TECHNICAL REPORT

ISO/IEC TR 24027

First edition
2021-11

Information technology — Artificial intelligence (AI) — Bias in AI systems and AI aided decision making

Technologie de l'information — Intelligence artificielle (IA) — Tendance dans les systèmes de l'IA et dans la prise de décision assistée par l'IA

Reference number
ISO/IEC TR 24027:2021(E)

Contents

Page

Foreword

ISO (the International Organization for Standardization) is a worldwide federation of national standards bodies (ISO member bodies). The work of preparing International Standards is normally carried out through ISO technical committees. Each member body interested in a subject for which a technical committee has been established has the right to be represented on that committee. International organizations, governmental and non-governmental, in liaison with ISO, also take part in the work. ISO collaborates closely with the International Electrotechnical Commission (IEC) on all matters of electrotechnical standardization.

The procedures used to develop this document and those intended for its further maintenance are described in the ISO/IEC Directives, Part 1. In particular, the different approval criteria needed for the different types of ISO documents should be noted. This document was drafted in accordance with the editorial rules of the ISO/IEC Directives, Part 2 (see www.iso.org/directives).

Attention is drawn to the possibility that some of the elements of this document may be the subject of patent rights. ISO shall not be held responsible for identifying any or all such patent rights. Details of any patent rights identified during the development of the document will be in the Introduction and/or on the ISO list of patent declarations received (see www.iso.org/patents).

Any trade name used in this document is information given for the convenience of users and does not constitute an endorsement.

For an explanation of the voluntary nature of standards, the meaning of ISO specific terms and expressions related to conformity assessment, as well as information about ISO's adherence to the World Trade Organization (WTO) principles in the Technical Barriers to Trade (TBT), see www.iso.org/iso/foreword.html.

This document was prepared by Technical Committee ISO/IEC JTC 1 *Information technology*, Subcommittee SC 42, *Artificial intelligence*.

Any feedback or questions on this document should be directed to the user's national standards body. A complete listing of these bodies can be found at www.iso.org/members.html.

Introduction

Bias in artificial intelligence (AI) systems can manifest in different ways. AI systems that learn patterns from data can potentially reflect existing societal bias against groups. While some bias is necessary to address the AI system objectives (i.e. desired bias), there can be bias that is not intended in the objectives and thus represent unwanted bias in the AI system.

Bias in AI systems can be introduced as a result of structural deficiencies in system design, arise from human cognitive bias held by stakeholders or be inherent in the datasets used to train models. That means that AI systems can perpetuate or augment existing bias or create new bias.

Developing AI systems with outcomes free of unwanted bias is a challenging goal. AI system function behaviour is complex and can be difficult to understand, but the treatment of unwanted bias is possible. Many activities in the development and deployment of AI systems present opportunities for identification and treatment of unwanted bias to enable stakeholders to benefit from AI systems according to their objectives.

Bias in AI systems is an active area of research. This document articulates current best practices to detect and treat bias in AI systems or in AI-aided decision-making, regardless of source. The document covers topics such as:

— an overview of bias (5.2) and fairness (5.3);

— potential sources of unwanted bias and terms to specify the nature of potential bias (Clause 6);

— assessing bias and fairness (Clause 7) through metrics;

— addressing unwanted bias through treatment strategies (Clause 8).

Information technology — Artificial intelligence (AI) — Bias in AI systems and AI aided decision making

1 Scope

This document addresses bias in relation to AI systems, especially with regards to AI-aided decision-making. Measurement techniques and methods for assessing bias are described, with the aim to address and treat bias-related vulnerabilities. All AI system lifecycle phases are in scope, including but not limited to data collection, training, continual learning, design, testing, evaluation and use.

2 Normative references

ISO/IEC 22989[1)], *Information technology — Artificial intelligence — Artificial intelligence concepts and terminology*

ISO/IEC 23053[2)], *Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)*

3 Terms and definitions

For the purposes of this document, the following terms and definitions given in ISO/IEC 22989 and ISO/IEC 23053 and the following apply.

ISO and IEC maintain terminological databases for use in standardization at the following addresses:

— ISO Online browsing platform: available at https://www.iso.org/obp

— IEC Electropedia: available at https://www.electropedia.org/

3.1 Artificial intelligence

3.1.1
maximum likelihood estimator
estimator assigning the value of the parameter where the likelihood function attains or approaches its highest value

Note 1 to entry: Maximum likelihood estimation is a well-established approach for obtaining parameter estimates where a distribution has been specified [for example, normal, gamma, Weibull and so forth]. These estimators have desirable statistical properties (for example, invariance under monotone transformation) and in many situations provide the estimation method of choice. In cases in which the maximum likelihood estimator is biased, a simple bias correction sometimes takes place.

[SOURCE: ISO 3534-1:2006, 1.35]

3.1.2
rule-based systems
knowledge-based system that draws inferences by applying a set of if-then rules to a set of facts following given procedures

[SOURCE: ISO/IEC 2382:2015, 2123875]

1) Under preparation. Stage at the time of publication: ISO/DIS 22989:2021.

2) Under preparation. Stage at the time of publication: ISO/DIS 23053:2021.

3.1.3
sample
<statistics> subset of a population made up of one or more sampling units

Note 1 to entry: The sampling units could be items, numerical values or even abstract entities depending on the population of interest.

Note 2 to entry: A sample from a normal, a gamma, an exponential, a Weibull, a lognormal or a type I extreme value population will often be referred to as a normal, a gamma, an exponential, a Weibull, a lognormal or a type I extreme value sample, respectively.

[SOURCE: ISO 16269-4:2010, 2.1, modified - added *<statistics>* domain]

3.1.4
knowledge
information about objects, events, concepts or rules, their relationships and properties, organized for goal-oriented systematic use

Note 1 to entry: Information can exist in numeric or symbolic form.

Note 2 to entry: Information is data that has been contextualized, so that it is interpretable. Data are created through abstraction or measurement from the world.

3.1.5
user
individual or group that interacts with a system or benefits from a system during its utilization

[SOURCE: ISO/IEC/IEEE 15288:2015, 4.1.52]

3.2 Bias

3.2.1
automation bias
propensity for humans to favour suggestions from automated decision-making systems and to ignore contradictory information made without automation, even if it is correct

3.2.2
bias
systematic difference in treatment of certain objects, people, or groups in comparison to others

Note 1 to entry: Treatment is any kind of action, including perception, observation, representation, prediction or decision

3.2.4
human cognitive bias
bias (3.2.2) that occurs when humans are processing and interpreting information

Note 1 to entry: human cognitive bias influences judgement and decision-making.

3.2.5
confirmation bias
type of human cognitive *bias* (3.2.4) that favours predictions of AI systems that confirm pre-existing beliefs or hypotheses

3.2.6
convenience sample
sample of data that is chosen because it is easy to obtain, rather than because it is representative

3.2.7
data bias
data properties that if unaddressed lead to AI systems that perform better or worse for different *groups* (3.2.8)

3.2.8
group
subset of objects in a domain that are linked because they have shared characteristics

3.2.10
statistical bias
type of consistent numerical offset in an estimate relative to the true underlying value, inherent to most estimates

[SOURCE: ISO 20501:2019, 3.3.9]

4 Abbreviations

AI artificial intelligence

ML machine learning

5 Overview of bias and fairness

5.1 General

In this document, the term bias is defined as a systematic difference in the treatment of certain objects, people, or groups in comparison to others, in its generic meaning beyond the context of AI or ML. In a social context, bias has a clear negative connotation as one of the main causes of discrimination and injustice. Nevertheless, it is the systematic differences in human perception, observation and the resultant representation of the environment and situations that make the operation of ML algorithms possible.

This document uses the term bias to characterize the input and the building blocks of AI systems in terms of their design, training and operation. AI systems of different types and purposes (such as for labelling, clustering, making predictions or decisions) rely on those biases for their operation.

To characterize the AI system outcome or, more precisely, its possible impact on society, this document uses the terms unfairness and fairness, instead. Fairness can be described as a treatment, a behaviour or an outcome that respects established facts, beliefs and norms and is not determined by favouritism or unjust discrimination.

While certain biases are essential for proper AI system operation, unwanted biases can be introduced into an AI system unintentionally and can lead to unfair system results.

5.2 Overview of bias

AI systems are enabling new experiences and capabilities for people around the globe. AI systems can be used for various tasks, such as recommending books and television shows, predicting the presence and severity of a medical condition, matching people to jobs and partners or identifying if a person is crossing the street. Such computerized assistive or decision-making systems have the potential to be fairer and the risk of being less fair than existing systems or humans that they will be augmenting or replacing.

AI systems often learn from real-world data; hence an ML model can learn or even amplify problematic pre-existing data bias. Such bias can potentially favour or disfavour certain groups of people, objects, concepts or outcomes. Even given seemingly unbiased data, the most rigorous cross-functional training and testing can still result in an ML model with unwanted bias. Furthermore, the removal or reduction of one kind of bias (e.g. societal bias) can involve the introduction or increase of another kind of bias (e.g. statistical bias)[3], see positive impact described in this clause. Bias can have negative, positive or neutral impact.

Before discussing aspects of bias in AI systems, it is necessary to describe the operation of AI systems and what unwanted bias means in this context. An AI system can be characterized as using knowledge to process input data to make predictions or take actions. The knowledge within an AI system is often built through a learning process from training data; it consists of statistical correlations observed in the training dataset. It is essential for both the production data and the training data to relate to the same area of interest.

