who are at a socio-economic disadvantage during their formative educational years,





thereby facilitating support measures to ensure equitable access to educational resources and opportunities. The aim is to mitigate future socioeconomic disparities that can emerge from these early disadvantages.

B.1.3 Testing and Validating the Methodology

Prior to the workshop, participants were provided with a TAIRA workshop template which outlined necessary sections and specific questions to facilitate the testing and validation process. Additionally, they received information regarding the use case, which included a concise overview of the AI system's description, its intended functionality and an account of the dataset it employs. It is important to note, that within this use case no AI system has been currently developed. However, as the TAIRA aims to address 'question zero' on whether an AI system should even be developed, it was sufficient to present the participants with general information on the systems' design and goals. Thus, the information provided was adequate to conduct the TAIRA.

The presentation on the "AI-assisted identification of disadvantaged students" AEQUITAS use case meant the problem had been already defined and the potential solution already identified so Steps 1 and 2 were pre-filled. The discussion then progressed to Step 3, where the Stakeholder and Factor Identification Methodology was employed to pinpoint relevant stakeholders impacted by the system. Subsequently, the group engaged in Step 4, assessing the AI system's alignment with the EGTAI requirements. The workshop concluded with Step 5, which involved discussing the AI system through an ethical lens, and steps 6 and 7, which involved a reflection of the process.

B.1.4 Results

B.1.4.1 STEPs 1 & 2

Prefilled prior to workshop

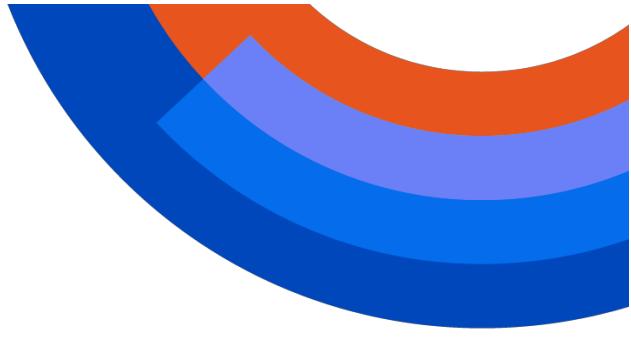
B.1.4.2 STEP 3: Identify Relevant Stakeholders and Factors

In this step, participants classified the various groups of stakeholders, such as those impacted by AI systems (referred to as Affectees), and those with authority over the development and deployment of the AI system (Technical- and Management Decision Makers). Additionally, participants highlighted stakeholders possessing the necessary expertise and insights for fostering the development of equitable AI systems (Domain Experts and AI Users). This part was perceived as fairly straightforward, with minimal ambiguity regarding the delineation of stakeholder roles.

The following stakeholders were identified:

Affectees

Students were identified as being both positively and negatively affected. The benefits were that the system might create a more meritocratic school system and help students



achieve their full potential. However, a drawback was that the system could potentially reinforce historic patterns of discrimination and intersectional vulnerabilities

Society and the school were also listed, as both could potentially benefit from an increase in meritocracy the system could create or suffer loss of talent if the system produces biased outputs.

Decision makers

Computer scientists, socio-economical experts, sales experts and compliance officers were identified to be decision makers during development

Deployers who customize the application, the non-technical experts deploying the system at the school and legal representatives were listed as those managing the system.

Final decision makers were principals, teachers, administration and management boards as well as public authorities

The governance of the AI system would be dependent on the educational institute in which the system is deployed.

Auditors were identified to be as external AI Act and other experts, but it was noted that in these cases school specific biases might go unidentified

The school administration was assigned the responsibility of supervising the system.

Domain Experts

The subjects that would be affected by the AI system, meaning students or their representatives were identified as essential experts.

Independent external experts who don't have a stake in the game, to keep credibility were also listed.

Participants also noted that teachers, professors and admission commissions should be consulted.

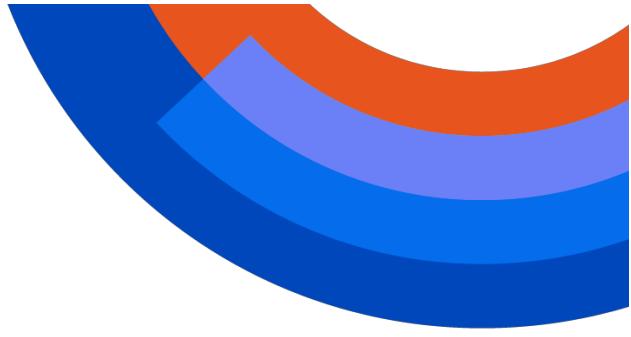
National educational institutes were also identified as essential stakeholders as they have an overview and interest in the entire education system.

