मानक
सं: टीईसी 57050:2023

STANDARD
No.: TEC 57050:2023

आर्टिफिशियल इंटेलिजेंस सिस्टम की निष्पक्षता मूल्यांकन और रेटिंग

Fairness Assessment and Rating of Artificial Intelligence Systems

ISO 9001:2015

दूरसंचार अभियांत्रिकी केंद्र
दूरसंचार विभाग, संचार मंत्रालय
खुर्शीदलाल भवन, जनपथ, नई दिल्ली – ११०००१, भारत
TELECOMMUNICATION ENGINEERING CENTRE
DEPARTMENT OF TELECOMMUNCATIONS, MINISTRY OF COMMUNICATIONS
KHURSHID LAL BHAWAN, JANPATH, NEW DELHI - 110001, INDIA
www.tec.gov.in

Release : July, 2023

FOREWORD

Telecommunication Engineering Centre (TEC) is the technical arm of Department of Telecommunications (DOT), Government of India. Its activities include:

- Framing of TEC Standards for Generic Requirements for a Product/ Equipment, Standards for Interface Requirements for a Product/ Equipment, Standards for Service Requirements & Standard document of TEC for Telecom Products and Services
- Formulation of Essential Requirements (ERs) under Mandatory Testing and Certification of Telecom Equipment (MTCTE)
- Field evaluation of Telecom Products and Systems
- Designation of Conformity Assessment Bodies (CABs)/ Testing facilities
- Testing & Certification of Telecom products
- Adoption of Standards
- Support to DoT on technical/ technology issues

For the purpose of testing, four Regional Telecom Engineering Centres (RTECs) have been established which are located at New Delhi, Bangalore, Mumbai, and Kolkata.

ABSTRACT

This Standard enumerates detailed procedures for accessing and rating artificial intelligence systems for fairness. Artificial intelligence is increasingly being used in all domains including telecommunication and related ICT for making decisions that may affect our day-to-day lives. Any unintended bias in the AI systems could have grave consequences. This standard provides a systemic approach to certifying fairness for AI systems. It approaches certification via a three-step process involving bias risk assessment, threshold determination for metrics, and bias testing. Bias testing includes scenario testing, where the system is tested in different scenarios to ensure that it performs equally well for all individuals. The standard is presently built for tabular data and intended to be expanded to other forms of data.

HISTORY SHEET

S. No.	*Standard No.*	*Equipment/ Interface*	*Remarks*
1.	TEC 57050:2023	Fairness Assessment and Rating of Artificial Intelligence Systems	July 2023

CONTENTS

1.0 Introduction

The increasing use of Artificial Intelligence (AI) and Machine Learning (ML) applications in all domains and the effectiveness of AI/ ML in public services delivery and e-governance by Government Organisations make it necessary for AI systems to be fair and unbiased. Unintended biases in AI applications lead to various ethical, social, and legal issues. National Digital Communications Policy 2018 [1] mandates synergizing deployment and adoption of AI and emphasizes leveraging AI technology to enhance the overall quality of service, spectrum management, network security, and reliability.

Toward achieving fairness in AI, some questions come to mind. What should fairness mean? What are the causes that introduce unfairness in machine learning? How best should we modify our algorithms and data to avoid unfairness? And what are the corresponding trade-offs with which we must grapple? [2]. However, this standard differs because it goes beyond mere questioning and attempts to develop a framework. The framework aims to ask some of these questions and provide a qualitative measure that helps to understand where the given AI system stands in terms of fairness. The fairness measure would also inform a non-expert if the AI were good enough to be allowed into production. This standard provides a framework to examine biases in different components of AI systems.

This standard helps in the fairness assessment of AI systems and provides a reference scale for their comparison. Fairness certification aids in compliance with the standard and trust, equity, and transparency among the people. As governments use AI applications to deliver citizen-centric services, determining the fairness of such applications would become a requirement, and citizens would be the beneficiaries of fairness certification. Fairness is a subjective topic; hence, having a deterministic yardstick, metrics or thresholds cannot be applied uniformly across use cases, domains, or industrial environments.

The objective of the standard is to promote bias assessment, enabling a standard procedure for bias assessment and its transparent disclosure and thereby enhancing

trust in the AI system. Governments, business enterprises, and non-profits can use this standard to demonstrate their efforts toward fairness. The standard can be used in two ways: (1) Self-certification, wherein the entity conducts an internal assessment of the AI systems and provides a report as per the requirements of the standard, and (2) Independent certification, wherein an external auditor conducts an assessment and provides a report as per the requirements of the standard. This standard details how bias will be assessed, measured, presented, and disclosed. The entity seeking certification can adopt the standard for internal assessment, while the external auditor can conduct an independent assessment to certify the product under this standard. Section 3 enumerates how this standard may be used.

Bias can creep into an AI system in many ways, in different stages of the data and AI lifecycle, and can affect other AI components. The standard describes such scenarios in Section 4 towards the goal of early detection of biases. Section 5 describes various fairness metrics and the necessary consideration for evaluating such metrics. Section 6 describes the standard operating procedure to determine biases. Section 7 provides the guidelines for producing a certification report, and Section 8 concludes with limitations and future work.

Our approach to creating the framework focuses on two impact groups: (a) Direct implications caused due to citizens (or primary ‘affected stakeholders’) being subject to decisions of a specific AI system, and (b) Indirect implications caused due to the overall deployment of AI solutions in society [3]. Further, the framework is built on the Principles of Responsible AI laid out by NITI Aayog in India, specifically equality, inclusivity, and non-discrimination.

2.0 Terminology

2.1 ML Definitions

- **Machine learning:** An approach for learning models from data.
- **Model:** A function that finds patterns or makes decisions.
- **Label:** A value indicating the outcome or category for a sample.
- **Score:** A continuous valued output from a classifier. Applying a threshold to a score results in a predicted label.
- **Feature:** An attribute containing information for making a decision
- **Classifier:** A model that predicts categorical labels from features.
- **Training/ Test Data:** A dataset from which a model learns/ is tested.
- **Supervised learning:** Supervised learning (or supervised machine learning) is a subcategory of machine learning and artificial intelligence. It is defined using labelled datasets to train algorithms to classify data or accurately predict outcomes.
- **Unsupervised learning:** Unsupervised learning (or unsupervised machine learning) uses machine learning algorithms to analyse and cluster unlabelled datasets. These algorithms discover hidden patterns or data groupings without requiring human intervention.
- **Positive class and negative class:** The output of the binary classification system, which corresponds to the yes decision to the classification question, is called the positive class and the other class is called the negative class.
- **True/ False Positive/ Negative:** A true positive (TP) is a sample correctly classified as belonging to the positive class. A true negative (TN) is a sample correctly classified as belonging to the negative class. A false positive (FP) is the sample that is predicted as positive but belongs to the negative class. A false negative (FN) sample is mispredicted to the negative class.
- **Precision, Recall, Accuracy:** Precision denotes what proportion of positive predictions are correct, i.e. TP/ (TP + FP). Recall identifies what proportion of

actual positives was identified correctly, i.e. TP/(TP + FN). Accuracy is the fraction of correct predictions, i.e. (TP + TN)/(TP + FP + TN + FN).

- **Explainability:** The property of an AI system to express essential factors influencing the AI system results in a way that humans can understand [4].
- **Recourse:** an understandable human description for individuals who received a negative outcome to change their features to obtain a positive outcome.
- **Interpretable and Non-interpretable models:** Interpretable models are those whose internal logic is easy to interpret - for example, decision trees and logistic regression models are open-box. Non-interpretable models are those whose internal workings are hard to interpret - examples are neural networks, gradient boosted trees.
- **Open-box/ closed-box/ grey-box access to AI models:** Open-box model means we can access the internal logic, parameters, and hyper-parameters along with the training data. For closed-box models, the internal logic of a model or the training data is not known, and only the input-output behaviour of the model is known. The training data in grey box models are known, but the model internals are unknown.

2.2 Definitions relating to fairness

- **Bias:** In the context of fairness, bias is an unwanted characteristic that places one group at a systematic advantage and another group at a systematic disadvantage in comparison to another group.
- **Favourable label:** In a binary classification system, the positive class refers to the outcome that individuals wish to achieve, and its label is called the favourable label/ outcome.
- **Protected Attribute:** The feature that must not influence the decision process of a machine learning algorithm. For example, the gender of a person should not affect the job candidature in general. These protected attributes may also include sensitive attributes (e.g. Sensitive Personal Information).

- **Privileged and Unprivileged groups:** Given a binary protected attribute, the unprivileged groups are individuals who experience systematic discrimination, whereas the privileged group is the rest.
- **Group Fairness:** The goal of groups (per protected attributes) receiving similar treatments or outcomes.
- **Individual Fairness:** The goal of similar individuals receiving similar treatments or outcomes.
- **Bias mitigation process:** A process for reducing unwanted bias in training data, models, or decisions.
- **Affirmative action:** A series of policies that aims to increase the opportunities provided to the underrepresented/ unprivileged members of society.

3.0 Usage of the standard

3.1 Scope

The standard attempts to cover the following:

1. Types of AI/ML systems for which the proposed fairness measurement framework is applicable.
2. A combined fairness rating metric for AI/ML systems under the framework.
3. Framing Standard Operating Procedures (SOPs) for evaluating and rating AI systems for fairness. This would form part of the framework for AI assessment.

