



AI Fairness-by-Design Multi-Stakeholder Methodology

—

**A Comprehensive Framework for
Fair AI Design and Development**

ANNEX II: Prohibited Social Scoring Assessment (PSSA)

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

Contents

ANNEX II: Prohibited Social Scoring Assessment (PSSA)	2
Introduction	4
Prohibited AI Practices – a methodological approach	5
Structure	6
Pathways to the prohibition of social scoring	7
Step 1 - the system requirements	7
Step 2 - The Consequence requirements – part (i) and part (ii).....	7
Questions for Step 1 – system-based requirements	8
Questions for Step 2 – consequence-based requirements	10

Introduction

Determining whether an AI system or practice could be considered (or holding elements of) the prohibited AI practice of social scoring as per the EU AI Act is crucial in establishing whether such system poses too great a risk to fairness. The EU AI Act aims to protect against adverse impact of AI systems and practices on health, safety and fundamental rights and hence prohibits AI practices that pose too great a risk to health, safety and fundamental rights. As elaborated in D6.1 and D6.2, the concept of fairness is enshrined in multiple fundamental rights laid down in the EU Charter on Fundamental Rights, such as non-discrimination, equality between women and men, and rights of children, elderly and people with disabilities.

The prohibition of social scoring of art. 5.1(c) is the most relevant prohibition for AEQUITAS' uses cases as they all involve some elements of 'scoring'.

Prohibited AI Practices – a methodological approach

This sub-methodology contains a set of key questions designed to guide in systematically identifying, reflecting and assessing the potential prohibitive characteristics of their proposed AI solution. This methodological approach is important for evaluating the viability of an AI solution before resources are invested, ensuring fairness risk prevention and regulatory compliance ex ante. This question-based methodology is structured as a decision tree that guides stakeholders through a systematic evaluation of their proposed AI solution's potential prohibited characteristics. It serves two primary objectives: firstly, to translate the legal norms of the AI Act into actionable criteria that can be practically applied, and second, to encourage critical engagement with, reflection and recognition of potential unacceptable risks their AI systems may pose to fairness.

This methodology **does not aim to 'codify' the law, nor does it aim to 'codify' interpretation of the AI Act**. Instead, it breaks down the complex language and potential broad interpretation of the prohibited AI practices into specific answerable questions. In doing so, it prompts to consider how the technical functioning and capabilities of an AI system may impact different groups across different domains in real-life contexts. As such it takes a more granular approach than existing AI Act risk classifiers that merely display the text of the law in a question format. It is important to note that the methodology aims to create an engaging process for analysing the AI system through the lens of these prohibitions and enhance their ability to recognise potential (fairness) risks. As a result, any conclusions drawn from this methodology, such as 'the proposed AI solution is likely prohibited under Article 5.1(c)' are intentionally phrased as guidance rather than definitive legal pronouncements. Rather, these suggestions act as a first step towards deeper engagement with the risks involved and act as an encouragement to seek legal counsel to fully understand the potential regulatory implications of their proposed AI system.

Structure

Article 2 of the AI Act specifies conditions under which AI used in certain domains or for certain purposes, such as the military, defence, national security, and scientific research and development, are exempt from the AI Act. Accordingly, the process begins with the **‘Am I Exempt from the EU AI Act?’** methodology designed as a classical decision-tree framework. This process involves asking a set of questions regarding a parties’ legal status, the context in which the proposed AI system will be used, its intended purpose, and its market conditions. Based on the responses, parties are categorised as either subject to the AI Act or exempt.