Participant also listed ombudsmen for equality and discrimination.

AI Users

The teachers were identified as the ones that would ultimately be using this AI system to assess student outcomes

Students were also listed as potential users, especially if the system had an interface that allowed them to check their status, outcomes or learning paths.



The second segment of the stakeholder identification process focused on constructing a Stakeholder Engagement Map. This map facilitated an assessment of each stakeholder's involvement, determining whether they functioned as experts, co-responsible actors, or fully responsible actors throughout six AI-lifecycle phases:

- Scoping
- Risk assessment
- Development
- Evaluation
- Deployment & monitoring
- Re-evaluation of the AI system.

Participants exhibited some uncertainty regarding the specific role of each stakeholder in these phases, mostly because all stakeholders were seen to be involved in almost all stages.

The affectees were assigned the role of 'expert' in all stages of the AI lifecycle but development. The governance decision makers – legal experts and schools- were listed as co-responsible for the first four stages of the lifecycle and as responsible for deployment, monitoring and re-evaluation. Development decision makers were viewed as fully responsible for the development stage, and co-responsible for the remaining ones. Supervisors and notified bodies were listed as experts across all stages but development, as participants were dubious whether their expertise would be relevant for technical teams in this phase. Domain experts and AI users followed a similar trend, with participants agreeing that their expertise was necessary in all stages but expressing doubt whether they would be included in development.

Due to time constraints in this session, participants did not have the time to elaborate on identifying other influencing or relevant factors.

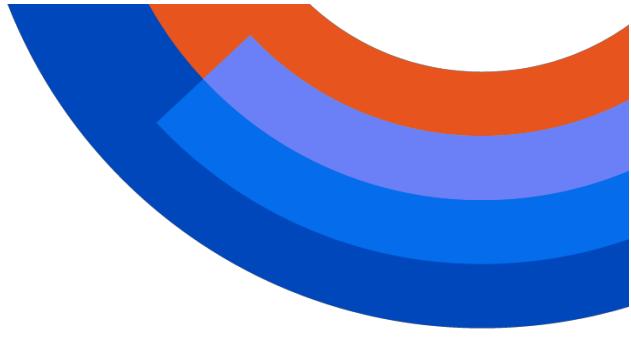
B.1.4.3 STEP 4: Assessing the AI Solutions Against the Ethics Guidelines for Trustworthy AI (EGTAI)

In Step 4, participants evaluated the AI system against the EGTAI requirements.

Regarding **human agency and oversight**, significant emphasis was placed on understanding who comprehends the AI system's outputs. It was highlighted that the AI system is not fully autonomous; it generates recommendations rather than making conclusive decisions. It was stressed that every suggestion or query produced by the AI system requires validation. Participants signaled the need for a classification system to define the required level of human oversight for the AI system.

In the context of **technical robustness and safety**, participants raised questions about the definition of accuracy. Specifically, they inquired whether a higher accuracy in the AI system correlates with a reduction in critical, adversarial, or harmful outcomes. They debated the relationship between increased accuracy and fairness in the outputs generated by the AI system. For example, it was noted that while the system might accurately predict that a student might under or over perform, these predictions could





reflect historic patterns of discrimination and if used maliciously could enforce or justify existing biases, rather than correcting them. This could hugely derail students' lives, affecting access to educational resources, opportunities and diminishing future economic prospects.

Privacy and data governance were also considered as potential concerns. The system uses multiple points of sensitive data, for example socio-economic status which, if not handled correctly could reinforce discrimination. Furthermore, exchange of datasets in this use case was flagged as a potential child protection issue, with participants emphasizing the need for clear guidelines on when, how and where the data would be used.

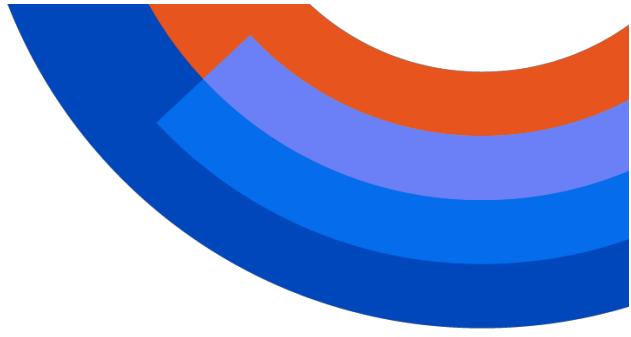
Participants also discussed **transparency** measures that could enhance trustworthiness of the system. Firstly, traceability and clear documentation of what was done and why were brought up as technically feasible means to bolster transparency. Participants also reflected on how to best ensure quality outputs of the system. Technical experts explained how established standards and fairness metrics could be used to assess outputs in numerous ways. It was also noted that a mechanism to overcome human biases – such as automation bias - should also be integrated to maximize human oversight.