The standard's end goal is to develop a set of SOPs that would be used to assess and arrive at fairness scores for different fairness metrics and a combined fairness score. The standard is intended to be used as a tool for validating the risks as part of self-assessment but can also be used as an assurance by a third party via an independent audit. Globally, independent audits of AI systems are gaining interest, with regulators increasingly mandating compliance and assurance with such audit mechanisms.

The standard can primarily help a developer assess the developed AI system. As the developer knows the most about the AI system and the data used to develop it, the framework is expected to help the developer self-certify the AI system based on the framework. However, the intention is not to make the framework a mandatory regulatory assessment tool. It is nearly impossible to fathom all situations and create a framework that flexes to cover each type of AI development process and still does justice when giving fairness scores. Thus, the framework intends to allow generous introspection through SOPs and to arrive at the fairness scores that a developer can confidently display; the auditor assesses the AI further wherever required.

3.2 Users of the Standard

3.2.1 Organisations/ individuals developing AI systems

One goal of the Standard is to help the AI developer arrive at a set of fairness scores for the AI system under development on a self-assessment basis through the SOPs

recommended under the framework. Hence, the first level user of the report would be the AI system developer.

The second level user would be the auditor or the tester responsible for auditing the AI system. The fairness scores, as indicated by the developer at the first level, and the evaluation of the developer's adherence to the SOPs given in the framework would provide a baseline for the auditor to proceed with further evaluations.

The third-level user would be the management and the key decision-makers. The key decision-makers may be the policymakers in the government, the regulators from a regulatory agency, civil society members who work in AI fairness or ethics, lawyers, and business leaders who need to decide to release the AI tool into production.

3.2.2 Third-party auditors

Independent third-party auditors, accredited by a certifying agency, may audit the AI systems and issue Fairness Certificates with fairness rating scores based on this standard. The sector regulators could voluntarily or mandate the certifications. The third-party auditors are also responsible for validating the assumptions and choice of parameters used by the AI tool developer during self-certification. The auditors are expected to be a team of domain experts, representatives from legal and regulatory bodies, and technology and data experts. The auditors should have sufficient domain knowledge to verify the context-specific choices (of protected attribute/metric/threshold selection) made by the auditee. The auditor may seek access to data or statistical properties of data, model, or metrics from the model as a way to evaluate the bias in the model if constrained by concerns over proprietary information and related intellectual property. However, the auditor shall appropriately explicitly document the same in the report along with specific limitations to comprehensively certify the AI system.

3.2.3 Procuring organizations

Many organizations follow a transparent tendering process for procurement. These include government departments, public sector undertakings, banks, international

bodies like World Bank, non-governmental organizations (NGOs), etc. Their future procurements might consist of AI-based applications. The services offered by these organizations might impact the lives of millions of citizens. It is, therefore, essential for them to deploy only those AI systems that are proven fair.

To benchmark, the solutions offered by various bidders at the time of bidding, the standardized fairness rating, such as the Fairness Score Certificate, could be asked as a qualification criterion. Also, these organizations might need more expertise to assess whether the delivered AI systems are fair. So, these procuring organizations may ask for a self-certification, or a third-party certification for fairness based on the SOPs enumerated in this standard.

3.2.4 Sector regulators

In specific verticals where fairness in AI systems is crucial, such as legal expert systems, medical diagnosis applications, self-driving cars, and autonomous aircraft, the sector regulators may mandate tolerance levels on relevant, carefully selected metrics. They may specify the minimum fairness rating score as a benchmark for different industry-specific use cases, including any specific tolerance levels if necessary.

3.2.5 Start-ups and SMEs

Developers, particularly start-ups and SMEs may get their systems certified for fairness from third-party auditors for broader acceptability of their products.

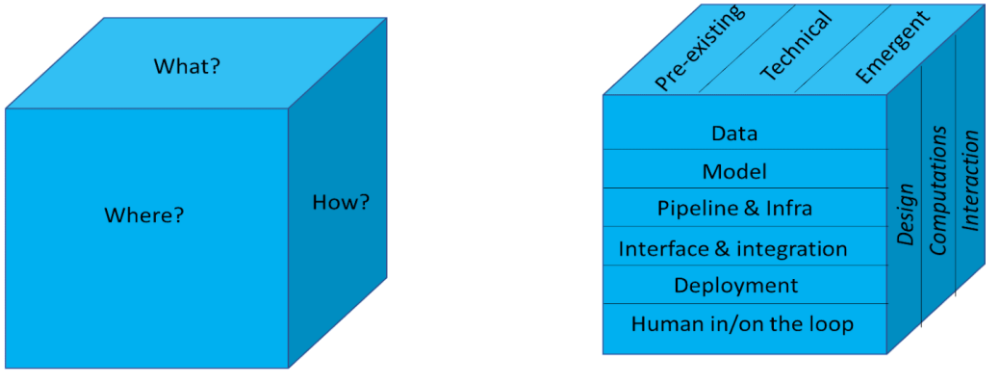
4.0 Sources of Bias in AI systems and AI lifecycle

4.1 Understanding sources of Bias

This section provides an overview of different types of AI systems, different kinds of biases that may occur in such systems, the generic components of AI systems, and how biases can originate and affect such subsystems in different phases of the AI lifecycle. This forms the basis for assessing fairness in each subsystem and the overall AI system.

Algorithms trained on biased data are perceived to contribute to inequalities in outcomes, thereby putting certain groups at a disadvantage [5]. Unfairness and discrimination can also be attributed to the model's parametric choices. Based on the impacted group, biases or unfair discrimination are classified into Individual bias, Group bias, and multi-Group bias. For instance, bias in a clustering algorithm may affect specific group(s). The bias may exclude individuals and groups for classification and regression problems, while ranking, matching, and recommendation algorithms may impact individuals, groups, and multi-groups. Knowing the type of AI system and the bias implication helps in scanning for such elements.

To understand bias or unfair discrimination, it is necessary to understand the sources of bias. Bias Cube [6] attempts to provide a systematic explanation of sources of bias, types of bias, and how it gets exhibited to a user, thereby helping in assessing the potential impact and possible solutions for addressing the discrimination.



Source: Bias Risk Assessment – A systematic approach Part 1 | Medium

Figure 1: Understanding Bias in the context of automated decision systems

Biases are also categorized as pre-existing, technical, and emergent biases [7]:

- **Pre-existing biases** are contributed by individual or societal biases in the environment. These get amplified by the selection and representation of the data gathered.
- **Technical bias** can arise from tools, lack of context for the model, and inconsistencies in coding abstract human concepts into machine learning models.
- **Emergent bias** arises from the feedback loop between human and computer systems.

4.2 Understanding factors that influence bias in each source

Bias can occur in different stages of the lifecycle from other sources, affecting other components. This section lists the key contributors of bias for each component of AI systems across the lifecycle. These contributors are classified into process factors and technical factors, defined as follows:

- **Process factors:** The activities that determine the need for specific actions in the data and AI lifecycle in this standard's context of bias examination are referred to as process factors.
- **Technical factors:** The decision activities in the data and AI lifecycle involving the model development that may result in bias are referred to as technical factors.

4.2.1 Data

Biases arising from data are prominent avenues that contribute to disparate impacts. Bias in data can come from various touch points, including data gathering, data augmentation, data merging, data cleaning, data pre-processing, data encoding, and data splitting. These actions can happen across multiple stages of the AI lifecycle. Instances of bias in data may arise due to the following:

Process factors	Technical factors
• Historical data and pre-existing bias • Relevance of data • Under-representation in data • Completeness (including missing values or inputs from uncalibrated sources) • Retraining data	• Imputations of missing data (including replacing or substituting values) • Duplicate data removal • Outlier treatment, including outlier removal, normalization, discretization, feature selection • Annotations, including labelling inconsistencies • Inferences and proxies associated with data • Unvalidated causal relationships

4.2.2 Model

Bias can arise from models due to model choices, feature engineering, training, parametric choices, and testing and tuning. Such biases may occur due to:

Process factors	Technical factors
• Pre-trained models (bias from transfer learning) • Associated or connected model's quality • The extent of training (e.g., number of hours of training) • Objective definition • Adversarial exposure, including model or system feedback process (e.g., leading to data poisoning)	• Choice of models (including the architecture, e.g., no. of layers) • Feature choices (inferences & proxies) • Seed, epoch, batch size, learning rate, and dropouts • Activation, loss, and optimizer choices • Metric choices for testing • Tuning choices • Metric choices for performance monitoring

4.2.3 Pipeline & infra

Biases can arise from pipelines and infrastructure due to pipeline, infrastructure robustness, infrastructure measures, and optimization choices. Instances of such biases may occur due to the following:

Process factors	Technical factors
• Pipeline quality (including errors and defects) • Uncertainty calibrations • Pipeline robustness against data leakage or exposures	• Optimization choices for throughput, latency, scalability, and resource usage.

