

Appendix D: Use case S1 detailed findings

D.1 FbD methodology: Prohibited Social Scoring Assessment

D.1.1 Use Case Summary

The AEQUITAS use case of AI Assisted Identification of Child Abuse and Neglect in Hospitals was evaluated during this workshop. The AI system aims to detect and assess the risk of child abuse and neglect in the hospital context, whilst reducing bias and discrimination on the basis of ethnicity and socioeconomic background. The system was developed using data from 18,000 records of emergency room pediatric patients from 2010 to 2020.

D.1.2 Attendees

- Social Sciences Expert from ThinkTankPeriod
- Social Sciences Expert from ThinkTankPeriod
- Technical Expert from UMU (Umeå University)
- Technical Expert from UMU (Umeå University)
- Technical Expert from UNIBO (University of Bologna)
- Technical Expert from UNIBO (University of Bologna)
- Sociotechnical Expert from Phillips
- Worker Representative from Eurocadres
- Moderator from ALLAI
- Moderator from ALLAI

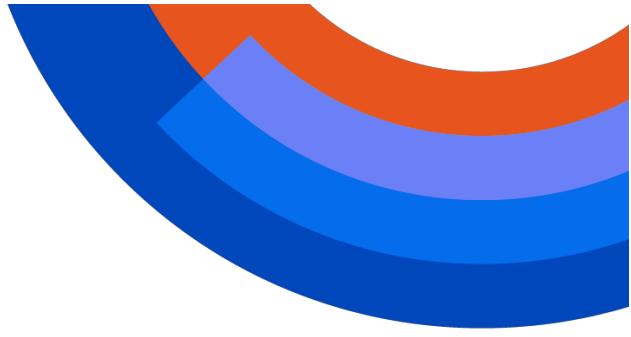
D.1.3 Methodology

The assessment followed the Prohibited Social Scoring Assessment (PSSA) methodology, grounded in the EU AI Act's definition of social scoring as the "evaluation or classification of natural persons or groups of persons over a certain period of time based on their social behavior or known, inferred or predicted personal or personality characteristics".

For the social scoring to be prohibited, it must meet one or both criteria:

- detrimental or unfavorable treatment of certain natural persons or groups of persons in social contexts that are unrelated to the contexts in which the data was originally generated or collected;
- detrimental or unfavorable treatment of certain natural persons or groups of persons that is unjustified or disproportionate to their social behavior or its gravity.

PSSA follows the structure of the AI Act definition, taking participants through a two-step process of:



- 1) identifying whether their AI system uses/constitutes a form of social scoring, and
- 2) assessing whether this social scoring meets the consequence requirements that would render the system classified as prohibited.

To do so, the participants answered a series of reverse-engineered and interlocking questions intended to elicit less overt forms of social scoring and guide participants in dissecting which elements of the use case, if any, would be considered prohibited.

D.1.4 Process

Prior to the workshop, participants received the PSSA Workshop Template. This document outlined the assessment methodology but also included a brief overview of the AI use case. The overview covered key details such as the system's description, its intended functionality, the specific problem it aims to address, and a general (albeit high-level) description of the dataset.

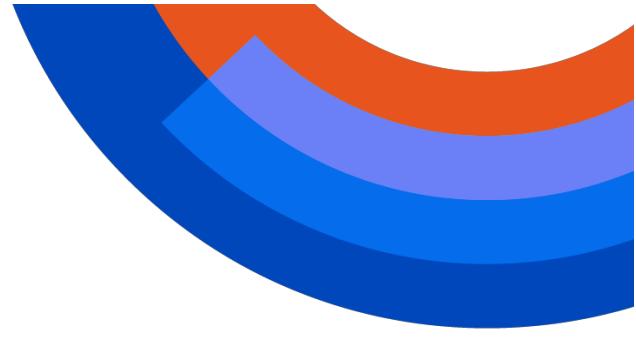
The workshop was conducted via WEBEX and began with a short introduction. During this session, the moderators outlined the objectives of the workshop and provided a summary of the methodology. Given the smaller group of participants, the session was held in a plenary format, with moderators facilitating an open discussion. Following the introduction, the moderators presented the use case in more detail and then guided the participants through the methodology. Each question was discussed in turn to assess whether the AI system met any of the necessary and sufficient conditions that would classify it as a form of prohibited social scoring.

D.1.5 Results

Based on the limited information available about the use case, we applied the methodology outlined in the PSSA Template to assess whether the AI system constituted a form of prohibited social scoring. Our overall conclusion was that the use case does not fall under the prohibition.

D.1.5.1 Part 1: System Requirements

We first assessed whether the AI system exhibited characteristics of evaluation or classification, as per the definition of social scoring. Actively discussing the aspects of the system, we found that the system likely performs a form of classification, as it flags children who may be at risk of abuse thereby dividing individuals into two categories: at risk and not at risk. This classification utilizes a flagging mechanism, where a predicted output identifies a child as potentially at risk. Whether the system also performs evaluation, with functions that could include rating, ranking, or appraising individuals or groups, remained unclear. As such, due to the lack of detail on the technical aspects of the system, we were unable to confirm whether the algorithm incorporated evaluative functions. Resultantly, we concluded the system involves classification but not necessarily evaluation.