The predictions made by AI systems can be highly varied, depending on the area of interest and the type of the AI system. However, for classification systems, it is useful to think of the AI predictions as processing the set of input data presented to it and predicting that the input belongs to a desired set or not. A simple example is that of making a prediction relating to a loan application as to whether the applicant represents an acceptable financial risk or not to the lending organization.

A desirable AI system would correctly predict whether the application represents an acceptable risk without contributing to systemic exclusion of certain groups. This can mean in some circumstances taking into account considerations of certain groups, such as ethnicity and gender. There can be an effect of bias on the resulting environment where the prediction can change the results of subsequent predictions. Examples of how to determine whether an algorithm has unwanted bias according to the metrics defined in Clause 7, are given in Annex A.

Uncovering bias can involve defining appropriate criteria and analysing trade-offs associated with these criteria. Given particular criteria, this document describes methodologies and mechanisms for uncovering and treating bias in AI systems.

Classification (a type of supervised learning) and clustering (a type of unsupervised learning) algorithms cannot function without bias. If all subgroups are to be treated equally, then these kinds of algorithms would have to label all outputs the same (resulting in only one class or cluster). However, investigation would be necessary to assess whether the impact of this bias is positive, neutral or negative according to the system goals and objectives.

Examples of positive, neutral and negative effects of bias are as follows:

— Positive effect: AI developers can introduce bias to ensure a fair result. For example, an AI system used for hiring a specific type of worker can introduce a bias towards one gender over another in the decision phase to compensate for societal bias inherited from the data, which reflects their historical underrepresentation in this profession.

— Neutral effect: The AI system for processing images for a self-driving car system can systematically misclassify "mailboxes" as "fire hydrants". However, this statistical bias will have neutral impact, as long as the system has an equally strong preference for avoiding each type of obstacle.

— Negative effect: Examples of negative impacts include AI hiring systems favouring candidates of one gender over another and voice-based digital assistants failing to recognize people with speech impairments. Each of these instances can have unintended consequences of limiting the opportunities of those affected. While such examples can be categorized as unethical, bias is a wider concept that applies even in scenarios with no adverse effect on stakeholders, for example, in the classification of galaxies by astrophysicists.

One challenge with determining the relevance of bias is that what constitutes negative effect can depend on the specific use case or application domain. For example, age-based profiling can be considered unacceptable in job application decisions. However, age can play a critical role in evaluation of medical procedures and treatment. Appropriate customization specific to the use case or application domain can be considered.

In ML systems, the outcome of any single operation is based upon correlations between features in the input domain and previously observed outputs. Any incorrect outputs (including for example, automated decisions, classifications and predicted continuous variables) are potentially due to poor generalization, the outputs used to train the ML model and the hyperparameters used to calibrate it. Statistical bias in the ML model can be introduced inadvertently or due to bias in the data collection and modelling process. In symbolic AI systems, human cognitive bias can lead to specifying explicit

knowledge inaccurately, for example specifying rules that apply to oneself, but not the target user, due to in-group bias.

Another concern about bias is the ease with which it can be propagated into a system, after which it can be challenging to recognize and mitigate. An example of this is where data reflects a bias that exists already in society and this bias becomes part of a new AI system that then propagates the original bias.

Organisations can consider the risk of unwanted bias in datasets and algorithms, including those that at first glance appear harmless and safe. In addition, once attempts at removing unwanted bias have been made, unintended categorisation and unsophisticated algorithms have the potential to perpetuate or amplify existing bias. As a consequence, unwanted bias mitigation is not a "set-and-forget" process.

For example, a resume review algorithm that favours candidates with years of continuous service would automatically disadvantage carers who are returning to the workforce after having taken time off work for caring responsibilities. A similar algorithm can also downgrade casual workers whose working history consists of many short contracts for a wide variety of employers: a characteristic that can be misinterpreted as negative. Careful re-evaluation of the newly achieved outcomes can follow any unwanted bias reduction and retraining of the algorithm.

The more automated the system and the less effective the human oversight, the likelihood of unintended negative consequences is heightened. This situation is compounded when multiple AI applications contribute to the automation of a given task. In such multi-application AI systems, greater demand for transparency and explainability regarding the outcomes it produces can be anticipated by the organisations deploying them.

5.3 Overview of fairness

Fairness is a concept that is distinct from, but related to bias. Fairness can be characterized by the effects of an AI system on individuals, groups of people, organizations and societies that the system influences. However, it is not possible to guarantee universal fairness. Fairness as a concept is complex, highly contextual and sometimes contested, varying across cultures, generations, geographies and political opinions. What is considered fair can be inconsistent across these contexts. This document thus does not define the term fairness because of its highly socially and ethically contextual nature.

Even within the context of AI, it is difficult to define fairness in a manner that will apply equally well to all AI systems in all contexts. An AI system can potentially affect individuals, groups of people, organizations and societies in many undesirable ways. Common categories of negative impacts that can be perceived as "unfair" include:

— Unfair allocation: occurs when an AI system unfairly extends or withholds opportunities or resources in ways that have negative effects on some parties as compared to others.

— Unfair quality of service: occurs when an AI system performs less well for some parties than for others, even if no opportunities or resources are extended or withheld.

— Stereotyping: occurs when an AI system reinforces existing societal stereotypes.

— Denigration: occurs when an AI system behaves in ways that are derogatory or demeaning.

— "Over" or "under" representation and erasure: occurs when an AI system over-represents or under-represents some parties as compared to others, or even fails to represent their existence.

Bias is just one of many elements that can influence fairness. It has been observed that biased inputs do not always result in unfair predictions and actions and unfair predictions and actions are not always caused by bias.

An example of a biased decision system that can nonetheless be considered fair is a university hiring policy that is biased in favour of people with relevant qualifications, in that it hires a far greater proportion of holders of relevant qualifications than the proportion of relevant qualification holders in the population. As long as the determination of relevant qualifications does not discriminate against particular demographics, such a system can be considered fair.

An example of an unbiased system that can be considered unfair, is a policy that indiscriminately rejected all candidates. Such a policy would indeed be unbiased, as not differentiating between any categories. But it would be perceived as unfair by people with relevant qualifications.

This document distinguishes between bias and fairness. Bias can be societal or statistical, can be reflected in or arise from different system components (see Clause 6) and can be introduced or propagated at different stages of the AI development and deployment life cycle (see Clause 8).

Achieving fairness in AI systems often means making trade-offs. In some cases, different stakeholders can have legitimately conflicting priorities that cannot be reconciled by an alternative system design. As an example, consider an AI system that decides the award of scholarships to some of the graduate programme applicants in a university. The diversity stakeholder in the admissions office wants the AI system to provide a fair distribution of such awards to applications from various geographic regions. On the other hand, a professor, who is another stakeholder, wants a particular deserving student interested in a particular research area to be awarded the scholarship. In such a case, there is a possibility that the AI system denies a deserving candidate from a particular region in order to meet the research objectives. Thus, meeting the fairness expectations of all stakeholders is not always possible. It is therefore important to be explicit and transparent about those priorities and any underlying assumptions, in order to correctly select the relevant metrics (see Clause 7).

6 Sources of unwanted bias in AI systems

6.1 General

This clause describes possible sources of unwanted bias in AI systems. This includes human cognitive bias, data bias and bias introduced by engineering decisions. Figure 1 shows the relationship between these high-level groups of biases. The human cognitive biases (6.2) can cause bias to be introduced through engineering decisions (6.4), or data bias (6.3).

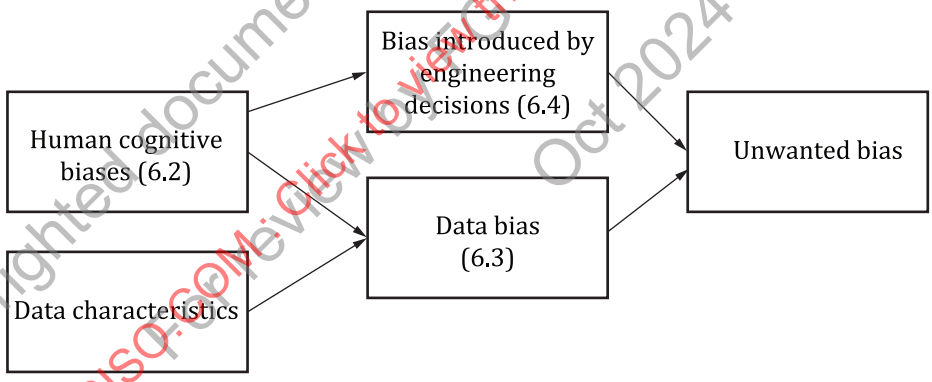

Figure 1 — Relationship between high-level groups of bias

For example, written or spoken language contains societal bias which can be amplified by word embedding models[4]. Because societal bias is reflected in existing language that is used as training data, it in turn causes non-representative sampling data bias (described in 6.3.4), which can lead to unwanted bias. This relationship is shown in Figure 2.