4.2.4 Interface and integrations

Biases arising from interfaces and integrations are widely examined in the user interface (UI/ UX) or application programming interface (API). Their relevance to discrimination should be seen from the perspective of disparate implications they may cause. Biases can arise from interface and integrations due to nudges, design, and integrations (with tools or other models). Instances of such biases may occur due to the following:

Process factors	Technical factors
• Social and technological accessibility (hardware requirement, disability, etc.) • Integration quality, including defects, failures, and adversities	• Interface design preferences that unfavourably position minority-related options (e.g., User choice architecture used for nudges and deceptive designs)

4.2.5 Deployment

Bias arising from deployment is often caused by the environment in which the model or application is implemented. Discrimination can arise from statistical distribution differences between training and deployment environments. It can also occur due to changes in the meaning of inferences and casualties. Instances of bias in interface and integrations can occur due to:

Process factors	Technical factors
• Statistical distribution differences (between training and deployment) • Changes in the meaning of inferences and choices	• User interactions and adaptiveness • User interaction or feedback collection choices

4.2.6 Human-in-the-loop/ Human-on-the-loop (HIL)

Disparate impacts arise from human-in-the-loop or human-on-the-loop due to the inherent biases of human actors in the process. Human decisions associated with inferences, proxies, causalities, outcomes, and subsequent actions (specifically human-in-the-loop and human-on-the-loop decisions on model outcomes) can contribute to biases. Instances of such biases may arise due to the following:

Process factors	Technical factors
• HIL fitment in the model lifecycle	• Approach towards observing outcomes • Conclusions and inferences reached (including unvalidated casualties) • Beliefs and their influence on actions

4.2.7 AI-based system

Bias contributed by the AI-based system is due to the holistic use of the application. This includes overall system design and associated choices that may impact the way the users of the AI-based system perceive the outcomes, contributing to bias. Bias

contributed by AI-based systems shall only be examined after considering the biases contributed by all the above sources. Instances of discrimination in AI-based systems may arise due to the following:

Process factors	Technical factors
• System design • Disparate errors • Accessibility	• User journey map

Note: Some of the biases may not be relevant for a given AI system.

5.0 Overview of Fairness Metrics

In general, fairness metrics evaluation requires the identification of protected attributes, unprivileged/ privileged groups, favourable outcomes, fairness metric selection, and computation.

5.1 Protected attribute selection

Protected attributes are part of the dataset to describe the identity (user profile) or provide information on the data subject/ user (demographic profile). The protected attributes may be race, ethnicity, nationality, skin colour, gender, age, sexual orientation, marital status, religion, political opinion, disability, etc. These attributes shall not be relied on for determining eligibility or ineligibility for goods or services that may impact the rights of these individuals. The selection of protected attributes should be based on legal, ethical, and application context and may therefore be chosen wisely. For example, pregnancy should not be considered a protected attribute for a classification problem that tries to predict the treatment procedure in the radiology department. However, it can be viewed as a protected attribute when employees are chosen for a particular non-laborious training.

There are also two other forms of protected attributes - 1) Statistics related to a protected attribute, for example, count of the number of senior citizens, 2) Correlated attribute or proxies-attribute, which has a direct correlation with the protected attribute, for example, in a particular dataset containing people between age 1 to 20, height may be considered as correlated to age. Similarly, zip code in some geographical regions may be correlated to religion, and if such a case exists, then zip code also needs to be considered a protected attribute. Note that one must attempt to identify all the protected attributes in the data.

Demographic variants in India are extensive; therefore, one must be careful in identifying protected attributes, especially correlated ones. For example, there is a strong correlation between religion and location in various parts of India. Factors such as ‘dependent parents’ do not exist in western societies. Therefore, such factors can

be considered a protected attribute to predict whether an individual will get health insurance.

5.2 Privileged and Unprivileged group selection

Once the protected attributes are selected, one needs to identify the groups or classes for each protected attribute experiencing systematic discrimination (unprivileged group). In contrast, the privileged group can be the rest (for binary protected attributes) or a particular group (for multi-class protected) that might experience systematic discrimination.

The task is more straightforward for categorical protected attributes as the categories are known and present in the data. However, the user must identify the ranges for the privileged and unprivileged classes for continuous attributes. Users should use acceptable demographic classes relevant to the fairness assessment. For example, age greater than 60 for senior citizens or the legal age for voting in India, etc. The user's rationale for choosing the potentially privileged and unprivileged classes for which to make a test is essential while assessing the system's fairness.

The next consideration is the definition of a privileged or unprivileged group which involves multiple more than one protected attribute. For example, for a particular application, 'Married Woman' might be worth considering as a potentially unprivileged class even though individually 'Married' for marital status and 'Woman' for gender may not be the unprivileged classes.

5.3 Identification of Favourable Outcome

A favourable outcome is a label value corresponding to an outcome that provides an advantage to the recipient. The opposite is an unfavourable label. The label is known for a binary classification setting, and the user needs to identify the favourable outcome. For example, in a loan prediction case, a favourable outcome corresponds to getting the loan, whereas a favourable outcome for cricket performance prediction may be 'not getting out.' Each label corresponds to a favourable or unfavourable outcome for a binary classification scenario. However, in a multi-class classification

scenario, one must choose the labels corresponding to favourable and unfavourable outcomes. Both sets should be exclusive.

Choosing favourable and unfavourable outcome ranges for a regression problem is non-trivial, and users are advised to document their understanding and rationale for making such a choice in the report.

5.4 Fairness Metrics

There are many different approaches to defining fairness in AI systems. Further, AI fairness is also context-dependent and culture-dependent. When the fairness of AI is derived from existing societal laws, it is compared with the anti-discrimination laws followed by various countries. It is classified under the following two major categories:

1. Disparate/ Adverse Treatment: An intentional decision-making process suffers from disparate treatment if decisions are based on the subject's protected attributes.
2. Disparate/ Adverse Impact: Outcomes of a seemingly neutral decision-making process that disproportionately hurts/ benefits people with specific protected attribute values. Note: Here, the term does not mean the disparate impact metric.

We present a set of fairness metrics for dealing with adverse impact scenarios. Additional metrics can be found in [26]. The fairness metrics quantify the level of discrimination in the outcome and treatment towards an individual or group(s) of the population. The specific metric may be used, keeping in mind the context and application of a given AI algorithm.

5.4.1 Group fairness metrics for classification

1. **Demographic (dis)Parity:** In the classification setting, a classification algorithm's rate of acceptance (fraction of individuals classified positively) rate across groups. A ratio or a difference quantifies this notion. The four-fifth rule, used in many countries, suggests that this ratio must not be less than 0.8 between any two groups. Though common in practice, this notion should be used carefully. This notion presupposes equal claim and equal qualification in all the groups

ignoring any natural advantages individuals from some groups may have over others. The ratio-based demographic parity is also called disparate impact metric.

2. **Minimum allocation guarantee (quota guarantee):** This notion guarantees an absolute minimum guarantee to each group. Demographic parity also ensures a quota proportional to the size of each group. This notion generalizes the quota guarantee by allowing different values for different groups, which may or may not be proportional to the size of the group. The quota is given as input.
3. **Predictive rate parity**: Predictive parity requires equal precision for protected and non-protected groups.
4. **Predictive equality:** Predictive equality requires an equal false positive rate FP/(FP + TN) between the groups, where FP is a “false positive” and TN is a “true negative” outcome.
5. **Equality of Opportunity**: This metric warrants an equal false negative rate between privileged and unprivileged groups. The false negative rate is the ratio of false negatives to the sum of false negatives and true positives.
6. **Equalised odds:** This notion conditions the acceptance rate on the true label of individuals. In other words, it requires equal positive and negative rates.

5.4.2 Individual Fairness Metrics for classification

This notion requires that similar individuals are treated similarly.

1. **Lipschitz condition:** This condition asserts that given similarity metrics on individuals (feature space) and outcome space, the individuals closer to each other in the feature space must be placed close in the outcome space by the algorithm. We can compute the outcome as the fraction of similar individuals which yields similar outcomes. One drawback of using this evaluation technique is the assumption that these metrics are agreed upon by all parties and given as input to the algorithm [8].
2. **Meritocratic Fairness:** This notion requires that less qualified individuals (according to some pre-decided criterion) are never preferred over more qualified individuals [9].

3. **Calibration:** In selection tasks, the probability of selecting a particular individual must match the probability that the individual is the best candidate among all available candidates under given (partial) information. Calibration in scoring tasks implies that the assigned score indicates the fraction of individuals with the same score having positive labels. That is, the assigned scores have a semantic meaning.
4. **Counterfactual Fairness:** Counterfactual fairness captures the intuition that a decision is fair towards an individual if it is the same in (a) the actual world and (b) the counterfactual world. The counterfactual world is defined as the one where the user's protected attributes were changed while all the other features that are not causally dependent on the protected attributes remain the same [10]. We compute the metric for a given set of individuals (test data), the fractions of the individuals whose decisions are the same in the actual and counterfactual worlds.

5.4.3 Multi-group fairness metrics for Matching, Ranking, and Recommendation Algorithms

In many online platforms today (such as Amazon, Netflix, Spotify, LinkedIn, and Airbnb), there are multiple stakeholders: (i) providers of goods and services, (ii) customers who pay for them, and (iii) the platform which provides the matching between the providers and customers. The platform lies in the centre of the ecosystem, enabling the providers and the customers to connect and do business. Crucially, the platform controls the exposure of service providers to potential customers and vice-versa. For example, Uber matches drivers with passengers, and Swiggy matches delivery drivers to food orders. In the case of Airbnb or different freelance websites, customers have more control over the choice of the provider. However, the platform still decides how much exposure and attention each provider gets and which customers they are shown to through their ranking algorithms. Similarly, recommendation algorithms determine providers' exposure from being recommended to the customers.

5.4.3.1 Ranking and matching algorithms

Incorporating fairness in ranking and matching algorithms is more challenging than classification algorithms. It is because assigning class labels to each observation is an independent task, whereas ranking or matching are dependent tasks that depend on who has been matched or ranked already, and in what order.