We then examined whether the classification relied on social behavior or known, inferred, or predicted personal or personality characteristics. Again, due to limited insight into the dataset, we could not determine the type and sources of data conclusively. However, a participant informed the group that the flagging relied on medical data, including medical history. As such, given that the system uses medical data, it is likely that it could draw on known or inferred personal characteristics, such as mental health history. We discussed a more extreme and hypothetical scenario whereby the system used parental social media data to infer personality traits, yet this was deemed outside the scope of the use case. More plausibly, we discussed whether mental health diagnoses of the child or parents could be interpreted as personality-related traits feeding into the classification process.

Next, we considered whether the data used originated from social contexts unrelated to where it was originally collected. We concluded the system does not meet this criterion as it was explained the AI system operates within the hospital setting, using data collected for medical purposes and thus remaining within the same context. However, we considered more complex cases, such as the use of patient addresses or zip codes, which could function as proxies for socioeconomic status or social behavior. We found the extent to which these types of data constitute a different “social context” is debatable. Nonetheless, under the General Data Protection Regulation (GDPR), both medical history and administrative data (e.g. addresses) are considered part of the protected patient record, with access managed through role-based controls. While medical history is classified as special category data requiring extra protection, administrative data, though not inherently special category, is still protected when associated with patient records. Therefore, we concluded that while the AI system likely relies on personal or personality-related characteristics, it does not use data from unrelated social contexts and thus does not satisfy Point (i) of the prohibition criteria.

D.1.5.2 Part 2: Consequence Requirements

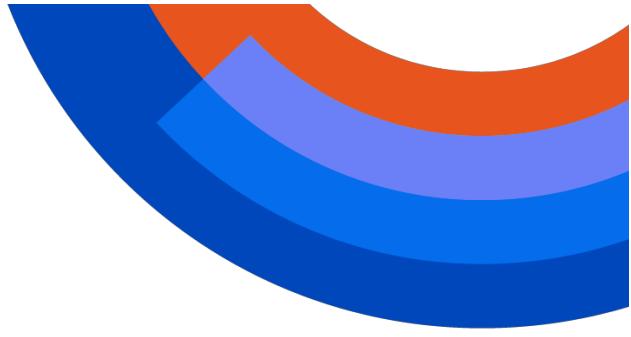
The second part of the methodology assessed whether the use of the system’s output, e.g., the classification of a child as at risk, results in detrimental or unfavorable treatment under the two conditions:

Point (i): Use in unrelated social contexts - We found that any action taken because of the AI system’s output would occur within the hospital setting, where the data was originally collected. Therefore, this condition is not met.

Point (ii): Unjustified or disproportionate treatment - We concluded that any resulting action, such as further investigation by medical professionals, would not be unjustified or disproportionate, given the gravity of child abuse as a social and medical issue. In fact, when a child is flagged as being at risk, it is likely to trigger proportionate and legally justified interventions, rather than arbitrary or excessive responses.

Moreover, the types of contexts in which unjustified detrimental treatment is a concern (such as education, employment, law enforcement, housing, and migration) are those listed in Annex III and Annex I of the AI Act. The AEQUITAS use case does not affect





individuals in these domains directly. Therefore, even if negative outcomes arise for parents (e.g., social services involvement), these outcomes are likely to be justified and proportionate, assuming the system's flagging is accurate and responsibly used.

D.1.6 Areas for Improvement

Overall, the methodology proved to be a suitable medium through which prohibited social scoring could be addressed. The questions flowed cohesively into each other and broke down the relevant aspects of the AI Act into comprehensible criteria. However, several areas of improvement were highlighted. Firstly, the European Commission has released additional guidelines on the prohibitions (including guidelines on social scoring) since the development of the methodology. Thus, to ensure that the methodology is up to date with recent legal developments, it needs to be updated in alignment with the newly released guidelines.

Secondly, participants claimed the wording of certain questions was lacking clarity. For example, participants found the distinction between 'unfavorable' and 'detrimental' treatment to be ambiguous. Whilst the methodology does provide an explanation for the difference between the two, this explanation comes in a later section of the methodology. Participants had similar qualms about the distinction between certain technical terms, such as 'classification' and 'evaluation'. As a result, we intend to include more thorough clarification of these concepts and more definitional explanations to the top of the methodology to ensure consistent understanding and clarity.

Thirdly, participants noted some questions, such as Question 2.d, were overly verbose and could be more concise. In the same vein, Question 6 was flagged for being 'too wordy' and 'repetitive'. To this end, we will review and redraft the relevant questions for better comprehension and concision.

The final area of improvement lies in providing more detailed and structured information about the use case, including the technical aspects of the AI system and the dataset involved. The current description did not align with participants' expectations because the description lacked sufficient depth, making it difficult for participants to fully grasp its relevance and application to the methodology. Additionally, the methodology itself would have benefited from a clearer articulation of its objectives and the rationale behind its design. To address this in future iterations, we will include an introductory note at the beginning of the methodology to outline its purpose, structure, and intended application.