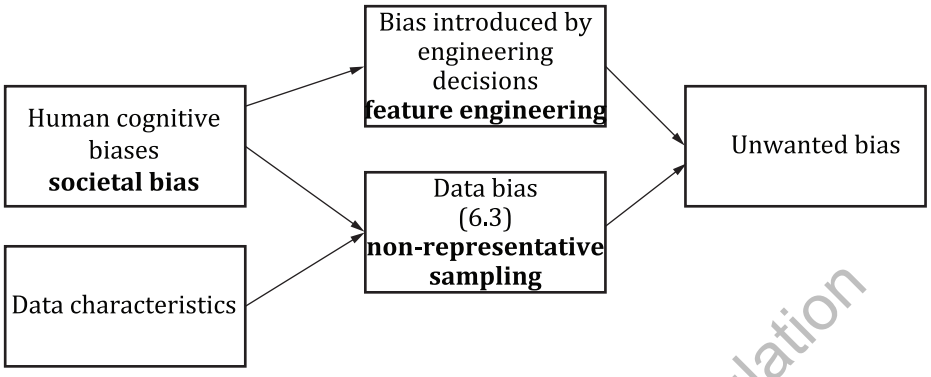

Figure 2 — Example of societal bias manifesting as unwanted bias

Systems are likely to exhibit multiple sources of bias simultaneously. Analysing a system to detect one source of bias is unlikely to uncover all. In the same example, multiple models are used for natural language processing. The outputs of the word embedding model that may be affected by non-representative sampling bias are then further processed by a secondary model. In this case, the secondary model is vulnerable to bias in feature engineering because a choice was made to use word embeddings as features of this model.

Not all sources of bias start with human cognitive biases; bias can be caused exclusively by data characteristics. For example, sensors that are attached to a system may fail and produce signals that can be considered outliers (see 6.3.10). This data, when used for training or reinforcement learning, can introduce unwanted bias. This is shown in Figure 3.

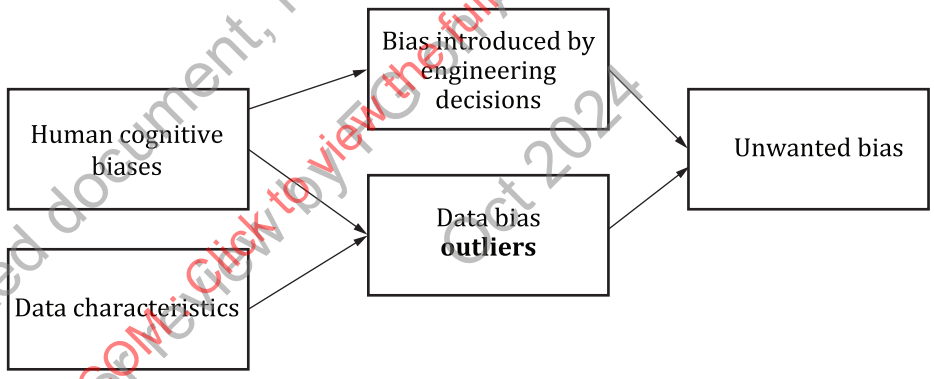

Figure 3 — Example of data characteristics manifesting as unwanted bias

6.2 Human cognitive biases

6.2.1 General

Human beings can be biased in different ways, both consciously and unconsciously, and are influenced by the data, information and experiences available to them for making decisions[5]. Thinking is often based on opaque processes that lead humans to make decisions without always knowing what leads to them. These human cognitive biases affect decisions about data collection and processing, system design, model training and other development decisions that individuals make, as well as decisions about how a system is used.

6.2.2 Automation bias

AI assists automation of analysis and decision-making in various systems, for example in self-driving cars and health-care systems, that can invite automation bias. Automation bias occurs when a human

decision-maker favours recommendations made by an automated decision-making system over information made without automation, even when the automation makes errors.

6.2.3 Group attribution bias

Group attribution bias occurs when a human assumes that what is true for an individual or object is also true for everyone, or all objects, in that group. For example, the effects of group attribution bias can be exacerbated if a convenience sample is used for data collection. In a non-representative sample, attributions can be made that do not reflect reality. This is also a type of statistical bias.

6.2.4 Implicit bias

Implicit bias occurs when a human makes an association or assumption based on their mental models and memories. For example, when building a classifier to identify wedding photos, an engineer can use the presence of a white dress in a photo as a feature. However, white dresses have been customary only during certain eras and in certain cultures.

6.2.5 Confirmation bias

Confirmation bias occurs when hypotheses, regardless of their veracity, are more likely to be confirmed by the intentional or unintentional interpretation of information.

For example, ML developers can inadvertently collect or label data in ways that influence an outcome supporting their existing beliefs. Confirmation bias is a form of implicit bias.

Experimenter's bias is a form of confirmation bias where an experimenter continues training models until a pre-existing hypothesis is confirmed.

Human cognitive bias, in particular this confirmation bias can cause various other biases, for example selection bias (6.3.2) or bias in data labels (6.3.3).

Another example is “What You See Is All There Is” (WYSIATI) bias. This occurs when a human looks for information that confirms their beliefs, overlooks contradicting information and draws conclusions based on what is familiar[6].

6.2.6 In-group bias

In-group bias occurs when showing partiality to one's own group or own characteristics. For example, if testers or raters consist of the system developer's friends, family or colleagues, then in-group bias can invalidate product testing or the dataset. This can be expressed in the evaluation of others, allocation of resources and many other ways.

It has been shown that people will seek to make more internal (dispositional) attributions for events that reflect positively on groups they belong to and more external (situational) attributions for events that reflect negatively on their groups.

6.2.7 Out-group homogeneity bias

Out-group homogeneity bias occurs when seeing out-group members as more alike than in-group members when comparing attitudes, values, personality traits and other characteristics. For example, Europeans can be seen as one homogenous group by Americans and vice versa. However, each group would be able to identify many subgroups and specific traits within each to prove the great diversity that exists in reality.

The out-group homogeneity effect is an individual’s perception of out-group members as more similar to one another than are in-group members, for example, "they are alike; we are diverse". The term "out-group homogeneity effect" or "relative out-group homogeneity” has been explicitly contrasted with “out-group homogeneity” in general, the latter referring to perceived out-group variability unrelated to perceptions of the in-group.

The out-group homogeneity effect is part of a broader field of research that examines perceived group variability. This area includes in-group homogeneity effects as well as out-group homogeneity effects. In-group homogeneity effects occur when in-group members are perceived as being similar with regards to positive characteristics. This area of research also deals with perceived group variability effects that are not linked to in-group or out-group membership, such as effects that are related to the power, status and size of groups.

The out-group homogeneity effect has been found using a wide variety of different social groups, from political and ethnic groups to age and gender groups.

6.2.8 Societal bias

Societal bias occurs when similar cognitive bias (conscious or unconscious) is being held by many individuals in society. Consequently, this bias can be encoded, replicated and perpetuated through organizations' policies.

It manifests in ML when models learn or amplify pre-existing, historical patterns of bias in datasets. This societal bias originates from society at large and can be closely related to other cognitive or statistical bias. It manifests as data available about society that reflects historical patterns. Societal bias can also be considered a type of data bias (6.3).

Societal bias also manifests when cultural assumptions about data are applied without regard to cross-cultural variation. For example, many groups treat genomics data as entirely secular, but some groups consider genomics to also contain sacred or spiritual properties. A model built on that data can predict disease across populations in a balanced way. However, if that data involves social groups who do consider the data to be sacred and the developer does not acknowledge or accommodate that cultural difference, the model can perpetuate societal bias regardless of the numerical output.

One example of societal bias is when historical data records are inappropriate for inferences being made, possibly reinforcing commonly held but inaccurate social views. For example, predicting whether a prisoner will commit another crime if released on parole (i.e. recidivism rate) depends on the availability of data about which previous prisoners committed which types of crime(s)[7], if any, after they also were released on parole. Available data, however, is restricted to former prisoners that were arrested for or convicted of a crime after they were released. It is well-documented that police arrests and judicial convictions are themselves heavily influenced by attitudes toward ethnicity, poverty and prior arrests. For example, any systematic over-arrest and conviction of a particular group of people would then lead to the systematic over-classification of recidivism among this population of prisoners.

Systemic bias, also called institutional bias, is a form of societal bias found in systems. Systemic bias is the inherent tendency of a socio-technical system or process to support particular outcomes.

The term systemic bias is historically used in the context of human systems and processes operating within organizations or within a society or culture and is discussed extensively in the field of industrial organization economics.

For example, systemic bias plays a part in systemic racism. Systemic racism is a form of racism that can be embedded within society, a particular culture, or an organization.

6.2.9 Rule-based system design

Developer experience and expert advice can have a significant influence on rule-based system design while also potentially introducing various forms of human cognitive bias. A developer can, for example, put in place an explicit rule based on an assumption about income that makes a split in a population such that separate models are applied for people who receive a regular income in their bank accounts versus those who do not. Such a split can be embedding a bias against those who are self-employed versus those employed by a third party. The rule can also unfairly discriminate against different demographics of peoples where there are links between type of employment and social demographics in a particular geographical location.

6.2.10 Requirements bias

Requirements creation presents occasions for the human cognitive biases listed in 6.2 to manifest. For example, implicit assumptions about hardware capabilities made by AI developers of high socio-economic status will not necessarily hold true for all the users of the AI system. In general, human cognitive bias will tend to draw the attention of AI developers towards conditions similar to their own that are not representative of the overall target user base. See Clause 8 for examples of treatment strategies for mitigating bias during requirements development.

The quantity being optimized during model training can also introduce requirements bias into the system. Naive translations of system requirements into utility equations can create requirements bias.

6.3 Data bias

6.3.1 General

A major source of bias is the data used to train and develop AI systems. The details in 6.3 elaborate on specific ways in which data can be biased. Data bias arises from technical design decisions and constraints and it can be caused by human cognitive bias, the training methodology chosen and variances in training infrastructure. These sources are not exclusive to AI systems and can be found in other applications. However, the way that they manifest in AI systems follows certain patterns. For example, bias caused by the training dataset can be based on an incorrect application or disregard of statistical methods and rules.