Most of the fairness metrics and definitions discussed earlier, including individual and group fairness metrics, are also applicable in these situations. In the Indian context, fairness is ever more critical due to the potential use of automated decision-making algorithms to rank candidates for jobs, selection processes, subsidy disbursal, and other government benefits. India's affirmative action policies aimed at uplifting certain castes may require adjustments to the rankings by design. Thus, in the Indian context, ranking algorithms should be given more care while looking at it from a fairness angle. To be more specific, inter-alia, the following fairness measures need to be assessed while using matching/ ranking algorithms, in addition to the fairness measures mentioned in individual and group fairness metrics:

1. **Proportional representation** in top-ranked or top-matched sets. This is the same as predictive parity, equality of opportunity, and equalized odds but is limited to the automated algorithm's top-ranked set used for selection.
2. **Diversity** in top-ranked or top-matched sets. This is to ensure that members of each group and sub-group (including gender, caste, religion, type, geographical, state, etc.) are adequately represented at the top-ranked or matched positions in proportion to their prevalence in the dataset used as input. Diversity will always have a socio-economic and political context. In the case of India, diversity may include things that are not usually included in the West but are guaranteed by the prevailing law (caste, for example). The definition of diversity can be the same as that of demographic parity stated above.
3. **Procedural (probability-based) fairness** is defined using statistical significance tests that ask how likely it is that a given ranking or matching was created by a fair process, such as by tossing a coin to decide whether to put a protected group or a privileged-group candidate at a given position i [11]

4. **Exposure-based fairness** is defined by quantifying the expected attention received by a candidate, or a group of candidates, typically by comparing their average position bias to that of other candidates or groups [11]

$$\text{Exposure}(\boldsymbol{\tau}(i)) = \mathrm{E}_{\boldsymbol{\tau}\sim\pi}\,[\mathrm{v}(\boldsymbol{\tau}(i))]$$

Here, π: $rnk(\mathrm{C}) \rightarrow [0,1]$ is the probability mass function over the ranking space, and position bias $\mathrm{v}(\boldsymbol{\tau}(i))$ refers to the observation that the customers or users of a ranking system tend to prefer candidates at higher positions and that their attention decreases geometrically or logarithmically with increasing rank. (i) refers to the ranking at position i [12].

When re-ranking or re-matching occurs to meet the fairness criteria, it may be essential to re-assess the algorithm to find the new loss in the objective function (or utility). This loss in performance due to fairness criteria should be within acceptable tolerance limits for the algorithm to function properly.

5.4.3.2 Recommendation Algorithms

Recommendation algorithms can be thought of as one form of ranking algorithm. A recommendation problem can be formalized as selecting a top-N list of items from a set of n items for each of the m users. The fairness of recommendation algorithms can also be divided into group fairness and individual fairness. Similarly, group discrimination for binary classification and group fairness also measures the disparity between two groups defined by the protected attributes in terms of metrics more relevant to the recommendation algorithms. Below we recall a few metrics from [13].

The first metric is value unfairness, which measures the inconsistency in signed estimation error across the user types. This is computed as the average of |(PDj - RDj) - (PAj - RAj)|, where PDj and PAj represent the average predicted score for the jth item from disadvantaged and advantaged users, respectively, and RDj and RAj are the corresponding items for ratings. The average is computed over all items. Value unfairness occurs when one class of users is consistently given higher or lower predictions than their true preferences.

The second metric is absolute unfairness, which measures inconsistency in absolute estimation error across user types. Average | |PDj - RDj| - |PAj - RAj||.

The third metric is underestimation bias, which measures inconsistency in how much the predictions underestimate the true ratings. This is defined as: Average | max(0,(RDj - PDj)) - max(0,(RAj - PAj)|. The fourth metric is overestimation bias, which measures inconsistency in how much the predictions overestimate the true ratings, computed as Average | max(0,(PDj - RDj)) - max(0,(PAj - RAj)|. Underestimation is important when missing recommendations are more critical than extra recommendations. At the same time, overestimation is important in settings where users may be overwhelmed by recommendations, so providing too many recommendations would be especially detrimental.

Other metrics compare the disparity between scores like F1 and NDCG [14] between the advantageous and disadvantageous users [15].

For detecting individual fairness, Li et al. [16] consider counterfactual fairness in the recommendation, which requires that the recommendation results for each user are the same in the factual and the counterfactual world.

5.4.3.3 Group fairness metrics for clustering

Balance-based metrics [17]: Say there are m protected groups, and k is the number of clusters. Given a protected group p and a cluster c, say p0 is the proportion of samples belonging to the group, and p1 is the proportion of samples in the cluster belonging to the group. The ratio b = p1/p0 for each protected group and cluster is computed. The minimum value of b, 1/b for all the combinations is the balance metric, whose value will range between 0 and 1. The value one essentially represents a completely fair balance. We advise the value of the balance metric to be more than 0.8 based on the 4/5th rule.

5.4.3.4 Combining Metrics

1. **Bias Index**: Different users might use different metrics to check the fairness of an AI system. Hence, it is crucial to standardise the bias measurement on a linear scale so that a uniform scale can be used to assess fairness and compare different AI systems [18]. Bias Index is defined for each protected attribute in the system as follows:

$$BI_{\mathrm{i}} = \sqrt{\frac{\sum_{j=1}^{n}(M_{\mathrm{ij}} - M_{\mathrm{j}}')^2}{n}}$$

where,
i: number of the protected attributes
j: number of fairness metrics used
n: total number of fairness metrics used
m: total number of protected attributes considered in the AI system
Mij: value of the jth fairness metric for the ith protected attribute
M′j: ideal value of the jth fairness metric i.e., 0 for difference metrics and 1 for ratio metrics

2. **Fairness Score:** Fairness Score [18] is defined for the AI system as follows:

$$FS = 1 - \sqrt{\frac{\sum_{i=1}^{m}(BI_{\mathrm{i}})^2}{m}}$$

Substituting the equation for BI, we get,

$$FS = 1 - \sqrt{\frac{\sum_{i=1}^{m}\sum_{j=1}^{n}(M_{\mathrm{ij}} - M_{\mathrm{j}}')^2}{mn}}$$

While Bias Index corresponds to the degree of bias for a particular protected attribute in a dataset or model, Fairness Score corresponds to the degree of fairness in the entire model, considering all the protected attributes together. For a fair system, the Bias Index for each protected attribute should be zero, and the Fairness Score should be one. AI systems may have more than one protected attribute, so there would be as many Bias Indexes as the number of protected attributes, but there will be only one Fairness Score for the model.

The AI system might be biased for some of the protected attributes and fair for the other protected attributes, which the corresponding Bias Indexes would reflect; Fairness Score, on the other hand, would reflect the overall fairness.

5.5 Metrics selection

1. Different metrics may be required to check for biases in the pre-processed training dataset and the outcomes.
2. A single metric for each part (pre-processed training dataset and the outcome being two parts) might not correctly identify bias in all cases, so using a combination of metrics is recommended.
3. The dataset might have multiple attributes affecting the system's fairness. It is necessary to check for fairness for each protected attribute.
4. Different AI systems may use different fairness metrics, so the combining fairness metrics used for rating should be flexible to accommodate distinct fairness metrics and a different number of metrics, to compare various supervised learning AI systems [18].
5. Determining appetite and tolerance for fairness in the process is essential as this helps decide the fairness assessment basis for the same. Refer to section 6 for details.
6. Affirmative actions by Government (both State and Central) bring demographic variations wherein standard group parity metrics may be ineffective.

5.6 Limitations of fairness metrics

1. Fairness metrics represent the statistics but do not exhibit the root cause of the bias. Miscalibration, sample bias, label bias, sub-group validity, disparate error rate, redlining, and even outlier removal can lead to bias. These root causes require context-specific examination of the process and model beyond the metrics.
2. Fairness metrics focus on a mathematical representation of bias and may not be best suited to measure edge cases and/ or long tail/ fat tail risks. Additional

consideration may be needed in special cases such as small positive class and/or insignificant minority groups, significantly skewed misclassification penalties.

6.0 Proposed Assessment Framework

The Fairness assessment framework has two parts: (A) the dimensional view of the contributors of bias, and (B) the approach towards conducting bias assessment [6].

6.1 Dimensional view of contributors of bias

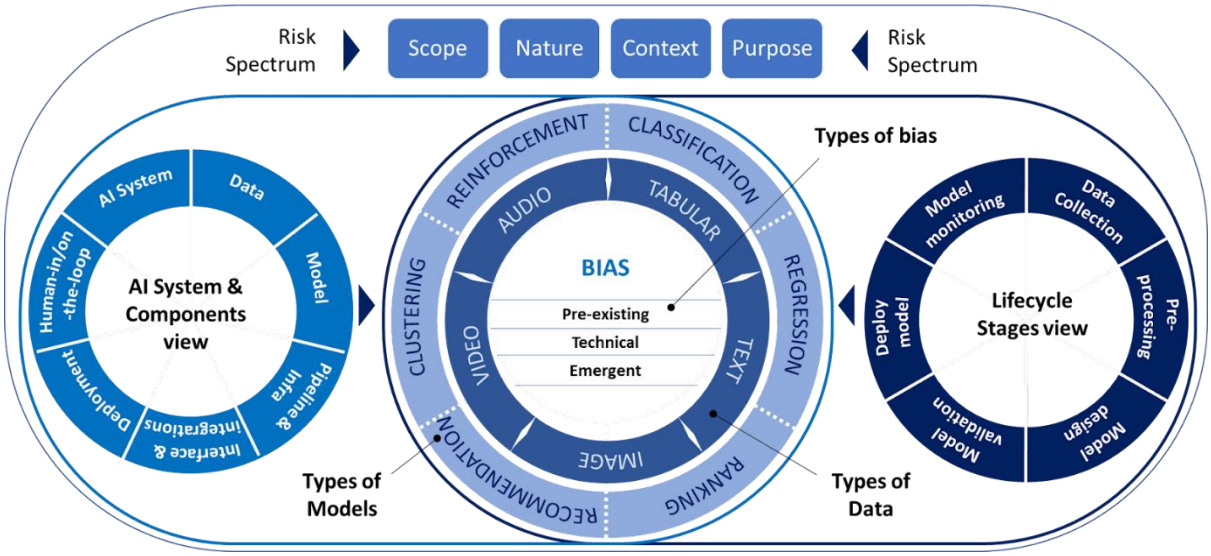


Figure 2: Dimensional view of contributors of bias

The fairness assessment framework provides a multi-dimensional approach to examining biases in an AI system.