If no exemption is established, the next set of questions provide guidance through the **‘Is My AI Practice Prohibited Social Scoring?’** part of the methodology. This part is built on a detailed analysis of the prohibition of art. 5.1(c), social scoring as it is related both to fairness and a number of the AEQUITAS use cases. Art. 5.1(c) prohibits the placing on the market, the putting into service or the use of AI systems for the evaluation or classification of natural persons or groups of persons over a certain period of time based on their social behaviour or known, inferred or predicted personal or personality characteristics, with the social score leading to either or both of the following:

- (i) detrimental or unfavourable 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;
- (ii) detrimental or unfavourable treatment of certain natural persons or groups of persons that is unjustified or disproportionate to their social behaviour or its gravity;

To flesh out the depth of the prohibition, we draw on the relevant recital to extract the lawmaker's intentions, specific details, nuances, and key aspects of the legal language, and where relevant on other regulatory norms. For example, Recital 31 clarifies that *‘AI systems providing social scoring of natural persons by public or private actors may lead to discriminatory outcomes and the exclusion of certain groups*. Moreover, the EU Charter of Fundamental Rights prohibits discrimination on multiple grounds, requires equal treatment of men and women, equal treatment before the law, respect for cultural, religious and linguistic diversity, protection of children and inclusion of the elderly and persons with a disability (Title III ECFR).

Pathways to the prohibition of social scoring

The following exemplar section demonstrates how we categorize potential pathways to the prohibited AI practice of social scoring of art. 5(1)(c). enabling the identification of potential areas of concern. The prohibited practice is described as follows:

“(c) the placing on the market, the putting into service or the use of AI systems for the evaluation or classification of natural persons or groups of persons over a certain period of time based on their social behaviour or known, inferred or predicted personal or personality characteristics, with the social score leading to either or both of the following:

(i) detrimental or unfavourable 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;

(ii) detrimental or unfavourable treatment of certain natural persons or groups of persons that is unjustified or disproportionate to their social behaviour or its gravity;”

By breaking down each clause of this prohibition into necessary and sufficient conditions, we can systematically assess whether a proposed AI solution is likely to fall under the prohibition.

Step 1 - the system requirements

Step 1 broadly relates to what we term ‘system requirements’ which addresses the system-related aspects that must be fulfilled for a system to engage in social scoring. Using this structure, we assess whether the proposed AI solution fulfils the following conditions:

- **Condition 1**, Evaluation/Classification: Whether the proposed AI solution evaluates or classifies individuals or groups.
- **Condition 2**, Basis of Evaluation/Classification: Whether the evaluation or classification is based on any of the following:
 - a. The social behaviour of individuals or groups;
 - b. Known personal or personality characteristics;
 - c. Inferred personal or personality characteristics; or
 - d. Predicted personal or personality characteristics.

If a proposed AI solution meets the two conditions of the system-based requirements, we proceed to assess its likelihood of fulfilling the ‘consequence requirements’ if it were to be placed on the market, put into service, or otherwise used.

Step 2 - The Consequence requirements – part (i) and part (ii)

The ‘consequence clauses’ are articulated in Step 2 **Part (i)** and **Part (ii)** with only one of the conditions from either subsection needing to be satisfied the proposed AI solution to be likely prohibited. Broadly, the ‘consequence clauses’ address the potential harmful outcomes that must result for the social scoring system to be deemed unacceptable. Specifically, it aims to

determine whether the social scoring produced by the AI system could lead to detrimental or unfavorable treatment of individuals or groups.

A proposed AI practice meets the consequence clause if it fulfills the following conditions outlined in **Part (i)**:

- **Condition 1**, Contextual Misuse: Whether the output is used in social contexts different from those in which the data was originally generated or collected.
- **Condition 2**, Negative Outcomes: Whether the output causes detrimental or unfavourable treatment of individuals or groups due to its unrelated transfer of use.

Similarly, a proposed AI solution meets the consequence clause if it fulfills the conditions outlined in **Part (ii)**:

- **Condition 1**, Treatment: Whether the social score is likely to result in detrimental or unfavourable treatment of individuals or groups.
- **Condition 2**, Unjustified / Disproportionate Outcomes: Whether the detrimental or unfavourable treatment is unjustified or disproportionate to the social behaviour being evaluated or classified.