6.3.2 Statistical bias

6.3.2.1 Selection bias

6.3.2.1.1 General

Selection bias occurs when a dataset's samples are chosen in a way that is not reflective of their real-world distribution. Selection bias can be attributable to human cognitive bias in the data selection process (6.2).

6.3.2.1.2 Sampling bias

Sampling bias occurs when data records are not collected randomly from the intended population.

If a dataset is biased in the number of samples it draws from different groups, then the model will not accurately reflect the environment in which it will be deployed. For example, a facial recognition system trained on only one gender or only one race of people, is likely to not be able to as successfully recognize the faces of the types of people not in the training dataset[8].

6.3.2.1.3 Coverage bias

Coverage bias occurs when a population represented in a dataset does not match the population that the ML model is making predictions about. For example, if building an ML model to predict enjoyment of dramatic movies was based on a survey of viewers of comedic movies, it would clearly have coverage bias that can be material.

6.3.2.1.4 Non-response bias

Non-response bias (also called participation bias) occurs when people from certain groups opt-out of surveys at different rates than responders from other groups.

6.3.2.2 Confounding variables

A confounding variable is a variable that influences both the dependent variable and independent variable causing a spurious association. Because of this, a perceived relationship between two variables can be proven as partially or entirely false.

6.3.2.3 Non-normality

Most statistical methods assume that the dataset is subject to a normal distribution. However, if the dataset is subject to a different distribution (e.g. Chi-Square, Beta, Lorentz, Cauchy, Weibull or Pareto) the results can be biased and misleading.

6.3.3 Data labels and labelling process

The labelling process itself potentially introduces the cognitive or societal biases described in 6.2 to the data. For example, by deciding to classify people into male or female, or old and young, people are cast into discrete categories that do not necessarily represent the full reality being modelled. Labels can be selected that can be too broadly interpreted or that reduce a continuous spectrum to a binary variable. Other times the labelling process naturally falls into a discrete space, but the true labels are inaccessible. Proxies for ground truth are often used in these cases that correlate with the true labels and are accepted as sufficiently close for most purposes. If the inaccuracies introduced by that proxy are not random, they can introduce bias into the system. For example, an AI system recommending parole eligibility as described in 6.2.8 can also be described as generally not having access to information about whether people who were not released can have committed further offences.

Finally, it is possible for the labelling process itself to be inherently flawed. During data labelling it is possible for the human cognitive bias of the data labellers to be introduced into the data. It is also possible for such bias to be incorporated into the labelling instructions.

6.3.4 Non-representative sampling

Bias can manifest in several ways during training data selection, as result of the human cognitive biases described in 6.2, or due to sampling or coverage bias as described in 6.3.2. Sometimes all available datasets have properties inherited from the human cognitive bias that produced them. Human cognitive bias in the selection process can prevent the use or creation of unbiased datasets. Non-representative sampling is an example of biased training data selection. Most modelling techniques treat the training data as a true and accurate picture of the phenomenon being modelled. If a dataset is not representative of the intended deployment environment, then the model has the potential to learn bias based on the ways in which the dataset is non-representative.

Representativeness can take different forms in different application domains. For example, in the domain of facial recognition there are several different ways for a dataset to be non-representative with respect to attributes such as skin tone. The number of images of people with a particular skin tone, the lighting conditions of images and the relative entropy of images of people with one skin tone are examples of how a non-representative dataset can introduce bias into a model.

Data features can be present that can allow an ML model to infer group membership indirectly, even if the group membership features themselves are not among the ML model input (see 8.3.3.1).

6.3.5 Missing features and labels

Real world data are rarely complete. In particular, features are often missing from individual training samples. If the frequency of missing features is higher for one group than another then this presents another vector for bias. For example, the patient history for certain groups of people is often less complete in comparison to other groups due to the more fragmented care they receive on average. This imbalance in data quality has the potential to lead to lower quality medical predictions.

6.3.6 Data processing

Bias can also creep in due to pre-processing (or post-processing) of data, even though the original data would not have led to any bias. For example, imputing missing values, correcting errors, removing outliers or assuming specific data distribution models can also lead to bias in the operation of an AI system. This can be caused by the human cognitive biases described in 6.2.

6.3.7 Simpson's paradox

Simpson's paradox manifests when a trend that is indicated in individual groups of data reverses when the groups of data are combined. The background to this observation usually lies in the different weighting of the individual groups.

6.3.8 Data aggregation

Aggregating data covering different groups of objects that have different statistical distributions can introduce bias into the data used to train AI systems[9]. This can be caused by human cognitive bias such as out-group homogeneity bias.

6.3.9 Distributed training

Due to privacy and related regulatory considerations, learning closer to the source of the data can become widespread using distributed methodologies and techniques. Distributed ML can introduce its own cause for data bias, as the different sources of data can have a different distribution of features. If all the data sources that cumulatively contribute to the completeness of the feature space do not participate in the training, bias corresponding to feature space of non-participating data sources can occur. Non-participation can happen due to network issues, lower capability of computing devices for respective data sources or non-selection of the data source.

6.3.10 Other sources of data bias

The data and any labels can also be biased by artefacts or other disturbing influences. This bias would be regarded by an AI algorithm as part of the model to be generalized and would thus lead to undesired results. For example:

— Outliers are extreme data values that, if real, represent very low probability events of the to-be-modelled data.

— Noise is distortion and is characterized by a statistically-distributed variation of a physical quantity. Noise is caused by stochastic processes and cannot be described deterministically. Noise can have a negative influence on the model if overfitting takes place. Furthermore, artificially generated noise can be used to create adversarial examples that will cause undesired results.

6.4 Bias introduced by engineering decisions

6.4.1 General

ML model architectures - encompassing all model specifications, parameters and manually designed features - can be biased in several ways. Data bias and human cognitive bias can contribute to such bias.

6.4.2 Feature engineering

During the feature engineering process in building an ML model, the AI developers can directly use any of the input features or can create complex features for the ML model from input features in such a way that they can be linear or non-linear combinations of some of the input features. Steps such as encoding, data type conversion, dimensionality reduction and feature selection are subject to choices made by the AI developer and can introduce bias in the ML model.

For example, the AI developer can choose to represent height of people through categorial values such as tall, average or short and then choose the ranges in such a way that a majority of one gender falls in the average and short category while the majority of another falls in the tall and average category. This can introduce an unwanted bias in the model. As another example, the AI developer can use a complex feature of body mass index (BMI) composed from the height and weight of a person and then create the model to use the BMI feature rather than the original height and weight features. This can introduce bias that is unfair to some groups such as professional sumo wrestlers and weightlifters.

Sometimes, hidden or implicit correlations across features can gain prominence due to underfitting or when there are insufficient model parameters. This can then reflect as an unwanted bias in the system predictions.

6.4.3 Algorithm selection

The selection of ML algorithms built into the AI system can introduce unwanted bias in predictions made by the system. This is because the type of algorithm used introduces a variation in the performance of the ML model.

In the simplest example, this can involve using a linear model for a non-linear problem. In a more complex example, there are different possible configurations of long short-term memory models and such models can consist of several layers. This directly influences the complexity of the function that the network is able to approximate. Other neural network architectures like transformer-encoder-decoder models have functionality that can introduce unwanted bias in the predictions made by the system.

There can be many sub-models inside an ML model that can be interacting with a linear combination or a more complex combination of the sub-models. This can introduce many complex issues, which can include unwanted bias in the AI system predictions.

For example, in an ML model for a natural language question answering system there can be a combination of a predicate-prediction model, a value-identification model, a predicate-value binding model and a constraints-identification model. The way these sub-models are combined or sequenced can introduce unwanted bias in the system predictions.

Gradient-boosting can also be used to combine a set of machine learning sub-models into a single strong learner in an iterative fashion. However, the ensemble of such sub-models can introduce unwanted bias in the final predictions. For example, a ML model can use a sequential construction of shallow regression trees to form an ensemble and give a prediction as a sum of the trees' prediction probabilities. The way the ensemble is constructed can introduce bias in the system.

6.4.4 Hyperparameter tuning

When creating a machine learning model, the design choices made define the model architecture. Often the optimal model architecture evolves through hyperparameter tuning. Hyperparameters include the number of network layers, the number of neurons in a layer (also called the width of each layer), the learning rate for gradient descent, the degree of polynomials to use for the linear model and the number of trees in a random forest etc.

Hyperparameters define how the model is structured and cannot be directly trained from the data like model parameters. Thus, hyperparameters affect the model functioning and accuracy of the model and thus can potentially lead to bias.

There are many possible activation functions for a neural network. The choice of activation functions can affect the accuracy and predictions made by the ML model. This can appear as bias in the predictions made by the system.

Further, it is often necessary to pick a decision threshold for a given model to perform some action. Frequently, such thresholds are manually set. Thus, if a model updates on new data, the previously manually set threshold can become invalid or can lead to bias in predictions. This is especially important for dynamic systems.

6.4.5 Informativeness

For some groups the mapping between inputs present in the data and outputs are more difficult to learn. This can happen when some features are highly informative about one group, while a different set of features is highly informative about another group. If this is the case, then a model that only has one feature set available, can be biased against the group whose relationships are difficult to learn from available data. This concept applies both in training and evaluating a model. Model expressiveness (6.4.7.2) is also a factor for informativeness.