- **Dimension 1:** Types of bias: There are three types of biases, namely, pre-existing, technical, and emergent biases. Data is the best place to look for pre-existing bias, whereas technical bias can be present in both data and model. Emergent bias can typically be checked in retrained data or retrained models.
- **Dimension 2:** Types of data: There are different data modalities, including tabular, text, image, video and audio, etc. The procedure for detecting biases may be different for different data types. For example, a common form of discrimination in text data is due to the encoding of the text input.
- **Dimension 3:** Types of models: There are four broad types of machine learning models, namely, supervised (classification and regression), semi-supervised, unsupervised (ranking, recommendation, and clustering), and reinforcement learning. Separate model-fairness assessments are proposed for different

types of models. Another dimension in models is about access - open, grey, or closed box, which can also determine the method performed for fairness assessment.

- **Dimension 4:** Types of components: The bias assessment algorithm is different for different components, viz. the AI system, data, model, interfaces, pipeline, infrastructure, etc. For example, the set of fairness metrics is different while testing the data than that for testing a model.
- **Dimension 5:** Types of lifecycle stages: Bias assessment at different stages of the AI lifecycle poses distinct challenges. For example, checking a bias post-deployment needs to be performed using the real workload, whereas, at build time, it can be done using the training and test data.
- **Dimension 6:** Types of risk: Understanding and determining the risk associated with bias in the AI system can help in deciding the amount of test data required, the variation of test data needed, and establishing the acceptable thresholds for risk evaluation based on the risk spectrum. The risk spectrum contains the AI system's scope, nature, context, and purpose under consideration.

6.2 Approach to bias assessment

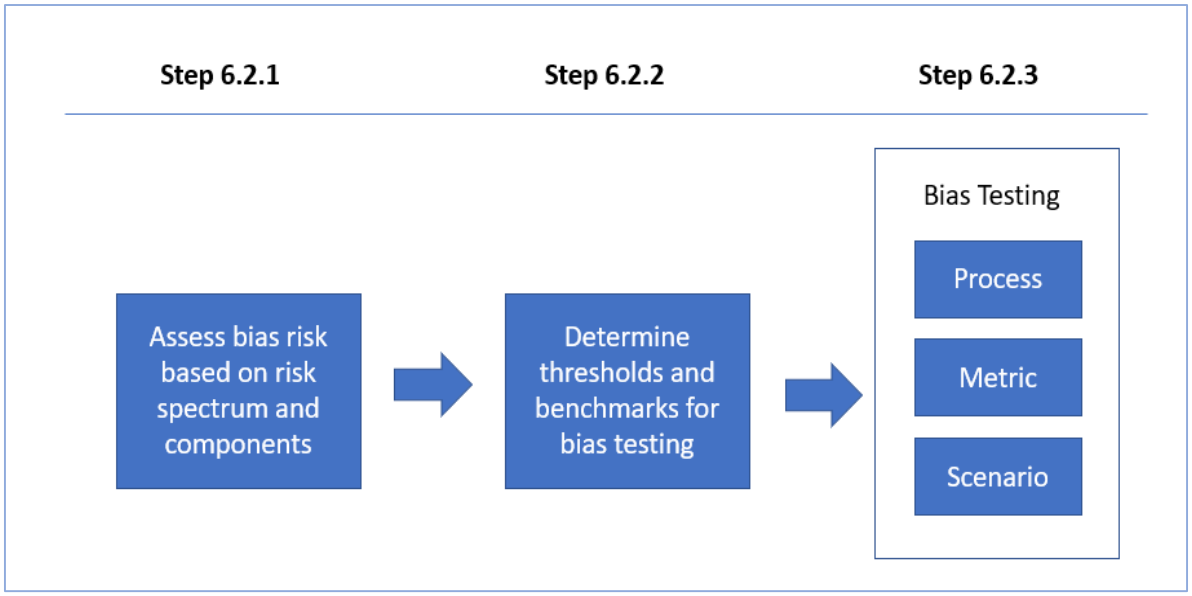


Figure 3: Three-step approach to bias assessment

Bias assessment shall be approached with three steps, namely, (a) bias risk classification of the AI system and assessing bias contributors in the AI system, (b)

determining the appropriate metrics to be used, their thresholds, and the benchmarks for bias, and (c) bias testing mechanism to validate and assess the extent of impacts caused by them. Throughout this process, it is assumed that the auditee and the auditor assess the bias for the same protected attribute(s). In most AI applications, the choice of the protected attribute is straightforward. However, the protected attribute must be a piece of common knowledge between the auditor and the auditee.

This standard proposes an approach to testing bias with methods based on (a) Process, (b) Metrics and measures, and (c) Scenarios. Details are as follows

Bias risk classification of an AI system and assessing bias contributors can be done using a preliminary questionnaire to gather contextual understanding. The auditee prepares the bias risk assessment (6.2.1) and determines the appropriate metrics, their thresholds, and the benchmarks for bias (6.2.2). The auditor then verifies the appropriateness of such classification based on transactional and documentary evidence. The auditor uses the bias risk assessment and evaluation of metrics, thresholds, and benchmarks to determine the tests (6.2.3) performed on the AI system to validate the extent of the impact caused by bias.

Section 6.2.1 uses risk analysis to determine the risk level. The accuracy of risk analysis depends on the quality of the risk inputs. Risk analysis is formulated based on the likelihood of harm and the seriousness or impact of such harm to people, community, and nation/ state, caused by the autonomous decision. The auditee prepares for steps 6.2.1 and 6.2.2 by conducting a risk analysis and defining appetite and tolerance levels for bias in metrics, thresholds, and benchmarks. The auditee shall document the facts relating to each question and assign a risk rating for each question in 6.2.1. The auditor shall assess the risk analysis for appropriateness and determine the extent of bias testing required for certification.

Risks shall be classified as high, medium, and low based on the following principles:

1. To what extent will the results affect the user’s life?

2. To what extent will the outcomes affect users' rights and freedom as per the constitutional and ethical considerations??
3. To what extent it intends to uphold the principles of equality and principles of inclusivity, and non-discrimination?

An illustrative representation of the risk consideration is provided below:

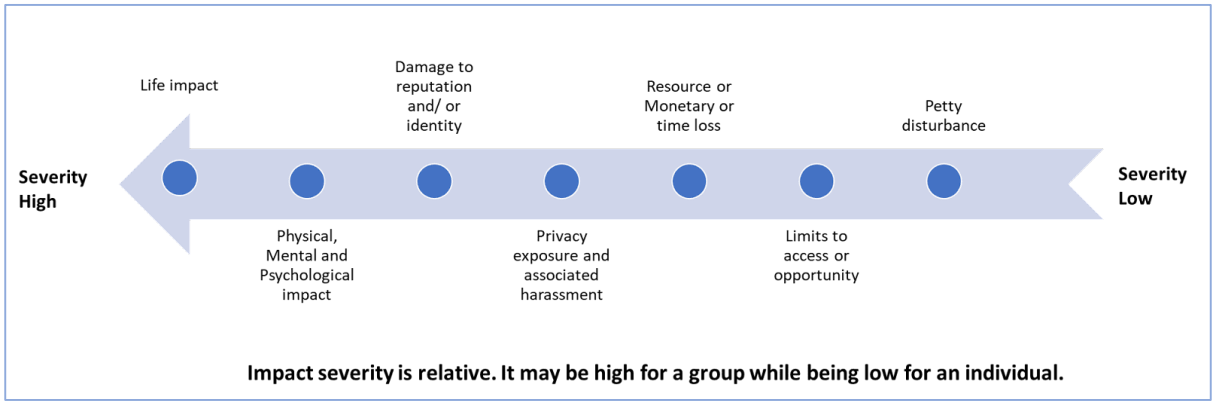


Figure 4: An illustrative representation of the risk consideration; Source: ForHumanity

1. For High risks:
    - AI results may have a bearing on the safety and security of individuals or may determine their critical life decisions (e.g., disease diagnosis algorithms and autonomous vehicles).
    - AI is anticipated to impact individuals' eligibility for certain benefits, thereby significantly impacting rights or freedom.
    - AI aims to address existing societal biases.

2. For Medium risks:
    - AI results may affect individuals' convenience or financial choices.
    - AI is anticipated to impact individuals' eligibility for certain benefits to a limited extent due to public interest requirements.
    - AI intends to support determining prioritisation for essential services.