Regarding **diversity, non-discrimination and fairness**, the discussions focused on avoidance of unfair bias. It was first established that beyond trustworthiness concerns, under the AI Act providers have a responsibility to avoid unfair bias both in data and in algorithmic design. The use-case was identified to be high-risk, and thus subject to the AI Act requirements. Participants noted that effective bias mitigation at the data governance level was achievable, with the challenging element being the algorithm. Technical participants put forth some techniques that could address is, such as finding aggregate biases.

The system's impact on **societal and environmental wellbeing** was also discussed. In particular, participants zoned in on whether, and how, the AI system could have a negative impact on society at large. Participants noted that the AI system could reinforce discrimination, leading to a reduced opportunity of education and potential rejections from universities. Social experts brought up the fact that this could amplify rather than reduce unfairness noting that the system might be a form of 'regular discrimination with extra steps. It was emphasized that ultimately AI is a tool, and its impact would depend on how it is designed and used.

Lastly, when discussing **accountability**, there was a lack of clarity on how to effectively ensure traceability of the AI system throughout its entire lifecycle. Participants lacking technical backgrounds found it challenging to formulate responses to these inquiries but expressed confidence that a subject matter expert could implement the necessary measures.





B.1.4.4 STEP 5: Testing the AI Solution through an Ethical Lens

Step 5 involved a series of questions aimed at evaluating the AI solution through an ethical lens. The questions encompassed 4 of the 5 tests: the harm test, the publicity test, and the virtue test and professional test.

During the **harm tests**, participants first discussed whether the AI solution was necessary to solve the problem. Again, it was emphasized that AI is ultimately a tool, if it is designed correctly, it was viewed as being useful in solving the problem, making it easier to achieve equal access to education. It was noted that it shouldn't be the sole means for solving the problem and should be paired with policy solutions and non-technical interventions. Participants also noted that if they were in the position of the students on which this system is being used, they would only support the solution if they were provided with some mechanism of collective feedback, frequent checks and transparency to let everyone know the inputs and determinants of the outputs.

For the **publicity test**, participants considered the questions and concerns such a solution might raise with the public. They noted that likely questions about fairness definitions and metrics and consequently the decision-making process would be raised. Additionally, students and parents might be worried about being discriminated against. The participants agreed that with sufficient transparency it would be possible to respond to these questions.

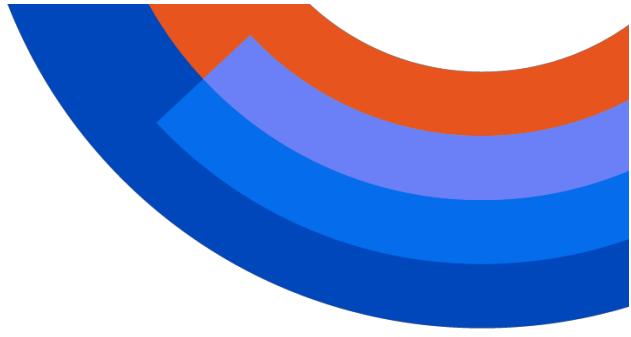
The **virtue test** looked at the beliefs, assumptions, attitudes, and values that the solution reflects. Participants emphasized the meritocratic ideals creating balance in educational settings and opportunities were core beliefs in the design process of the AI. Equity and fairness as equal opportunity were flagged as key assumptions.

For the **professional test**, participants had to consider what peers, classmates, colleagues say about the ethical alignment of the AI solution. Participants noted that questions might come up about how fairness is conceived within the AI solution and on which features (gender, race, socio-economic status) would discrimination be defined.

B.1.4.5 STEP 6 and 7: Choose your solution, Reflect & Document

Step 6 involved comparing the solutions assessed and choosing the best one. As participants only analyzed the AI system presented in the use case, they did not actively engage with this step in the workshop. They were instead provided with a brief explanation and overview of how this step would have progressed in a real-world context where multiple solutions might have been on the table.

Step 7 involved reflection on the solution and the process itself. Participants reflected on the process of assessing the solution using the TAIRA. They noted that the methodology was intuitive and that the questions were easy to follow. The technical participants especially emphasizes that such a tool would be useful for technical teams during risk assessment and design stages. The solution itself was noted to be viable, so long as the identified concerns were adequately addressed.