We note that we do not consider the phrase '*with the social score leading to*' a separate condition but rather a consequence of the fulfilment of either or both of the above conditions. This is because any system not specifically aimed at 'social scoring', but mere evaluation or classification could, by virtue of contextual misuse of information or unjustified or disproportionate outcomes, 'become' a social scoring system. For example, a system that is ostensibly referred to as a 'credit scoring system' that integrates a personality trait such as a perceived interest in online gambling based on someone's social media feed, should count as a social scoring system.

Questions for Step 1 – system-based requirements

The following outlines the questions that correspond to eliciting **condition 1**: whether the proposed AI solution evaluates or classifies individuals or groups, and **condition 2**: whether the evaluation or classification is based on social behavior or known, inferred, or predicted personal or personality characteristics of individuals or groups.

Consider the functioning and purpose of the AI solution:

Q.1a Would the proposed AI solution produce any form of classification, such as categorizing, grouping, sorting, of individuals or groups?

Q.1b Would the proposed AI solution produce any form of evaluation, such as assessment, rating, ranking, appraising of individuals or groups?

Q1.c Would the classification or evaluation reflect a certain period of time?

Consider all information used for the functioning at individual datapoint, variable or feature level:

Q.2a Is any data(point/variable/feature)/information related to an individual's or group's on- or offline social behavior?

Q.2b Is any data(point/variable/feature)/information **related to** personal or personality characteristics of an individual or group?

Q.2d Would any data(point/variable/feature)/information be used **to (supposedly) infer** personal or personality characteristics of an individual or group that are not directly observed or documented as input for evaluation/classification?

Q.2e Would any data(point/variable/feature)/information be used **to predict** personal or personality characteristics of an individual or group?

Q.3 Would the proposed AI solution use any data(point/variable/feature)/information as described in Q2 for, or as part of, the classification or evaluation of individuals or groups?

Q.4 Would the proposed AI system produce a score, such as a ranking position, a categorical range, a prediction, a blacklisting, a flagging, a binary classification, an estimate, a percentage, etc. for an individual or group?

Summary: If (any of the) question(s) under 1 and 2, and questions 3 and 4 are answered affirmatively the proposed AI solution in question has likely utilized one or more of the following data types as input for the classification or evaluation of an individual or group, resulting in an associated evaluative or classificatory score:

- The social behavior of individuals or groups
- Known personal or personality characteristics
- Inferred personal or personality characteristics
- Predicted personal or personality characteristics

At this stage, such a system would not yet be classified as prohibited. While it may generate a social score, it remains unclear whether this score will be applied in a context different from where the data was originally collected, or if the score could lead to detrimental or unfavorable treatment that is unjustified or disproportionate given the social behavior. designed to evaluate whether a proposed AI solution's social score is transferable across different contexts, whether the input data originates from various unrelated sources, and whether the social score is likely to result in harm, disadvantage, or unfavorable treatment due to its application in unrelated contexts.

Questions for Step 2 – consequence-based requirements

The following **Part (i)** outlines the questions that correspond to eliciting **condition 1**: Whether the data/information included in the classification or evaluation process is used in social contexts different from those in which the data was originally generated or collected, and **condition 2**: whether the output leads to detrimental unfavorable treatment of individuals or groups due to its unrelated transfer of use of such data/information.

The questions of **Part (ii)** help determine whether the proposed AI solution's outcome could result in detrimental or unfavorable treatment (**condition 1**) that is unjustified or disproportionate in light of the type of behavior or the gravity of the behavior (**condition 2**).

Part (i) Out-of-context data/information

Consider all data and information used for the AI solution's functioning, including by looking at each separate datapoint and variable/feature to determine if any data(point/variable/feature)/information was generated or collected in a different context than the context in which the AI solution will be used.