6.4.6 Model bias

Given that ML often uses functions like a maximum likelihood estimator to determine parameters, if there is data skew or under-representation present in the data, the maximum likelihood estimation tends to amplify any underlying bias in the distribution. For example, if the distribution of men and women represented in the dataset is 60 % men and 40 % women, a model can represent this skew at 80 % men and 20 % women by using thresholds that do not consider the initial bias. Downstream activation functions like the sigmoid function can amplify small differences in features that are the result of data bias.

6.4.7 Model interaction

6.4.7.1 General

It is possible for the structure of a model to create biased predictions. For example, assume that variables *X* and *Y* are relevant to predicting outcomes in two groups but are independent in one group and interactional in the other. A model where the two variables are present but cannot be isolated will potentially yield biased outcomes.

6.4.7.2 Model expressiveness

Models have different expressive capacity and some embody a wider variety of functions than others. The number and nature of parameters in a model as well as the neural network topology can affect the expressiveness of the model. Any feature that affects model expressiveness differently across groups has the potential to cause bias.

Model architectures that allow for recursion can also allow for more expressiveness. The properties of some groups can be wholly understood through a static representation of the current state. The same properties of other groups can be understood as the result of a sequence of states. In this case, a non-recurrent model will perform better for the former group than for the latter.

7 Assessment of bias and fairness in AI systems

7.1 General

When developing and deploying an AI system, it is important to be aware of possible bias (including statistical and societal) that can lead to unfair system behaviour. One way to uncover evidence of unwanted bias is to assess the system's outputs using one or more fairness metrics. Unwanted bias that is discovered using this assessment, can be treated using the techniques described in Clause 8.

Metrics of statistical bias seek to evaluate differences between average observed values and true values. With the proliferation of AI systems and concerns relating to their fairness, there is also growing awareness that such metrics of statistical bias are insufficient to detect unfair or discriminatory behaviour. This has led to the development of metrics[10], that aim to capture various notions of fairness.

Such metrics are described in the literature on "algorithmic fairness"[11] and are referred to as "fairness metrics" or "metrics of algorithmic fairness". For example, some fairness metrics are designed to compare different types of error rates between different groups of people.

Note that there is not a one-to-one correspondence between the broad notion of bias (as defined in this document) and statistical bias metrics. There is also no one-to-one correspondence between the broad notion of fairness (as discussed in 5.3) and fairness metrics. The main challenge remains to determine the metrics that are most appropriate in any given context[12].

To date, most work on fairness metrics has focused on the fairness of classification- or regression-based AI systems with respect to groups defined in terms of one or more demographic attributes. Approaches for assessing the bias and fairness of classification-based AI systems are introduced in this clause. Similar concepts exist for AI regression systems - see[13][14] for examples.

Bias in classification systems can be detected through measurements of different types of errors with respect to various groups. The approach of dividing data into training, validation and test datasets is augmented by subdividing each of those datasets based on the characteristics with respect to which the system is expected to be fair. If there are multiple characteristics relevant to detecting possible biases in a particular system, then those characteristics can be considered as independent or as intersectional. For example, a system that is unbiased with respect to gender and race independently can be biased towards a specific combination of the two.

Prior to testing, fairness objectives can be made explicit, this includes determination of relevant demographic characteristics, selection and justification of fairness metrics to be used in detecting bias and fixing the allowable margin of difference ("delta").

Once the data has been appropriately divided, the pre-determined fairness metrics are calculated on each group and comparisons are made between groups. A classification system can be considered sufficiently fair (or "unbiased") with respect to relevant characteristics, if metric-based measurements across groups are within a sufficiently small delta.

7.2 Confusion matrix

A confusion matrix[15] (see Figure 4) is a tool that can be used to evaluate the performance of a classifier. It reports the number of false positives, false negatives, true positives and true negatives and includes further performance criterion derived from these values. Since a confusion matrix contains and compares multiple metrics, it allows a detailed analysis of the performance of a classifier and is helpful in circumventing or uncovering the weaknesses of individual metrics.

		true condition			
	total population	condition positive	condition negative	Prevalence = $\frac{\sum condition\ positive}{\sum total\ population}$	Accuracy (ACC) = $\frac{\sum TP + \sum TN}{\sum total\ population}$
predicted condition	prediction positive	true positive (TP) Power	false positive (FP) Type I error	Positive Predictive Value (PPV), Precision, Relevance = $\frac{\sum TP}{\sum prediction\ positive}$	False Discovery Rate (FDR) = $\frac{\sum FP}{\sum prediction\ positive}$
	prediction negative	false negative (FN) Type II error	true negative (TN)	False Omission Rate (FOR) = $\frac{\sum FN}{\sum prediction\ negative}$	Negative Prediction Value (NPV), Separation Ability = $\frac{\sum TN}{\sum prediction\ negative}$
		True Positive Rate (TPR), Sensitivity, Recall, Probability of Detection = $\frac{\sum TP}{\sum condition\ positive}$	False Positive Rate (FPR), Fall-out, Probability False Alarm = $\frac{\sum FP}{\sum condition\ negative}$	Positive Likelihood Ratio (LR+) = $\frac{TPR}{FPR}$	Diagnostic Odds Rate (DOR) = $\frac{LR+}{LR-}$
		False Negative Rate (FNR), Miss Rate = $\frac{\sum FN}{\sum condition\ positive}$	True Negative Rate (TNR), Specificity, Selectivity = $\frac{\sum TN}{\sum condition\ negative}$	Negative Likelihood Ratio (LR-) = $\frac{FNR}{TNR}$	F1 Score = $\left(\frac{TPR^{-1} + PPV^{-1}}{2}\right)^{-1}$

Figure 4 — Confusion matrix and derived classification performance metrics[16]

7.3 Equalized odds

Equalized odds means that an algorithm's decisions are independent of a category *A* given the input *Y*.

A predictor $\hat{Y}$ satisfies equalized odds with respect to category A and outcome *Y*, if $\hat{Y}$ and *A* are independent conditional on *Y*:

$$P(\hat{Y}=\hat{y}|Y=y, A=m) = P(\hat{Y}=\hat{y}|Y=y, A=n)$$

for all values of *Y*, all values *m*, *n* of *A*.

This implies that true positive rates (TPR) are equal across demographic categories and false positive rates (FPR) are equal across demographic categories.

Note that this definition allows the models to take demographic information into account. TPR is equal to [1 minus false negative rate (FNR)], so this also encourages equal false negative rates across demographic categories. To see trade-offs between false negatives and false positives, comparing false negative rate and false positive rate can help.

7.4 Equality of opportunity

Equal opportunity means that an algorithm's decisions that $\hat{Y}$=1 are independent of a category *A* given the input *Y*=1.

A binary predictor $\hat{Y}$ satisfies equal opportunity with respect to *A* and *Y* if $\hat{Y}$=1 and *A* are independent conditional on *Y*=1. Formally:

$$P(\hat{Y}=1|Y=1, A=m) = P(\hat{Y}=1|Y=1, A=n)$$

for all values of *m*, *n* of *A*.

This implies equal True Positive Rates (TPR) across demographic categories.

7.5 Demographic parity

Statistical parity means that there are equal prediction rates between categories. Demographic parity (also known as group fairness) says there are equal prediction rates between demographic categories, like ethnicity. Demographic parity, which is a case of statistical parity, means that a decision - such as accepting or denying a loan application - be independent of a demographic attribute. Formally, given demographic variable A:

$$P(\hat{Y}=\hat{y}|A=m) = P(\hat{Y}=\hat{y}|A=n)$$

for all values m, n that A can take.

Parity does not capture cases where the output decision is correlated with one of the groups or attributes being evaluated[17] and there is no guarantee that the predictions made will be equally good for each category.

7.6 Predictive equality

Predictive equality implies equal false positive rates (FPR) across demographic categories.

Formally:

$$P(\hat{Y}=1|Y=0, A=m) = P(\hat{Y}=1|Y=0, A=n)$$

for all values m, n that A can take.

7.7 Other metrics

Alternative metrics can include minimax fairness and Pareto fairness[18].

8 Treatment of unwanted bias throughout an AI system life cycle

8.1 General

An AI system or service typically goes through a life cycle from the business need and inception stage, through design and development, verification and validation, to operations and retirement. The AI system life cycle is defined in the ISO/IEC 22989 standard being developed by SC 42[19]. There different ways a life cycle is instantiated for a specific service or product. This clause describes the life cycle stages important to this document only.

In many AI system implementations, parts of the system will be procured rather than developed by the same organization. Recognizing this, different parts of this clause will apply to different implementation contexts and there can be intellectual property, transparency or commercial considerations that hinder bias identification and reduction.

Supply chain risks related to unwanted bias can occur, in particular where there is no transparency of source code, models, training data provenance or labelling processes. It can be beneficial to have bias-related considerations added to any commercial agreement.

8.2 Inception

8.2.1 General

System requirements analysis is an important activity in mitigating unwanted bias. It is the stage where the internal and external requirements are analysed, stakeholders of the system are determined and

system goals are assessed. By this milestone, risks posed by the system have been identified, impact to identified stakeholders has been assessed and stakeholder levels of engagement defined.

The considerations and potential requirements described in this clause are not applicable to the treatment of unwanted bias alone. A formal analysis and a more complete list of considerations exist in a set of international standards in the area of governance and management. These are being developed by SC 42[19], and include:

— ISO/IEC 38507[20];

— ISO/IEC 42001, *Information technology — Artificial intelligence — Management* system (in preparation);

— ISO/IEC 23894, *Information technology — Artificial intelligence — Risk Management* (in preparation).

All activities for bias mitigation are implemented based on the policy set by the governing body and through management activity.