B.1.5 Key Insights

There was a recurring discussion through-out the session on how this AI system should only be conceived as a tool and not as a sole solution to the issue of inequitable education access. Emphasizing the AI is a tool rather than a comprehensive solution, allowed participants to identify concerns not only from a technical side, but also to consider how the broader social context in which it would operate would affect trustworthiness.

From the technical side, the system was flagged as having the following concerns: lack of clarity around human oversight, risk of reinforcing existing biases through training data, challenges in algorithmic bias mitigation, and insufficient transparency and traceability mechanisms. The use of sensitive socio-economic data also raised concerns about privacy and data governance and surfaced child protection issues

From the broader social side, participants questioned how the system might be used within educational institutions, including the risk of legitimizing exclusion or deepening structural inequalities. Ethical acceptance was seen as dependent on transparency, feedback and redress mechanisms, and alignment with non-technical interventions.

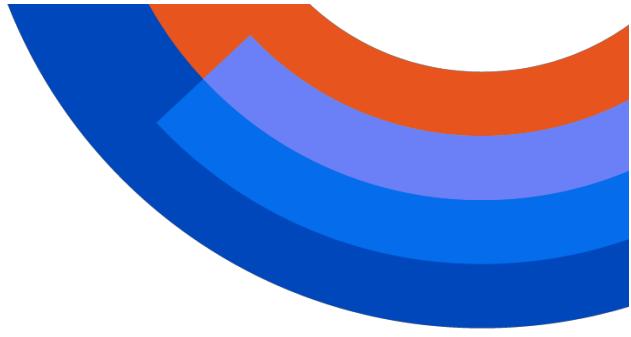
More broadly, the session surfaced the importance of considering trustworthiness within broader socio-technical frameworks. Ultimately, no final decision in this session was reached on whether the system should be developed, but participants noted that it could be a good solution if trustworthiness concerns are addressed

B.1.6 Assessment of TAIRA

Overall, a core finding of the session was the importance of conducting a TAIRA early in the AI lifecycle. Attendees, regardless of their expertise or professional background, engaged collaboratively, contributing diverse perspectives to each posed element. This cross-disciplinary dialogue proved extremely valuable, fostering a rich exchange of ideas that illuminated critical thought processes relevant to the AI system's development.

The questions posed during the workshop were designed to be intuitive, prompting meaningful reflections on ethical, legal and societal implications even among participants with a strong technical focus. No substantive changes of TAIRA were deemed necessary. Technical experts especially noted that the TAIRA was easy to follow and that the questions could serve them when brainstorming and designing an AI system.

As a result, it was evident that integrating TAIRA early in the AI lifecycle is essential, as it facilitates discussions between technical and non-technical stakeholders, a healthy exchange of knowledge, and understanding of positions and interests. This collaborative approach ensures that the design of AI systems is informed by ethical and legal considerations and societal implications, ultimately leading towards trustworthy AI solutions.



B.1.7 Areas for Improvement

Although overall the participants found TAIRA to be a well-developed and useful tool during the risk assessment stage, it should be noted that it best functions when it is used to compare several AI, non-AI and hybrid solutions to an issue. This allows stakeholders to better consider alternatives and avoid the techno-solutionism trap. However, due to the time and case study restraints, in our workshop only one solution was assessed. Resultantly, we would encourage further consideration of and an investigation into other solutions for the issue of unequal educational aspects. In particular, it may be useful to consider a hybrid alternative that combines school-based and policy-based solutions, alongside the proposed AI system.

Additionally, the owners of the use-case were not present in this workshop. Their presence could have provided additional context and insights and likely resulted in a more productive discussion. Lastly, for the TAIRA to be truly effective, it should involve conversations with more diverse stakeholders, including affectees and system users, which in this case would be students, teachers and parents.

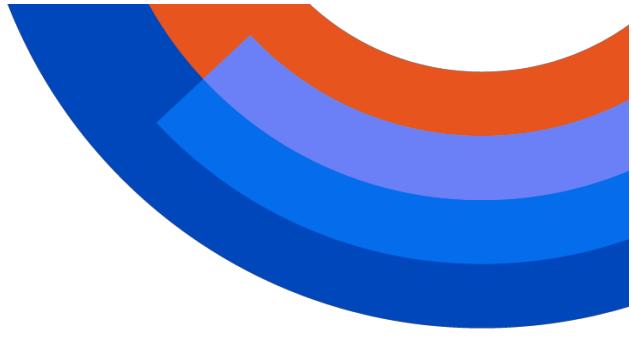
B.2 Fundamental Rights Impact Assessment

The Fundamental Rights Impact Assessment (FRIA) for AI fairness (FRIA-F) has been developed as part of the 'Fair-by-Design Sub-Methodologies' to assess the implications of AI systems on fundamental rights. This methodology supports the implementation of trustworthy AI by integrating fairness considerations into the design, development, and deployment of AI systems.