3. For Low or no risks:
    - AI results may have limited or no bearing on individuals and may not impact them physically or financially.

- AI intends to have limited or no impact on potential access to services or opportunities (social or technical accessibility).

The auditee shall document the risk score of the AI system by considering the risk contributed by each assessment referred to in Steps 1a and 1b below.

6.2.1 Bias risk classification

Bias risk classification is a self-assessment questionnaire that helps the auditee to determine the bias risk level associated with the identified AI system and the component contributors of bias. It helps distinguish AI applications according to their potential impact on individuals, society, and the planet.

The users shall assess risks for each of these questions in each section (AI System, Data, Model, Pipeline, Interface & Integration, Human-in-the-loop, Deployment) based on facts and provide detailed responses for each of these questions. The risk assessment shall be consolidated at a section level to see whether they have a majority of questions representing a specific risk level (High/ Medium/ Low). The threshold determination (6.2.2) and testing process (6.2.3) shall be proportionally considered based on the assessed risk levels for relevant components and the AI system as a whole. Regarding testing, this assessment can help with the amount of test data required, the variation of test data required, and testing frequency, and support in establishing the acceptable thresholds for risk evaluation based on the risk spectrum.

A representative set of questions for Bias risk classification [19, 20] are as follows:

AI System
• Business model: Is the system a for-profit use, non-profit use, or public service system? (Note: Public services may increase risk). • Impacts critical functions/ activities: Would disrupting the system's function or activity affect essential services/ life or death decisions on behalf of users?

(Note: supporting life or death decisions may increase risk. For example, decision support systems including chatbots especially used in legal and medical advice.)

- Impacted stakeholders: Who are impacted by the system (e.g., consumers, workers, government agencies)? (Note: Impact on the wider public may increase risk)
- Autonomy: Does the system impact the autonomy of the individuals?
- Business considerations: Does the AI system have business considerations inconsistent with generally understood fairness expectations? (Note: e.g. Lending business rules that limit transactions with customers from certain geography or community)
- Regulatory attention: Is the industry or the use case known to have significant regulatory attention? (Note: Increased regulatory attention may increase risk)
- Adverse incidents: Are there any known adverse incidents in systems of this type (specifically regarding fairness)?

Data

- The provenance of data and input: Are the data and inputs from experts provided, observed, synthetic, or derived? (Note: less provenance increases risk)
- Dynamic nature: Are the data dynamic, static, dynamic, updated from time to time, or real-time? (Note: More dynamic nature may increase risk)
- Timeliness: Is the data timely and relevant given the context of the AI system's purpose?
- Data sources: Where the data is gathered from multiple sources, whether the source tools are calibrated (e.g., sensors for gathering data)?
- Nature of data: Is the dataset collected from a known and verifiable data repository?
- Data labelling: In the case of manual labelling, are the people labelling the data trained and aware of the context of the problem?

- Data appropriateness: Is the training data relevant and representative of the use case? Is the deployment environment consistent with the training environment?
- Data representativeness/sufficiency: How distributed is the dataset? Does it include data from all sections of society? Is it skewed in favour of certain groups?
- Data quality: Whether the quality of data is exposed to the risk of bias? Whether the treatment of synthetic data, data imputations, outliers, and duplicate data point to the potential for bias? How noisy is the data? (Note: Data quality and risk are inversely proportional)
- Detection and collection: Are the data and input collected by humans, automated sensors, or both? (Note: Quality issues associated with the human collection and sensor calibration will determine the risk)
- Inferences and proxies: Are there inferences and proxies in the model? Whether the inferences and proxies are validated with events in the real world? (Note: higher number of inferences or proxy variables increases risk)
- Retraining process: Whether the methods adopted for retraining are consistent? Are there measures to avoid possible leakage of prediction caused by retraining methods? What were the deciding factors for retaining or discarding the old data while retraining the model?

Model & Pipeline

- Deterministic and probabilistic: Is the model used in a deterministic or probabilistic manner? (Note: Deterministic nature may lower risk)
- Model-building from the machine or human knowledge: Does the system learn based on human-written rules, from data, through supervised learning, or reinforcement learning?
- Single or multiple model(s): Is the system composed of one model or several interlinked models?
- Model robustness: Is the model exposed to data poisoning attacks that may lead to potential biases? (Note: adversarial vulnerability is directly proportional to the risk)

- Parametric choices: Whether the choice of metrics, parameters, and benchmarks well documented and available for external auditing and explainability? (Note: Parametric choices that do not consider or evaluate bias implication usually increase risk)
- Causality: Whether the model has verifiable causality? Is there any unvalidated causality within the data? (Note: Models with less causal verifiability may increase risk)
- Pre-trained model: Is the AI system using a pre-trained model or API? (Note: use of untested pre-trained models may increase risk)
- Adverse impacts: Are there any foreseeable adverse impacts for the domain, or use case resulting in bias? (Note: Known potential for adverse incidents may increase risk)
- Feature selection: Were any features containing a protected attribute or proxy dropped? If yes, why and what effect do they impact bias?
- Model development: Whether sufficient considerations for fairness are undertaken in the process relating to model selection, objective definition, tuning, and metric choices?
- Pipeline quality: Is the model assessed for pipeline quality, including errors, defects, and inconsistent uncertainty calibrations?

Integration and Interface

- Combining tasks and actions into composite systems: Does the system combine several tasks and actions (e.g., content generation systems, autonomous systems, control systems)?
- Integration quality: Is the AI system assessed for integration quality (on model or application integrations), including defects, failures, and adversities resulting in bias?

Deployment

- Deployment environment: Is the AI system employed for public purposes or in government/ enforcement activities? (Note: Public purposes will increase risk)

• Model deployment: Whether the model has any social or technical accessibility challenges in the deployment environment?
Human-in-the-loop or Human-on-the-loop
• Action autonomy: How autonomous are the system's actions, and what role do humans play? • HITL: Are the models evaluated for the human-in-the-loop (HITL) effectiveness in the process and the bias contributed by the HITL? • Decisions on tradeoffs: Are there trade-offs relating to fairness? How are they adjudicated?

6.2.2 Determining the fairness metrics, thresholds, and benchmarks

Metrics represent the absolute values of the outcomes. Measuring bias would require defining bias at a metric level also. This is usually done by using a threshold in the values, where a breach (downward or upward, depending on the context) of the threshold represents bias.

In the US, [21] guidelines provide a method to identify disparate impact. Disparate impact, also called adverse impact, assesses the differences in the selection rate of groups.

It is common to determine the algorithm as unfair if the ratio of the selection rates for the groups is less than 4/5th of the group with the highest rate (80%), then this scenario is considered an adverse impact (has a discriminatory effect) on a group. This threshold helps in assessing the models for minimum acceptability. However, such a metric may create more discrimination when applied in a certain context than others. For instance, AI used for disease diagnosis or patient treatment suggestions may require an equal threshold (equal selection rates) between groups, as using a 4/5th rule may lead to misdiagnosis or mistreatment resulting in harm to certain groups.

For this reason, metrics such as the U.S. Equal Employment Opportunity Commission (EEOC) metrics are useful but do not fit the need for a standard. Given the limitations of such metrics, a 3-step approach as described below is proposed to enable better evaluation of the disparate impact.

Step 1: Determine the risks relating to the AI system based on risk assessment and determine metrics, measures, and/ or thresholds or benchmarks based on the risks. Refer to sections 6.2.1 and 6.2.3 for details.

Step 2: Gather any Sectoral or domain or geography or use-case specific requirements including benchmarks or thresholds, or additional fairness considerations placed by the appropriate authority or the body demanding conformity to the standard (e.g. Government or United Nations or a Private entity buying an AI system). Assess if the results of Step 1 meet the requirements.

Step 3: Report the absolute metric values allowing the report readers to interpret the results in context.

6.2.3 Bias testing

6.2.3.1 Process testing

Testing methods to examine the process include data collection, annotation, cleaning, pre-processing, testing, validation, and post-market monitoring. Broad-level guidance is provided as follows. Refer to section 5 for details regarding the contributors of bias for determining the process-level tests that need to be undertaken in the context of specific AI systems.

1. Public impact: In circumstances where the model is intended for public service, supports life or death decisions, reduces autonomy, impacts rights, or uses personal data for determining benefits, conduct various scenario-based tests to identify instances of bias caused by the edge cases or adversarial examples.
2. Regulatory focus: In cases with significant regulatory attention for bias in the industry, examine the history of reported and assessed violations by AI developers or enforcement actions by the regulators. Assess the applicability of such circumstances and include such instances in test data.

3. Adverse incidents: Collate the list of reported adverse incidents relating to bias (as applicable to the AI system's use case, domain, and industry). Test for bias using such scenarios to identify if the model is robust against reported adverse incidents in the marketplace.
4. Data Provenance: Conduct specific tests on the data with common provenance for bias risk.
5. Dynamic data: Modulate the frequency of testing for bias based on whether the data is updated time-to-time or in real-time.
6. Data quality: Conduct specific tests on data derived from the suboptimal quality of labels or the labelling process, inconsistent clustering of users for profiling, and suboptimal quality of pre-processing (including imputation, outlier treatment, etc.) for bias.
7. Data Privacy: Gather specific considerations aligned to the privacy requirements as guided by relevant regulations or industry requirements.
8. Proxies and inferences: Ensure to include proxies or inferences as protected categories to understand if they contribute to bias.
9. Deterministic models: Inspect the rules considered in the deterministic models and evaluate if they have a condition on protected attributes and/ or contribute to unfair treatments or bias.
10. Multiple models: Test models independently when they are an ensemble or multiple models contributing to one outcome.
11. Parametric choices: If testing represents the bias in the model, then examine the parametric choices for the model (including hyper-parameters) to determine if the choices or combinations thereof contributed to the model's bias.
12. Pre-trained models: Compare the reported bias of pre-trained models (basis system cards or model cards, if exists) with the bias contributed by using pre-trained models.
13. Retraining: Develop test data for testing data or model quality issues contributed by the retraining process.