8.2.2 External requirements

The identification of external requirements as part of system analysis activity is a normal part of the systems development and procurement life cycle. Special consideration can be given during this process to the following regulatory frameworks:

— International human rights, equality and indigenous rights instruments that place obligations upon entities to ensure that certain freedoms, for example the provision of financial services, are provided without discrimination.

— Specific laws and guidance relating to the provision of technical solutions, for example regulating the accessibility of software for users with different abilities or regulating a specific sector (such as HIPAA[21] in the United States).

— Data protection and privacy legislation[22] can include provisions relating to automated decision-making. This can be supranational, national or regional legislation. At the time of writing, examples of data protection and privacy legislation include: California Consumer Privacy Act of 2018[23], the Japanese Act on the Protection of Personal Information 2016[24] and the EU and EAA's General Data Protection Regulation[25].

— Competition and business law.

Below are examples of possible types of obligations upon the accountable entity:

— The need for a risk assessment, that can include societal concerns from the perspective of affected stakeholders.

— Notification to users that they are subject to an automated decision, the requirement to gain explicit consent and to provide a non-automated alternative when consent is not given.

— Ensuring a certain level of auditability or explainability in the solution, in order to support analysis of a particular decision or event.

— Activities to quantify or mitigate risks, such as collecting meta-data about data sources to understand provenance and quality[26].

— Provision for the meaningful involvement of a human in the decision-making process.

— The equivalent provision and pricing of servicing for groups of people with certain characteristics. This can include the ability to demonstrate that equality is achieved in practice.

8.2.3 Internal requirements

In addition to regulatory requirements, many other factors can contribute to a stakeholder's desire to mitigate bias, such as:

— internal goals, strategies and policies of an organization;

— moral or cultural values;

— avoiding societal concerns or reputational damage.

The analysis process can give special regard to five specific areas: the inclusion of trans-disciplinary experts, the identification of stakeholders, the selection of data sources, external change and specification of acceptance criteria including acceptable levels of bias.

8.2.4 Trans-disciplinary experts

Whilst unwanted bias is a relatively new topic in the context of technology, it is a well-understood topic in the social sciences. As part of the requirements analysis process (and indeed the whole system life cycle), it is relevant to consider the expertise available to fully mitigate societal concerns about bias and account for various perspectives. This can include:

— social scientists and ethics specialists;

— data scientists and quality specialists;

— legal and data privacy experts;

— representatives of users or groups of external stakeholders.

For example, the designers of a facial recognition system can place importance on the face contour feature in their design and miss the fact that the contour can be (partially or completely) covered for people with particular cultural or religious backgrounds. A sufficiently diverse team is more likely to identify such limitations with designs, assumptions and datasets.

8.2.5 Identification of stakeholders

Traditional requirements analysis includes the identification of stakeholders. However, in order to comply with aspects of the aforementioned regulatory frameworks and to properly mitigate societal concerns, this traditional definition of stakeholder can be broadened to include those directly and indirectly affected by the implemented system.

Based on the types of data being used to make automated decisions, designers can further decompose lists of stakeholders into groups of people who are differently affected by bias in the system, have different abilities in the use of the system or have different levels of knowledge and access. It is important to consider which biases, negative experiences or discriminatory outcomes can occur.

These can be considered through a variety of participatory design or ethnographic methods, both of which involve active outreach to and discussion with, affected groups. Typically, it is insufficient to note who can theoretically be impacted without direct input from those groups. Often, it is also insufficient to have a member of the affected group (who possibly also works as part of the design team) assess how that group is affected. Groups are not monoliths and one person cannot always appropriately represent the range of perspectives possible.

Stakeholder identification and engagement can be included in a formal description and documentation of anticipated areas of concern and potential consequences for affected groups, positive and negative. From these, more qualified and quantitative requirements can be derived. A human impact assessment conducted in later phases can then revisit these areas of concern and assess whether the concern was successfully mitigated.

8.2.6 Selection and documentation of data sources

The selection of data used to form either explicit rules within an expert rule-based system or data that is used to train ML models, is an essential activity that has a significant bearing on bias.

In the case of explicit knowledge, it is important to consider the human cognitive bias already present in those specifying the knowledge. Human cognitive bias can be present in a human judgement, that then becomes codified in a rule-based system, then propagated for the full system lifetime at a larger scale. Whilst it can be considered acceptable that such cognitive bias is present in a single human decision, propagating that bias through automated decision-making can have a much larger impact.

Statistical AI systems that learn from data without explicit knowledge being specified suffer from many risks. To the extent possible, collection can obtain data that is fair for each group, especially with regards to outcomes. For example, to reduce bias for a binary classifier, collected data can aim for equal ratios of positive to negative classification training examples for every group of interest.

Consideration can be given to individual data sources to determine:

— Completeness. A data source that excludes certain records because they do not hold the same features for all records can provide an incomplete picture and result in defects in the training process. Publicly available data (for example, from the internet) is unlikely to have equivalent distribution across groups of people.

— Accuracy. A data source that contains inaccurate data will propagate those inaccuracies into an ML model. This can result in general accuracy issues, but these issues can also be skewed towards certain groups of people, for example, those with less credit history.

— Collection procedures. It is important to understand the lineage of data, how it is collected, how it is input and whether these processes affect completeness and accuracy. The location and environment that data records are obtained from can be considered.

— Timeliness. The frequency that the data records are collected or updated can be relevant to ensure their accuracy. Conversely, updating can require re-evaluating. A system that passes an audit can drift out of conformity, especially if the updates come from a different process than the initial data.

— Consistency. The consistency with which input data items (or labels) are determined can be important. For example, if a human is categorizing items that do not have a clear boundary between categories, different categories or labels can result from the same data.

8.2.7 External change

Attention can be given to how changes can occur in the use of the developed or procured system.

Examples include:

— Deployment of an existing system to a different environment, including different users, target markets and data sources, can change the risks relating to the system.

— Over time, the relationship between inputs and outputs of the system can change. For example, a system using an ML model to make decisions based on correlations establishing during the initial ML model training can suffer if those correlations change over time. This is known as data drift.

— Use cases for the system can develop, either deliberately or organically, requiring re-assessment of risks.

— Societal norms change over time (for example, attitudes towards gender behavioural norms, ideal body shape or smoking). Bias in AI systems can be re-evaluated to consider resulting changes (such as to metrics, risks, stakeholders or requirements) and those changes can be addressed accordingly.

8.2.8 Acceptance criteria

Effective requirements are testable, in that when they are evaluated it is possible to determine whether a system complies with them. Often, AI system performance is compared to that of humans. Still, it is beneficial to be able to specify performance in a statistical manner. For example, stakeholders can specify a limit for false positive or false negative decisions in addition to an overall accuracy metric.

Establishing acceptance in terms of specific system characteristics and the degree of their compliance upfront allows for effective evaluation and decision-making.

Failure criteria can be the lower bound of acceptance, in effect, setting clear limits for acceptable performance for a model. Without these criteria being set and monitored, an AI system can drift such that unwanted bias arises without being noticed or remediated. The manner in which a system fails can also be carefully considered as a part of the design process to prevent extreme cases of unwanted bias occurring.

8.3 Design and development

8.3.1 General

Models themselves can contain unwanted bias if care is not taken to prevent this. Human bias can be encoded into ML systems through implicit assumptions that make their way into the design. Thus, it helps to identify and make implicit assumptions explicit.

In addition to the content of this clause, open-source tools listed in Annex A can assist with the treatment of bias issues in the design and development process.

8.3.2 Data representation and labelling

8.3.2.1 General

A key step in the development of an ML system is in deciding how to best represent the training data in features that are interpretable by the model. This is also called feature engineering and 6.3 describes some types of data bias that can affect this process. There are several often-implicit criteria that go into this process, including the criteria by which data are judged to be “good” or “bad” (for example, whether an overexposed photo can be kept in a dataset). These criteria about which data records are included in training data and what features are selected, can be made explicit. It is important to consider how the data relates to the purpose of building the system, the process by which features are chosen and the individuals that are choosing features and their rationale (including any associated explicitly identified assumptions).It is important to evaluate the chosen features for any data and human cognitive biases, such as missing feature values, unexpected feature values or data skew. Any of these can indicate that certain groups or characteristics are not accurately represented in the data.

Missing feature values can be the result of implicit bias in the data collection process, which can be identified and mitigated.

In deep learning algorithms where features are created during training, correct labels are critical. Annotations done by humans to create labels for the data can be prone to bias due to human cognitive bias or errors arising from to difficulties in the labelling task itself. It is important to ensure the labels are correct by both evaluating the labellers and the final labelled data. Additionally, even with correct labels, the types of labels specified by the labellers can be the cause of unwanted or undetected bias in the final model.

8.3.2.2 Using crowd workers for labelling

Crowd workers often annotate and label data to be used for the training of supervised ML. Errors or human cognitive bias that manifest during that process are propagated into a trained model[27].

Where crowd workers are used for data labelling it can be useful to understand the diversity and goals of the people who annotate the data and how they are incentivised. For example, it can be considered what success looks like for different workers, what they are paid for (quality or quantity) and the trade-offs between time spent on task and enjoyment of the task, as well as their cultural and socio-demographic backgrounds

Designers can develop tasks that account for human differences (and cognitive bias) in annotation, for example, by using golden test questions with known answers or other forms of participant pre-selection.

Clarity of instructions, as well as obtaining feedback from crowd workers on potentially confusing tasks, can be important to reducing unwanted bias. Human variability, including accessibility, muscle memory and human cognitive bias in annotation can be accounted for by using a standard set of questions with known answers.