In the context of AI, Rights Impact Assessments are methodical, structured, and systematic processes that aim to identify, analyze, assess, and manage the effects that result from the development, implementation, and use of an AI system. According to the Article 3(2) of the AI Act, risk refers to the "combination of the probability of an occurrence of harm and the severity of that harm". From this perspective, risk can be conceptualized as the potential rather than the certainty of harm. While health, safety, and fundamental rights are widely recognized as high-risk areas for AI development and deployment, risk can manifest in other categories. Nevertheless, the scope of this methodology centers around the analysis of risks to fundamental rights, whereby any risks posed thereto may be considered as high.

In 2019, the AI High Level Expert Group on AI presented Ethics Guidelines for Trustworthy AI. These guidelines define trustworthy AI as being lawful, ethical and socio-technically robust. For the ethical element of trustworthy AI, the guidelines explicitly take fundamental rights as a basis for AI ethics. This stemmed from the recognition that AI never operated in a lawless world and that existing laws, regulations and treaties apply to AI as much as they apply to any other technology. The principal aim of the AI Act regulation is to safeguard health, safety and fundamental rights from the adverse effects and risks derived from and within the AI lifecycle. This occurs through the AI Act's categorization of AI systems based on the perceived risks they pose to health, safety and fundamental rights. The higher the risk, the stricter the rules for such systems to





enter the EU market. Once an AI-system is categorized as high risk, providers and public sector deployers of such a system must perform a fundamental rights impact assessment (FRIA) (article 9(2)(a) and art. 27 of the AI Act respectively).

Though no further explanation is provided on how to perform a FRIA, providers need to document detailed information regarding the foreseeable unintended outcomes and sources of risks to health and safety, fundamental rights and discrimination in view of the intended purpose of the AI system (art. 11 jo. ANNEX IV para. 3), while public sector deployers must describe and analyze the following elements (art. 27(1)):

- The processes wherein the AI system will be used according to its intended purpose
- Temporality of AI system deployment (time and frequency)
- Affected population
- The risks of harm that are likely to impact the identified affected population
- Human oversight measures aligned with the instructions relayed by the provider
- Response plan in case of risk materialization and attributed internal governance structure including complaint mechanisms

Here, the described interconnection between the (risk-prevention) provisions of the AI Act and the rights granted in the ECFR are important, as one cannot be achieved without respect of the other.

First and foremost, we note that all fundamental rights enshrined in the ECFR may be equally, and concomitantly affected by the design, development and deployment of AI systems. The primary focus of the FRIA-F however, lies on fairness related legal norms that stem from the AI Act and the Charter of Fundamental Rights of the European Union. Further inspiration is drawn from the Council of Europe's European Convention on Human Rights for the Protection of Human Rights and Fundamental Freedoms and the recently adopted Council of Europe's Convention on the Impact of Artificial Intelligence on Human Rights, Democracy and the Rule of Law (AI Convention).

By conducting FRIsAs, all parties involved in the AI lifecycle can proactively identify and eliminate or mitigate potential unfairness in their AI systems, safeguarding fundamental rights. Conducting FRIsAs for AI fairness helps prevent potential biases, as well as unintended consequences and engenders the protection of rights through the understanding of the consequences that technical, organizational or even commercial choices might have for the protection of fundamental rights.

To this end, a multi-disciplinary group session was conducted in collaboration with the providers of an AI-based education prediction tool. The session aimed to assess potential impacts on fundamental rights using a FRIA. Participants were guided by targeted "triggers" questions designed to highlight where fundamental rights, as outlined in the Charter of Fundamental Rights of the European Union, may be at risk.

Throughout each session, there was a diverse range of participants, some from the providers of an AI-based education prediction tool, but the majority from different





disciplines and fields. All participants, guided by a moderator, provided insights based on their individual expertise. The process culminated in the documentation of results, findings, and a discussion of options to avoid or mitigate negative impacts to the fundamental rights of those stakeholders affected by the AI-based education prediction tool.