Q5.a Could any behavioral, personal or personality data(point/variable/feature)/information have been collected, generated or provided in a different context than where it will be used?

Q.5b Could any behavioral, personal or personality data(point/variable/feature)/information have been shared between private and/or public actors, government agencies, financial institutions or other organizations?

Q.5c Could any behavioral, personal or personality data(point/variable/feature)/information have been collected, generated, accessed or accessible across different platforms and services?

Q.5d Could any data(point/variable/feature)/information have been transferred across industries or jurisdictions?

If any data(point/variable/feature)/information was collected or generated in a different context than the context in which the AI solution will be used, it should be determined whether this could cause unfavorable or detrimental treatment in such context of use. Contexts in which detrimental treatment is likely are those listed in ANNEX III (education, recruiting, work, access to essential private and public services, law enforcement, migration, asylum, border control, housing, (health) insurance, the judiciary, democratic processes) and in ANNEX I (particularly medical devices). Other contexts, such as e-commerce, advertising, and (social) media, should also be investigated, for detrimental but particularly for unfavorable treatment.

Q.6a Could the proposed AI solution's outcome, by using any of the out-of-context behavioral, personal or personality t data(points/variables/features)/information as described in Q1, lead to the differential treatment of an individual or group, potentially leading to exclusion or discrimination or unequal treatment?

Q.6b Could the proposed AI solution's outcome, by using any of the out-of-context behavioral, personal or personality data(points/variables/features)/information, lead to unfavorable treatment of an individual or group?

Q.6c Could the proposed AI solution's outcome, by using any of the out-of-context behavioral, personal or personality data(points/variables/features)/information, lead to the detrimental treatment of an individual or group?

Summary: If any of the questions under 5 respectively 6 are answered positively it is likely that the output of the AI solutions is considered, or contains elements of, social scoring that will negatively impact individuals or groups.

Part (ii) - Unjustified or disproportionate treatment

For this condition to be fulfilled, it should be determined whether any unfavorable or detrimental treatment is unjustified or disproportionate, in light of the social behavior as referred to in Q2.a. Contexts in which detrimental treatment is likely are those listed in ANNEX III (education, recruiting, work, access to essential private and public services, law enforcement, migration, asylum, border control, housing, (health) insurance, the judiciary, democratic processes) and in ANNEX I (particularly medical devices). Other contexts, such as e-commerce, advertising, and (social) media, should also be investigated, for detrimental but particularly for unfavorable treatment.

Q.7a Could the proposed AI solution's outcome, by using any behavioral, personal or personality data(points/variables/features)/information as described in Q2, lead to the differential treatment of an individual or group, potentially leading to exclusion or discrimination or unequal treatment?

Q.7b Could the proposed AI solution's outcome, by using any behavioral, personal or personality data(points/variables/features)/information as described in Q2, lead to unfavorable treatment of an individual or group?

Q.7c Could the proposed AI solution's outcome, by using any behavioral, personal or personality data(points/variables/features)/information as described in Q2, lead to the detrimental treatment of an individual or group?

Q8 Can you provide a justification why each behavioral, personal or personality data(points/variables/features)/information as described in Q2, should lead to differential, unfavorable or detrimental treatment?

Q9.1 Would the behavioral, personal and/or personality data(point/variable/feature) / information as described in Q2 be considered trivial, irrelevant, inconsequential, unimportant, negligible, or otherwise insignificant, in light of the differential, unfavorable or detrimental treatment?

Summary: If any of the questions under 7 & 8, respectively 9 are answered positively it is likely that the output of the AI solutions is considered, or contains elements of, social scoring that will negatively impact individuals or groups.

Answering the above questions provides an indication whether the proposed AI solution is likely to perform (or hold elements of) social scoring. Please note however, that even if no social scoring is established, the proposed AI solution could still make assessments of individuals or groups based on their personality traits or characteristics in order to assess or predict the risk of a criminal offense, which is prohibited under art. 5.1(d).