8.3.3 Training and tuning

8.3.3.1 Training data

In many cases, preparing or curating a balanced dataset can occupy most of the development time. A seemingly straightforward approach to bias mitigation is to remove the relevant features that can be responsible for the unwanted bias directly. For instance, in a use case that automatically short-lists candidates based on resume information, examples of such features can be ethnicity, gender and age. At the same time, features that are relevant to the use case include experience, skills, qualification, certification and professional membership. While removing ethnicity, gender and age can appear to address the issue, other features, acting as proxy variables, can indirectly reflect bias. For example, features such as salutation or prefix (Mr/Ms/Mrs), or occupation can represent a proxy variable for gender. Thus, removing only some of the obvious features that are associated with unwanted bias does not always result in bias mitigation. Other examples of proxy variables include music tastes and age, shopping patterns and gender, zip codes and race and income levels, family status and gender, education (which university or college a person graduated from) and race, weight or height and gender etc.

Data-based methods can be used to mitigate bias in the training data. Data re-weighting, for example, can up-weight samples that align with an objective. Such techniques include:

— Sampling techniques to measure the representativeness of samples from different sources so that selection bias is identified and mitigated;

— Stratification sampling to overcome a rareness phenomenon. Stratified sampling can be used by increasing the relative frequency for the positive cases as compared to the negative cases. This can be done using several techniques including synthetic minority oversampling techniques (SMOTE) [28];

— Careful features selection in cases where sample features have strong correlation with the bias to be excluded (e.g. gender or colour).

Another approach is to find out the amount of unwanted bias present in the data and offset the bias from the result. Using a series of steps, it is possible to find out the feature contribution and the relative significance of each feature in a model's prediction. It is possible to then offset all the influence by a feature causing the bias. The process of determining relative feature significance can include the following:

— Iterative Orthogonal Feature Projection (IOFP). Given the input and output to an ML model, the method seeks to produce an input ranking that corresponds to the Machine Learning system's dependence on each input in its decision-making process and thus can detect bias involving certain features[29].

— Minimum redundancy, maximum relevance (MRMR). Feature selection identifies subsets of data that are relevant to the parameters used. One scheme is to select features that most strongly correlate to the classification variable whilst also being mutually distant from each other. This

scheme, termed as Minimum Redundancy Maximum Relevance (MRMR) selection has been found to be more powerful than other modes of feature selection[30].

— Ridge or LASSO regression. LASSO and Ridge regression are linear regression methods with regularization to prevent overfitting to training data. These techniques are also used to assist with feature selection[31].

— Random forest. Random forest is an approach that combines several randomized decision trees and aggregates their predictions by averaging. This technique has shown excellent performance in settings where the number of variables is much larger than the number of observations[32].

Note that although these approaches have been used to perform feature selection or feature relevance, they are not always directly applicable to determining biases present within data.

8.3.3.2 Tuning

Bias mitigation algorithms have been created to attain various objectives. Bias mitigation algorithms (also sometimes referred to as fair algorithms) can be classified as follows:

— Data-based methods, such as the up-sampling of under-represented populations or the use of synthetic data.

— Model-based methods, such as the addition of regularization terms or constraints that enforce an objective during optimization, or representation learning to hide-out or reduce the effect of a specific variable.

— Post-hoc methods, such as identifying group-specific decision thresholds based on predicted outcomes to equalize false positive rates or other relevant metrics.

Examples of bias mitigation algorithms that are applied are:

— Disparate impact remover: A pre-processing technique that edits values that will be used as features in such a way to reduce different treatment between the groups.

— Individual bias detector and remover: A technique that creates a new ML model for individuals in the disfavoured group receiving a different decision as compared to similar individuals in the favoured group. Sometimes it can apply different thresholds for the positive classification across groups.

— Decoupled classifiers: A technique to train a separate classifier on each group. The separate classifiers can equivalently be thought of as a single classifier that branches on the group feature.

— Joint loss function: A technique to capture group parity by using a joint loss function that penalizes differences in classification statistics between groups.

— Transfer learning: A technique[33] to mitigate the problems of low data volume for groups where there is a smaller population of data.

8.3.4 Adversarial methods to mitigate bias

One method for mitigating bias is to incorporate an adversarial unit into the model's architecture[34]. In these methods an "adversary" is predicting some property or characteristic defining groups towards which fairness is desired. The output from the model for which bias is being mitigated is the input to the adversarial model. The weight update for that model is then modified so that in addition to being optimized for the task it is performing, it is also reducing the amount of information it makes available to the adversary useful for its prediction. The net effect of this system is that the system learns how to perform its task in ways that are orthogonal to the characteristics for which bias is unwanted.

8.3.5 Unwanted bias in rule-based systems

Diversity in the background and experiences of the designers along with leveraging transdisciplinary experts (see 8.2.4) can help in reducing chances of introducing unwanted bias into the design of a system. For instance, consider a system for automatic identification of potential smugglers with hard-coded rules based on the knowledge of a few very experienced subject-matter experts. If the subject-matter experts are mostly experienced in a particular smuggling area, use of the system in a different area can result in an unintentional profiling of specific classes of people. This in turn can lead to a systematic difference in how these classes will be treated relative to other classes in the new context. A common consequence in such a scenario would be imbalanced chances of flagging someone mistakenly between different cohorts.

8.4 Verification and validation

8.4.1 General

The verification and validation of a newly developed ML model can identify and mitigate potential unwanted bias prior to deployment. A hold-out dataset obtained from a data source independent of the training dataset is typically used in verification and validation. This safeguard for model generalizability is also important for safeguarding against any unwanted bias implicit in the training dataset. In general, any steps taken during dataset processing and model training would be beneficial to apply to the validation data and procedures where applicable.

Whilst verification of ML systems is undertaken intensively using training and testing datasets (see 8.3.3.2), it is limited to verification of the results based on selection and variation in the data available. An AI system can be evaluated in a specific context. Having separate teams working on training and evaluation, a common practice in software development, can also avoid the influence of individual cognitive bias.

The investigation of apparent defects in the model can reveal why it is not maximizing for overall accuracy. The resolution of these defects can then improve overall accuracy. Datasets under-representative of certain groups (see 6.3.4) can be augmented with additional training data to improve accuracy in the decision-making and reduce biased results.

Software testing traditionally relies upon a "body of knowledge used as the basis for the design of tests and test cases"[35]. The success of any empirical testing activity is typically limited by the degree to which the surrounding requirements or risk management processes have explicitly identified potential unwanted bias or sources of unwanted bias. Further information on risk management in relation to AI is outlined in Annex B.

The techniques outlined in this subclause are meant to be conducted at a statistically significant scale. The techniques usually measure the sensitivity of outcome to a sub-group not explicitly included within the input data (see 8.3.3.1).

Bias in AI systems is measured in comparable ways to how other properties such as aggregate performance is measured. However, aggregate performance metrics against the entire test set does not necessarily indicate whether unwanted bias is present in the model. Overall metrics in the confusion matrix (see 7.2) can appear to work well on the entire set. However, calculating precision and recall on subsets of demographically important or certain categories can often reveal bias such as lower accuracy for one identified gender over another, or lower accuracy for a specific demographic group. These differences in performance likely indicate that undetected bias is present at earlier stages of the development process. For example, a certain group can be under-represented (see 6.3.4) in the training data. This subclause is meant to apply to development of new systems, to deployment of existing systems and to evaluating whether systems maintain quality over time. A change in the relationship between the expected and actual input data can be cause for evaluation. Evaluation can also include the outcomes of deployment on system users and bystanders (such as people or objects who are incidentally present but are not the target or subject of an AI system deployment). For example, a system that is unfair with respect to gender and ethnicity independently can be fair towards a specific combination of the two.

8.4.2 Static analysis of training data and data preparation

Analysis of training and production data can reveal data bias, such as that described in 6.3. Clustering and visualization methods on the training data, derived features or resultant predictions can help detect imbalances or potential bias in the training data or system.

Evaluators can identify the profile of the training and production data and validate whether the spread of a certain variable represents the expected real dataset. An example of this is identifying that records of a certain age group have been used for training, when a different spread of ages is expected in real datasets. This activity can aim to validate the potential for selection bias, sampling bias and coverage bias, but cannot do so exhaustively, as it is limited by the knowledge of the evaluator.

Evaluators can identify stages in the data preparation process that can potentially introduce bias through "missing data". For example, if specific data items are not available consistently across an input dataset, engineers can impute that information for the remaining records or they can remove it. If the absence of that data item is correlated with specific groups of records, this can result in unwanted bias that would not normally be detected in model testing.

8.4.3 Sample checks of labels

The risk of incorrect labelling described in 6.3.3, that is, human labellers incorrectly specifying the labels for a set of input data then used to train the model, can be assessed through sample checks of submitted labels.

Labelling based on expert judgement can be more complicated to evaluate. Double-blind reviews can be conducted, or consideration can be given to the evaluation of multiple experts, in order to assess the quality of the initial label.

8.4.4 Internal validity testing

Internal validity testing evaluates the correlation between individual input data items and the system outputs. Internal validity testing then reviews whether these correlations are adverse in the context of specific requirements or acceptance criteria.

This process relies on the data items that cause unwanted bias to be included in the input data domain. It can detect bias in models and their interaction, such as model expressiveness described in 6.4.7.2, non-representative sampling as described in 6.3.4 or data processing issues as described in 6.3.6.