B.2.1 Methodology

The Fundamental Rights Impact Assessment (FRIA) methodology was used to evaluate potential impacts the AI system might have on fundamental rights. The methodology establishes an 'if-this-then-that' logic', employing reverse engineering tactics, by identifying first any risk-triggers in the AI systems' design, development and use, and connecting those triggers to potential fairness related fundamental rights violations. Because the AI Act aims to protect against these violations, the methodology also helps identifying the relevant prohibited practices (art. 5) or risk elimination and mitigation requirements (art. 9 – 15 AI Act) related to the trigger.

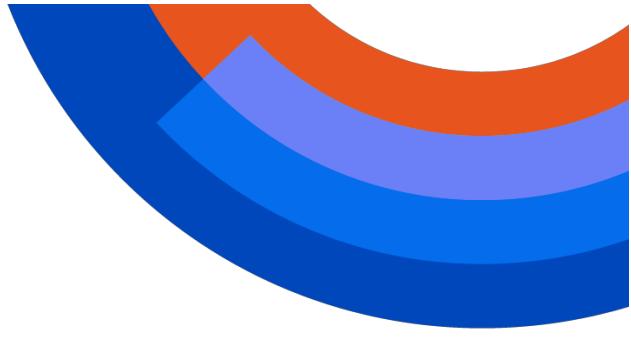
Prior to the meeting, participants were sent the database of the system, a brief description of what the system does and the FRIA methodology booklet to familiarize themselves with the methodological process and use-case. During the workshop, participants were divided into breakout rooms, each consisting of 4–5 members. Every group featured a mix of individuals from technical and non-technical backgrounds, as well as a technical decision-maker representing the 'education tool' providers.

B.2.2 Attendees

- Technical Decision-maker from the 'education tool' provider
- Technical Decision-maker from the 'education tool' provider
- Technical Decision-maker from the 'education tool' provider
- Technical Expert from UNIBO (University of Bologna)
- Technical Expert from UNIBO (University of Bologna)
- Technical Expert from UNIBO (University of Bologna)
- Socio-technical Expert from UMU (Umeå University)
- Technical Expert from UMU (Umeå University)
- Worker Representative from Eurocadres
- Social Sciences Expert from Period ThinkTank
- Technical Decision-maker from a hiring company
- Moderator from ALLAI
- Moderator from ALLAI

B.2.3 Session Summary

The goal of this session was to complete the FRIA. Participants were provided with specific 'triggers', questions designed to highlight areas where fundamental rights could



be impacted. Participants then assessed how these triggers might impact those rights and proposed appropriate mitigation measures.

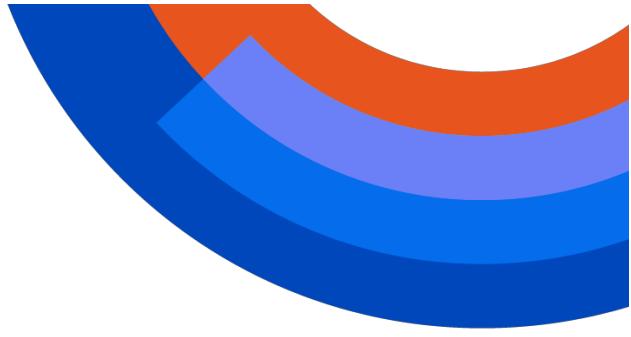
B.2.4 Results

The assessment revealed several areas of concern regarding fundamental rights impacts. **Human Dignity (Art. 1)** could potentially be affected by multiple triggers. In Trigger 1, participants emphasized that the system's lack of human oversight would violate human dignity, noting "the system does not (should not) make a prediction by itself - it should be used always with the oversight of a human." The dignity implications became particularly clear in Trigger 8, where participants noted that the system might make decisions based on "not measurable information" or misinterpreting missing values, potentially reducing students to incomplete data points rather than recognizing their fullness as human beings. The participants highlighted how this could affect a student's dignity through misinterpreting information from missing values about "the talent of the student or the child's state of mind at the time of the test". Throughout the FRIA assessment, participants consistently highlighted the importance of maintaining the student as the central focus rather than allowing the system to reduce them to data points, noting that automated recommendations without human validation compromise dignity through depersonalized decision-making.

The right to **Non-discrimination (Art. 21)** was implicated in several triggers. The assessment highlighted risks around data processing methods (Trigger 7), where normalization and standardization could inadvertently discriminate against disadvantaged groups if not properly managed. The handling of protected characteristics in the training data (Trigger 4) was also identified as a potential trigger for Non-discrimination, raising concerns about discrimination while acknowledging the paradox that this sensitive data was necessary to detect and address bias. Triggers that affected non-discrimination (Art. 21) were also found to harm other fundamental rights such as **Cultural, Religious and Linguistic Diversity (Art. 22)**, **Equality Between Women and Men (Art. 23)** and **Integration of Persons with Disabilities (Art. 26)** as they are closely interrelated.