6.2.3.2 Metrics and measures

Testing methods that examine the AI system to assimilate key metrics and measures to test for bias, including group and individual fairness metrics, monotonic risk metrics, etc. Broad-level guidance for bias testing using metrics and measures is as follows. Refer to section 6 for details of metrics that can be considered.

The following fairness assessment procedures typically cater to the three types of personas - 1) developer of the system, 2) internal tester or auditor, and 3) external auditors. Each persona has a varied level of information access and therefore the following procedures cater to different levels of information in the Data and AI lifecycle.

Data Fairness Assessment

Typical data lifecycle consists of 1) data collection/ annotation and data merging, 2) data quality checking and pre-processing, 3) encoding, and 4) splitting to generate training and test data. Developers/ data scientists typically have access to all such phases, but other personas may have limited access.

Developers using this standard are advised to run the data fairness assessment procedures after each data lifecycle stage to reveal the sources of bias. At least applying and reporting fairness assessment on the unencoded training data is highly recommended. A recommended standard data assessment procedure for building a tabular classification model is outlined below:

1. Identify protected attributes. See Section 5.1
2. Identify privileged and unprivileged classes. See Section 5.2
3. Identify favourable outcomes. See Section 5.3.
4. Compute the demographic parity metrics. See Section 5.4. Other metrics require predicted labels and, therefore, cannot be checked on static data.
5. Compute individual discrimination with the training and test data (see Section 5.4), considering the label present in the data but not in the model.
6. Check the result with the allowable thresholds (see Section 5.5).

7. Report all metrics results.

The above assessment should be carried out after every stage in the data lifecycle that performs any change in the data, except after standard encoding steps that typically do not introduce any bias. A detailed description of such scenarios is described in Section 6 - 6.2.2. Such checking determines the sources of bias in different data processing stages. In addition, the following two scenarios require special attention.

If the data source represents a protected attribute (e.g. different samples are collected from different countries), compute the ratio of favourable outcomes per data source. Take the maximum variation (either difference or ratio based) of such ratio; such variation should be in the allowable threshold for demographic parity. Note that the data source is not part of the data but a piece of extra metadata. In this case, evaluation of demographic parity, including such metadata, is recommended.

The train-test split divides the processed data into two parts containing training data and test data, or three parts containing training, validation, and test data. In general, such splits should be done uniformly at random such that the distribution of data remains the same in all the splits. However, the split can result in bias even if the processed data does not contain bias. Recheck the above methods separately to ensure that none of the splits (especially training data) contains any bias.

Note that access to the data is only possible under open-box and grey-box scenarios (as defined in section 2). In the case of closed-box access, the users may directly perform the model assessment phase.

Model Fairness Assessment

Model fairness requires the user to evaluate different metrics described in Section 5.4. As in the case of data, this version focuses on the fairness assessment of models built on tabular data. The standard operating procedure for tabular classification model assessment is as follows:

1. Identify protected attribute(s) (refer to section 5.1). One can continue with the same attributes used in data fairness evaluation. The same goes for the next two steps. Note that protected attributes can also be proxies.
2. Identify privileged and unprivileged classes (refer to section 5.2).
3. Identify favourable outcomes (refer to section 5.3).
4. Identify test data for evaluation. Consider inputs from the procedures in 6.3.2 in developing the test data.

This step depends on the level of access available to the evaluator. The aim is to identify sufficient test data that resembles the training data distribution. Different scenario-based testing with other goals of forming test data is discussed later.

If open-box or grey-box access is provided to the model, test data generated from the train-test split can be used for testing fairness. The test data volume is recommended to be at least 20% of the training data with a lower limit of 1K samples.

In case of insufficient available test data or closed-box access (without access to training or test data), use synthetic data generators [22] to create sufficient test data. Ensure that test data have a similar distribution to the training data. The data generation process can be offline or online. Offline data generation performs the evaluation procedures (5 and 6) on the available or generated test data. However, an online procedure starts with the evaluation procedure with available test data. It generates more test data in the neighbourhood of the failure samples, thereby obtaining more failure samples iteratively [23]. This process ensures finding more failure samples, especially for individual discrimination. Note that the process of generating synthetic data may not generate any gold standard labels as metrics such as demographic parity (group discrimination) and flip rates (individual discrimination) do not require gold standard labels. The objective of using synthetic data is to perform fairness testing and not to use it for bias mitigation purposes.

1. Compute the group discrimination metrics described in section 5.4, considering model output.

2. Compute the individual discrimination metrics described in section 5.4, considering model output.
3. Plot the various metrics on a graph and see if all the metrics are within the tolerance band.
4. Calculate the Bias Index for each protected attribute and the Fairness Score for the overall AI system [18].
5. Report all metrics results.
6. Select appropriate threshold values based on the justification of the risk factors associated with the application (refer to section 6.1) and evaluate whether bias exists for each metric. Document the justifications and assumptions for threshold value selection for self-certification. The certifying auditor should validate such assumptions.
7. As the system is operationalized and keeps learning from real-world data, it is necessary to check periodically for any biases introduced after the initial tests. As such, periodic recertification is recommended [18]. Refer to section 8 for details.

6.2.3.3 Scenario testing

This subsection explores the testing methods that examine the AI system on scenario-based testing, including test examples that represent counterfactual fairness, edge cases, long tail/ fat tail risk, and adversarial testing. Scenario testing aims to ensure that previously unidentified categories or instances of bias are identified through scenario-based testing.

1. Data characteristics: Develop test data for scenarios that consider source data sensor failures or calibration issues, incomplete or non-representative data, and unusual operating scenarios to test for bias.
2. Model and pipeline quality: Develop test data for testing scenarios that represent the inconsistent inputs channelized from other models, and realistic data distributions to understand its impact on the bias.

3. Counterfactual: Develop test data of selected protected attributes that may influence the distribution shift of the existing training set for counterfactuals.
4. Causality: Develop test cases for testing the model for causality including counterfactual fairness.

The standard operating procedure described in the previous subsection may not be suitable for some of the scenarios. The following procedure may be followed for test data generation in such scenarios. The rest of the procedure remains the same as above.

Simulating test data for user-defined scenarios:

1. Edge case - Alter or modify the distribution of scenarios based on known boundaries of the model.
    a. Create an open-box surrogate model
    b. Establish the model boundaries based on the surrogate model
    c. Create the test data in the regions of the boundary [24, 25]
2. Adversarial examples - Perturbations contributing to model misclassifying outcomes
    a. Alter or modify the protected attribute and generate samples to test as an adversary
    b. Determine the perturbation threshold for each feature (other than the protected attribute)
    c. Create test data based on the perturbation threshold
3. Change in distribution
    a. Use the method referred to in [21, 24] or alternative methods for simulating synthetic data for user-defined constraints.

The evaluators should document the test data selection process and attach the test data used for evaluation.

7.0 Fairness evaluation outcome report

Once the evaluation has been completed, a report may be generated bringing out all the assumptions, observations, metric values, risk profiles, types of tests performed, limitations, etc.

7.1 Structure of the report

The report may have three broad sections:

1. Summary
2. Tabulation of all metrics, their thresholds, and measured values
3. Detailed report of the assumptions, processes followed, tests performed, etc.

7.1.1 Summary

The summary section of the report may indicate the following:

1. A synopsis from the auditor of the overall assessment, risk profiles, processes and scenarios tested, etc.
2. A brief overview of the AI application covering its use case, target audience, the geographical and demographic spread of users, and so on.
3. List of all protected attributes.
4. Privileged and unprivileged classes for each protected attribute.
5. What was considered the favourable outcome?
6. Who was the auditor - whether it is self-testing or by an independent auditor?
7. Level of access to data to the auditor - training, validation, and test datasets.
8. Type of testing - open, grey, or closed box.
9. Dependencies on the developer for the evaluation process.
10. Limitations – what kind of checking was not done?

7.1.2 Tabulation of all metrics, their thresholds, and measured values

This section of the report provides the details of various fairness metrics used at different stages and for different protected attributes, their thresholds, and measured values. It may contain:

1. Tabulate all the fairness metrics used for each protected attribute for the training dataset, validation dataset, testing dataset, and AI model testing and evaluation. Mention the threshold and the measured values of each metric.
2. Mention the Bias Index value for each protected attribute by combining the individual fairness metrics, wherever applicable.
3. Indicate the overall Fairness Score of the model by combining the various Bias Index values.