This can include evaluation in a fully integrated environment, in order to detect any unwanted bias in the data collection or preparation activities used during AI system development. Integration can also detect non-representative sampling. For example, the data collected in an integrated environment can have varying characteristics such as lighting levels or sensor update frequency. Those variations can influence the input data to the AI system.

8.4.5 External validity testing

External validity testing can involve re-evaluating prior observations using external data sources. This is a useful technique because it can detect many types of unwanted bias that have been described in this document, including indirect bias. The aspect of the input data the indirect bias relates to is not explicitly contained within the features but is a second-order derivation[36].

For example, some media reports of AI bias[37] have focused on research that correlated model outcomes with census or postal code data to results in order to illustrate disparity of outcomes.

External validity testing can also include integrating new input data and validating that the results are consistent with internal validity testing.

External validity testing is particularly important for indirect bias introduced by proxy variables. If a model designer attempts to mitigate bias by simply removing demographic information from the input, unwanted bias is likely to still exist through proxy variables. For example, the model can perpetuate

"colour-blind" racism[38][39], a sociological concept that describes how claims to not "see" a person's skin colour prevents understanding and addressing the persistence of racial inequality in society. To avoid this outcome, external validity testing can in fact include demographic data originally excluded, or an exploration of the effects of the proxy variable. Deeper inquiry can be conducted to understand why such proxies exist and whether the purpose can be fulfilled without it. External validity testing can include qualitative data that demonstrates disparate impacts of the same classification. For example, if a certain model is used to identify people before boarding an airplane, the emotional harm of a false negative can be greater for those groups stereotyped as likely "terrorists" than for other groups. In this context, integrating input datasets with additional data points can be useful to properly evaluate the system for unwanted bias.

8.4.6 User testing

Testing with different types of end-users can be helpful when a user's interaction with the system influences outcomes and predictions in a fashion correlated with the user's membership of a group.

Evaluating user experience in real-world scenarios across a broad spectrum of users, use cases and contexts of use is a useful technique to detect unwanted bias in model interactions (see 6.4.7), data processing issues (see 6.3.6) and issues with the data labels (see 6.3.3).

8.4.7 Exploratory testing

System developers can organize a pool of trusted, diverse testers who can act as adversaries to test the system and incorporate a variety of potentially harmful inputs into unit or functional tests. This can help in uncovering unanticipated ways that a system can be biased, especially if the pool of testers includes representatives of groups which can be impacted by the system and who can be its end-users.

8.5 Deployment

8.5.1 General

Once deployed, proper training and support for the AI system is important for the users to enable effective use of the product. This includes guidance for system developers on what constitutes appropriate and inappropriate deployment of an AI system. For example, an attention tracking system can be perceived as inappropriate if used in an educational system to monitor student behaviour, but that can be different if the same system is used as a research tool in a psychology experiment.

Deployed systems can also include guidance for end-users. For example, it is desirable that recruiters using a hiring recommendation system understand the capabilities and limitations of the system. Both AI developers and end-users can be made aware of known areas of unwanted bias. This can be achieved through a transparency tool (see 8.5.3), that contains information about the data the model was trained on, distributions for populations of its false positive and false negative errors and other associated information.

Data subjects, the people to whom the training data refers, are not necessarily users of the system. They do not need training, but they can be informed of any bias within the system that can impact them, in language appropriate for the context. Failures in either training or support can result in additional bias that can be difficult to detect earlier.

8.5.2 Continuous monitoring and validation

Models can lose performance over time. Performance degradation can be attributable to changes in the environment, such as societal trends, practices and norms, emerging new behaviours, changing input population composition and changes in requirements. Further, a system can be biased towards a historical position[17].

Ongoing performance, using the techniques in 8.4, can be monitored when the system is deployed. This includes checking the performance of the system, for example outlier results, using visual data

exploration and techniques for assessing bias and fairness, including automated means. See Clause 7 for more information about measurement. If indications of unwanted bias are present, the system can be retrained or re-engineered. Monitoring is a known process in many industries that leverage automated decision-making in their processes. For example, in banking, scorecard models are being developed and introduced along with their approved monitoring processes. Monitoring processes can not only be applied to accuracy and performance of models (or systems) but can also be used for identification and tracking of unwanted bias in systems or models.

8.5.3 Transparency tools

To clarify the intended use cases of ML models and minimize their usage in contexts for which they are not well suited, released models can be accompanied by documentation detailing their performance characteristics. Model transparency tools can provide a framework for transparent reporting of ML model provenance, usage and fairness-informed evaluation. Model documentation can include:

— qualitative information, such as ethical considerations, target users and use cases;

— quantitative information, consisting of model evaluation that is disaggregated (split across the different target subgroups) and intersectional (including evaluation on multiple subgroups in combination, e.g. ethnicity and gender). See Clause 7 for further information on metrics;

— data information, if possible, that can be formalized as a data transparency tool.

The usefulness and accuracy of a transparency tool relies on the integrity of the creator(s) of the tool itself and can be stored as documentation or meta-data associated with each model. Model cards[40] are one transparency tool among many, that can include, for example, algorithmic auditing by third parties (both quantitative and qualitative), "adversarial testing" by technical and non-technical analysts and more inclusive user feedback mechanisms (see 8.4.6).

When using (or providing) third-party models and applications (also known as AI as a service or machine learning as a service), "FactSheets"[41] can be used to increase transparency and trust around such offerings.

Annex A
(informative)

Examples of bias

A.1 Example 1

Consider an algorithm that is used in the loan application process, to make a prediction as to whether the applicant represents an acceptable risk or not.

A desirable AI system would correctly predict whether the application represents an acceptable risk without contributing to systemic exclusion of certain groups.

A hypothetical example of an incorrect prediction would be rejecting a loan application from an "acceptable risk" candidate. In such a scenario, an AI system is automating an existing process in which loan officers determine whether applications represent acceptable risks. Over the course of many years, this human-driven process can result in 25 % of applications from immigrants being rejected. Prior to testing their algorithm for bias, it was agreed, given the existing levels of societal bias and since false negatives (i.e. rejecting applicants who were, in fact, creditworthy) were considered more important than false positives, that demographic parity and equality of opportunity to within 2 % would be sufficient to accept a prediction model as "unbiased".

In a first version, the AI system, in operation and trained on all previous loan applications, rejected 20 % of immigrant applications, but only 10 % of non-immigrant applications. The AI system has learned from and is emulating human decision-making and, while the algorithm is performing better than the historical human-driven process, it fails the demographic parity test and is rejected as unsuitable. After removing sensitive factors from training data, the new model performed as shown in Table A.1 and Table A.2 in this annex. In these tables, a positive condition represents a creditworthy applicant and a negative condition represents a credit risk.

Table A.1 — Confusion matrix for immigrant applications in Example 1

		True condition		
	Total population	**Condition positive**	**Condition negative**	**Total prediction**
Predicted condition	**Prediction positive**	88	0	88
	Prediction negative	2	10	12
	Total condition	90	10	100
		TPR = 0,98	TNR = 1,00	ACC = 0,98
		FPR = 0,00	FNR = 0,02	

Table A.2 — Confusion matrix for non-immigrant applications in Example 1

		True condition		
	Total population	**Condition positive**	**Condition negative**	**Total prediction**
Predicted condition	**Prediction positive**	89	1	90
	Prediction negative	1	9	10
	Total condition	90	10	100
		TPR = 0,99	TNR = 0,90	ACC = 0,98
		FPR = 0,10	FNR = 0,01	

Since the rejection rate for immigrants lies within 2 % of the rejection rate for non-immigrants, Demographic Parity holds. And since the TPR for immigrants lies within 2 % of the TPR for non-immigrants, also Equality of Opportunity holds. Therefore, according to the bias criteria established in advance of testing, the new models can be considered unbiased and are ready to be deployed.

A.2 Example 2

In another hypothetical scenario, the business wants to apply AI in order to enter a new business area based on anonymized data it believes to be useful, but that does not provide full insights into creditworthiness. The business decides to use Demographic Parity as a metric to determine the degree of unfair bias in the trained model, setting a threshold of 1 % difference in order to accept the model. Tests on the AI system reveal that it rejects 30 % of applications – regardless of immigration status. Even though the AI system rejects far too many applications, this is not tied to any specific group; according to their choice of metric and threshold, the model is unbiased. The business decides that the high overall error rate is an acceptable risk to take given the marginal cost of entering this market space is low, it can be worth the price of automation. However, this decision can be quite controversial. For example, in the setting in Table A.3 and Table A.4, the model indeed satisfies Demographic Parity, but fails every other test of algorithmic fairness defined in Clause 7. It also performs appreciably worse in terms of accuracy for immigrant versus non-immigrant populations, leaving the organisation open to accusations of discrimination.

Table A.3 — Confusion matrix for immigrant applications in Example 2

		True condition		
	Total population	**Condition positive**	**Condition negative**	**Total prediction**
Predicted Condition	**Prediction positive**	65	5	70
	Prediction negative	15	15	30
	Total condition	80	20	100
		TPR = 0,81	TNR = 0,75	ACC = 0,80
		FPR = 0,25	FNR = 0,19	

Table A.4 — Confusion matrix for non-immigrant applications in Example 2

		True condition		
	Total population	**Condition positive**	**Condition negative**	**Total prediction**
Predicted condition	**Prediction positive**	65	5	70
	Prediction negative	5	25	30
	Total condition	70	30	100
		TPR = 0,93	TNR = 0,83	ACC = 0,90
		FPR = 0,17	FNR = 0,07	