The **Right to Education (Art. 14)** was central to many identified risks. Participants noted that the main predictor being 3rd grade scores for 6th grade performance could affect educational access if historical bias in academic performance was amplified without considering how "lack of support or resources may have influenced those past results." (Trigger 9). Participants also identified Trigger 13 as harming students' right to education, noting that the system could produce outcomes that "systematically differ across demographic groups." Specifically, they found that predictions for test scores could vary significantly between students from different socioeconomic backgrounds or between migrant and non-migrant students, potentially affecting their educational opportunities. They also emphasized the importance of clear mechanisms allowing students and parents to understand and challenge the system's decisions that affect their educational opportunities (Trigger 21).





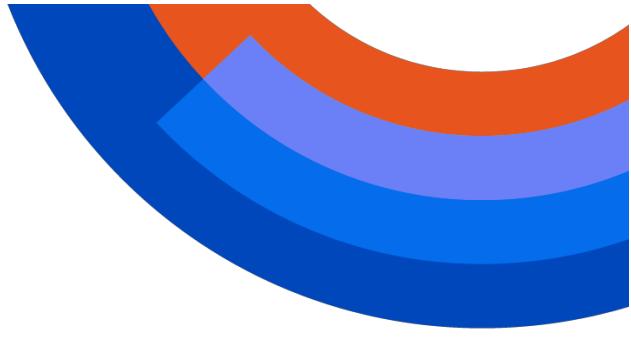
When participants were asked to propose measures to mitigate risks to the identified fundamental rights, they faced challenges due to the limited information available about the technical functioning of the AI education tool. As a result, the assessment was based primarily on a hypothetical scenario evaluating what the risks would be if the system exhibited the identified triggers, rather than an evaluation of the actual system itself. However, we found participants were still able to propose mitigation strategies at both the model-level and the data-level to address risks to fundamental rights in light of the identified triggers.

To mitigate negative impacts on **Human Dignity (Art.1)** participants suggested that predictions regarding students' status as disadvantaged or not should be evaluated against outcomes from a diverse range of students, rather than being limited to the training data alone. Additionally, participants recommended conducting dataset-level analysis as a pre-processing step to determine whether the AI system collects excessive student data beyond what is necessary to identify educational disadvantages or includes arbitrary information in the form of redundant variables. For example, data such as detailed family histories, comprehensive household information, or extensive behavioral records that are irrelevant to academic support needs should be avoided. The analysis should ensure that only the variables strictly required are included. Furthermore, to uphold Human Dignity, participants stressed the importance of ensuring the system's output is both transparent and interpretable to users, enabling informed decision-making. As such, the system should support effective and transparent decisions that positively influence the educational processes affecting students. Finally, users should have the ability to challenge or question the system's outputs. To facilitate this, appropriate measures should be incorporated into the finalized iteration of the AI system, such as feedback mechanisms or flagging tools, that allow users to clearly identify and assess the quality of predictions (e.g., flagging or scoring systems).

Regarding the **Right to Non-Discrimination**, participants acknowledged the complexity of data features related to protected characteristics. They highlighted a tension between the goal of creating a bias-free system and the necessity of including data on protected characteristics such as ethnic or social origin, national minority status, or disability to enable the AI system to accurately identify students who may face disadvantages within the context of education. To address potential negative impacts on the Right to Non-Discrimination, participants advocated for a nuanced approach. They emphasized the importance of addressing biases at the dataset level by carefully identifying the sensitive variables that are essential for the system's functionality. Furthermore, they proposed an iterative process to continually reassess the AI predictor's performance, ensuring that it does not amplify existing biases inherent in the data, which reflect societal biases present at the time of data collection. In the case of bias identification, the participants claimed the predictions should be corrected using state-of-the-art pre-processing or estimation techniques.

To mitigate potential negative impacts on the right to **Cultural, Religious, and Linguistic Diversity** arising from the system's use of Natural Language Processing (NLP) techniques, participants proposed several measures. They suggested employing





semantic extraction to accurately interpret the content of written feedback from parents who use non-standard English, local expressions, or a mix of languages. Additionally, they recommended that human reviewers verify the text in special cases to ensure accuracy. Participants also emphasized the importance of exercising particular caution when collecting data related to immigrant children. They noted the risk that the system might evaluate students based on misinterpreted information rather than their actual capabilities. This concern is particularly relevant for migrant communities, where most individuals in the database speak Spanish and are from Latin America, potentially increasing the likelihood of linguistic or cultural misunderstandings.