7.1.3 Detailed report

The detailed report should provide the support material for the contents mentioned in the previous two sections of the report. It should also include the assumptions made, justifications for decisions taken, disclosures from the auditee, etc. This section may include the following:

1. AI system description.
2. Type of data, model, interfaces, pipelines, etc.
3. Response to the questionnaire provided by the auditee (section 6.1).
4. Risk assessment outcome (section 6.1).
5. Basis of deciding the protected attributes, privileged and unprivileged classes for each protected attribute, a favourable outcome, indirect proxies, etc.
6. Basis of selecting the fairness metrics (section 6.3).
7. Basis of determining the thresholds (section 6.2).
8. How the data was split into testing, validation, and test data?
9. Whether synthetic data was used? If yes, then how was it generated? Report characteristics of synthetic data like proportions of minorities/majorities in protected attributes and distribution of other attributes.
10. What scenarios were tested?

11. Details of testing the intermediate steps, input, and output processes (section 6.3)
12. Details of the developer's involvement and dependencies on the developer for the assessment.
13. Whether the report is a result of self-certification or independent third-party audit along with their details along with other lineage data such as dates of evaluation.

7.2 Validity of the report

1. Self-certified: Self-certification shall be valid for three months or less as decided by the certifier.
2. Audited: Maximum period of 6 months or as indicated on the certificate, whichever is less.
3. Mandated: Maximum period of 1 year or as indicated on the certificate, whichever is less.

The certification validity period mentioned above would apply to systems that are not continuously trained (online systems). In cases where training is continuous during deployment, the certificate needs to be renewed at a higher frequency as decided by the certifying agent. Such frequency shall be at least once a quarter. The preparator of the report is accountable for keeping it up to date and sending the updated report to the consumer of its previous version.

8.0 Limitations and Scope

Section 6 mentions the various dimensions contributing to bias in an AI system. The scope of this version of the standard covers the limited points in that space as mentioned below:

1. Types of bias: Section 6 covers the data/ model assessment procedures for all types of bias, viz, pre-existing, technical, and emergent. Section 5 covers both group and individual bias.
2. Types of data: This version covers tabular data where each row is independent of the other. The standard covers test data generation procedures for tabular data. The future versions of this standard may cover other types of data, such as text, image, speech, etc., and different models built on the data.
3. Types of models: This standard presents metrics for evaluating bias in all models for tabular data. It covers the method for bias testing open, grey, and closed box models. Future versions may cover other types of models e.g. Reinforcement, GAN, and Autoencoder.
4. Types of components: The current draft covers fairness testing of data and models. Future versions may cover other components such as interfaces, pipelines, infrastructure, and deployments.
5. Type of lifecycle stages: This version covers the data lifecycle, model build lifecycle, and counterfactual deployment scenarios.
6. Types of risk: Section 6 presents the questionnaire related to risk evaluation in AI systems, models, and data. Future versions may cover risk evaluation for other components also.
7. Type of metrics: While this version of the standard covers some examples of different types of bias and metrics, it's not feasible to cover all types of biases that can arise in different procedures and lifecycle phases, depending on the nature and domain of the system and its underlying use cases.
8. Bias Mitigation: This version does not prescribe how bias mitigation should be done and is left to the developer of the AI system.

9.0 References

[1] National Digital Communications Policy 2018 https://dot.gov.in/sites/default/files/EnglishPolicy-NDCP.pdf

[2] The Frontiers of Fairness in Machine Learning, by Alexandra Chouldechova and Aaron Roth, Communications of the ACM, 2020, ACM New York, NY, USA.

[3] Responsible AI by NITI Aayog, India. https://www.niti.gov.in/sites/default/files/2021-02/Responsible-AI-22022021.pdf

[4] ISO/IEC 22989:2022

[5] Wachter, S., Mittelstadt, B. and Russell, C. (2021). Bias Preservation in Machine Learning: The Legality of Fairness Metrics under EU Non-Discrimination Law. SSRN Electronic Journal. [online] Available at: https://papers.ssrn.com/sol3/papers.cfm?abstract_id=3792772

[6] Bias Risk Assessment- A systematic approach, Sundar Narayanan, https://medium.com/@sundar-narayanan/bias-risk-assessment-a-systematic-approach-part-1-2-79e209d0e162

[7] Friedman, B. and Nissenbaum, H. (1996). Bias in computer systems. ACM Transactions on Information Systems, 14(3), pp.330–347.

[8] Dwork, Cynthia, Hardt, Moritz, Pitassi, Toniann, Reingold, Omer, and Zemel, Richard. Fairness through awareness. In Proceedings of the 3rd Innovations in Theoretical Computer Science Conference, pp. 214–226. ACM, 2012.

[9] Meritocratic Fairness for Cross-Population Selection by Michael Kearns, Aaron Roth, Zhiwei Steven Wu Proceedings of the 34th International Conference on Machine Learning, PMLR 70:1828-1836, 2017.

[10] Counterfactual Fairness, Matt J. Kusner, Joshua R. Loftus, Chris Russell, Ricardo Silva, https://arxiv.org/abs/1703.06856

[11] Meike Zehlike, Ke Yang, Julia Stoyanovich, Fairness in ranking, a survey, 2022, arXiv:2012.14000v

[12] Ricardo Baeza-Yates. 2018. Bias on the web. Commun. ACM 61, 6 (2018), 54–61. https://doi.org/10.1145/3209581

[13] Yao, S. and Huang, B., 2017. Beyond parity: Fairness objectives for collaborative filtering. Advances in neural information processing systems, 30.

[14] SK Learn Metrics https://scikit-learn.org/stable/modules/generated/sklearn.metrics.ndcg_score.html

[15] Yunqi Li, Hanxiong Chen, Zuohui Fu, Yingqiang Ge, and Yongfeng Zhang. 2021. User-oriented Fairness in Recommendation. WWW (2021)

[16] Yunqi Li, Hanxiong Chen, Shuyuan Xu, Yingqiang Ge, and Yongfeng Zhang. 2021. Towards Personalized Fairness based on Causal Notion. SIGIR (2021)

[17] Chierichetti, F., Kumar, R., Lattanzi, S. and Vassilvitskii, S., 2017. Fair clustering through fairlets. Advances in Neural Information Processing Systems, 30.

[18] Agarwal, A., Agarwal, H. & Agarwal, N. Fairness Score and process standardization: framework for fairness certification in artificial intelligence systems. AI Ethics (2022). https://doi.org/10.1007/s43681-022-00147-7

[19] OECD Framework for the Classification of AI Systems: a tool for effective AI policies, https://oecd.ai/en/classification

[20] Agarwal, A., Agarwal, H. A seven-layer model with checklists for standardising fairness assessment throughout the AI lifecycle. AI Ethics (2023). https://doi.org/10.1007/s43681-023-00266-9

[21] US Equal Employment Opportunity Commission (EEOC)

[22] Synthetic Data generator https://github.com/sdv-dev/SDV

[23] Aniya Aggarwal, Pranay Lohia, Seema Nagar, Kuntal Dey, Diptikalyan Saha. Closed box fairness testing of machine learning models. ESEC/SIGSOFT FSE 2019: 625-635

[24] ISO/IEC TR 24027:2021 Information technology — Artificial intelligence (AI) — Bias in AI systems and AI aided decision making https://www.iso.org/standard/77607.html

[25] Diptikalyan Saha, Aniya Aggarwal, Sandeep Hans: Data Synthesis for Testing Closed-Box Machine Learning Models. COMAD/CODS 2022: 110-114

[26] Bellamy, Rachel KE, Kuntal Dey, Michael Hind, Samuel C. Hoffman, Stephanie Houde, Kalapriya Kannan, Pranay Lohia et al. "AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias." IBM Journal of Research and Development 63, no. 4/5 (2019): 4-1.

10.0 Abbreviations

Abbreviation	Expansion
AI	Artificial Intelligence
API	Application Programming Interface
BI	Bias Index
CAB	Conformity Assessment Body
DOT	Department of Telecommunications
EEOC	U.S. Equal Employment Opportunity Commission
ER	Essential Requirement
FN	False Negative
FP	False Positive
FS	Fairness Score
GAN	Generative Adversarial Network
HIL	Human-in-the-loop or Human-on-the-loop
ICT	Information and Communication Technology
ML	Machine Learning
MTCTE	Mandatory Testing and Certification of Telecom Equipment
NDCG	Normalized Discounted Cumulative Gain
NGO	Non-governmental Organization
RTEC	Regional Telecommunication Engineering Centre
SME	Small and Medium Enterprises
SOP	Standard Operating Procedure
TEC	Telecommunication Engineering Centre
TN	True Negative
TP	True Positive
UI/ UX	User Interface and User Experience

11.0 Acknowledgements

- Dr. Diptikalyan Saha, Senior Technical Staff Member, IBM Research, India
- Dr. Abhijnan Chakraborty, Assistant Professor, Department of Computer Science & Engineering, IIT Delhi
- Dr. Anand Mahalingam, Vice President - Data Labs HDFC Life, Bangalore
- Dr. Ganesh Ghalme, Assistant Professor, Department of Artificial Intelligence, IIT Hyderabad
- Mr. K V Tirumala, Joint DGFT, DGFT, New Delhi
- Ms. Kavita Bhatia, Scientist 'F' Digital Economy & Digital Payment Division, Ministry of Electronics & Information Technology, New Delhi
- Mr. Pinal Patel, Sr. Research Manager, Raapid.ai.
- Mr. Sanjay Madan, Joint Director, Applied Artificial Intelligence & Analytics Division, C-DAC, Mohali
- Mr. Sundar Narayanan, Researcher & Domain Expert
- Dr. Vinosh Babu James, Technical Standards Director, Associate, Qualcomm India
- Mr. Avinash Agarwal, DDG (Convergence & Broadcasting), Telecom-munication Engineering Centre, New Delhi