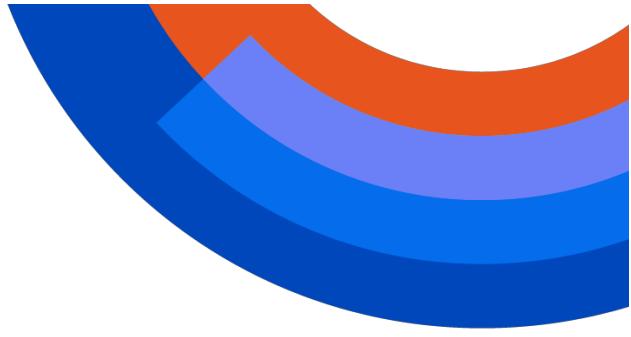
Participants discussed the potential negative effects on **Human Dignity** arising from the system being used at critical decision points (such as grade advancement, special education placement, or access to advanced programs) even if it was originally designed for low-stakes scenarios. To mitigate these impacts, participants emphasized the importance of correctly classifying AI systems and fully understanding their intended use and implications. This aligns with the requirements outlined in the AI Act, which mandates proper labeling and documentation for high-risk applications. When deploying AI tools in high-risk settings, the deployer holds the responsibility for ensuring their appropriate use, even if the tool itself was not explicitly designed for such purposes.

B.2.5 Findings

In taking on positions as both moderators and participants in the FRIA exercise session, ALLAI was able to identify key overarching themes throughout these sessions. In addition, ALLAI's focus on the legal, societal, and ethical perspectives also allowed us to identify some points which were not always elicited throughout the sessions.

One of the challenges throughout the session was limited visibility into the technical functioning of the AI-based education tool under review. Several participants observed that the system appeared to be only a rudimentary implementation, consisting largely of a few predictive algorithms, rather than a fully integrated AI system. As one participant remarked, the tool felt "barely a system. We found this lack of technical detail constrained the ability from the participants to fully assess the potential impacts on fundamental rights. Without a clear understanding of how the system functioned, end-to-end participants were limited in their ability to evaluate the triggers effectively or to determine the specific rights that may be at risk. This likely indicates that the FRIA, in its current form, is best suited for later stages of development, as the 'triggers' provided require evaluators to have knowledge of the systems' inner workings such as data processing techniques, model design and functioning as well as outputs.

Moreover, we found the absence of system-level information also hindered the participants' ability to propose meaningful mitigation strategies, particularly those related to technical interventions. For example, without knowing the architecture of the predictive models, data processing pipelines, or feedback mechanisms, it was difficult to identify what model-level changes might reduce unfairness or bias. Consequently, much of the



discussion remained at a conceptual or speculative level, focusing on hypothetical risks rather than concrete system behaviors.

Overall, we found this experience highlighted the importance of involving an interdisciplinary team when AI providers conduct a FRIA on their proprietary systems. Without sufficient transparency or access to detailed technical documentation, there is a risk that assessments will remain largely theoretical focused on abstract risks rather than yielding concrete, actionable insights. As such, we emphasize the value of multidisciplinary collaboration in AI impact assessments. Meaningful evaluation of potential risks to fundamental rights requires input not only from legal and ethical experts but also from those with deep technical understanding of the system, as well as a commitment to transparent documentation throughout the AI lifecycle

B.2.6 Areas for Improvement

Participants found that the current workshop and template structure created several challenges. First, they emphasized the need for additional preparation time. While the system database, a brief description of its functionality, and the FRIA methodology booklet were shared prior to the session, participants felt that a formal pre-workshop 'practice' exercise was necessary. This would allow participants to independently familiarize themselves with the use case. Additionally, they suggested the possibility of holding a pre-workshop session to review the material in more detail and acquire a deeper understanding of the system under analysis.

Secondly, participants suggested that the FRIA questions be organized according to the different stages of development. For example, data evaluation questions should precede model evaluation questions in the template. They also noted that the triggers were too technical and recommended that the FRIA assessment include triggers addressing the socio-technical aspects of the system. This would cover areas such as the types of decisions made by the system, the users and their interactions with the system, the individuals affected by these decisions, and the potential impact of these decisions. The legal terminology was generally clear, with the blue text examples offering helpful structure and guidance. However, participants requested clearer definitions for certain terms, such as "arbitrary".

Additionally, participants suggested the FRIA analysis would benefit from expanding the questions beyond those focused solely on the technical model and data of the system. They recommended including questions about the system-human interface and interactions to gain a deeper understanding of how the system, even though its basic user interactions, could impact fundamental rights.