2022 IEEE International Conference on Big Data (Big Data)

2022 IEEE International Conference on Big Data (Big Data) | 978-1-6654-8045-1/22/$31.00 ©2022 IEEE | DOI: 10.1109/BigData55660.2022.10020735

Standardization on Bias in Artificial Intelligence as Industry Support

Ewelina Szczekocka
Orange Innovation Poland
Orange Polska S.A.
Poland
0000-0003-4704-2112

Christèle Tarnec
Orange Innovation
Orange S.A
France
000-0001-9718-9396

Janusz Pieczerak
Orange Innovation Poland
Orange Polska S.A.
Poland
0000-0002-8218-8259

***Abstract*— Industry strives for trustworthy Artificial Intelligence (AI) systems through recognizing and implementing Responsible AI principles. Solutions supporting that goal are of the utmost interest in that context. Standardization is an essential element here, as it provides a platform for industry to discuss and facilitate not only the development of practical rules and requirements but also ways to implement AI based systems. One of Responsible AI principles is fairness, and bias is a serious obstacle against it. First, we explain the concept of Responsible AI and highlight results of our analysis on bias and fairness in ongoing international standardization works and AI Act (AIA). We identified a gap between the principles defined by high-level studies, including the AIA, and their practical implementations, and differences within standardization and research works. Second, we draw a standardization map for AI works. Finally, we state how international standardization bodies may fill this gap?**

***Keywords*— fairness, bias, Responsible AI, standardization**

I. Introduction

Bias is not a new concept. It has been known in the technical (algorithmic bias) and social fields (e.g. stereotypes) for a long time. Originally, the concept was not related specifically to Artificial Intelligence (AI) systems, but to such branches of science as statistics, mathematics, analytics in general. It acquires particular importance in the face of developing AI systems.

First of all, AI is of a great interest since few years, and it is its great comeback. This is because the data volumes available for training models and extracting valuable information, as well as computing capacities have skyrocketed over the last decade.

The first intensive works on AI date back to the 1950s (John McCarthy[1], Natural Language Processing, the first chatbot ELIZA[2]). However, the first significant names for AI are even earlier, the distinguished mathematician Allan Turing (and his famous test of AI called "The Imitation Game"[3]), and even before that Rene Descartes, the great French philosopher, and his visions and predictions about thinking and decision-making by machines.

AI developed significantly in the 1990s, particularly with the increase in computing power. One of the most interesting moments was the victory of a computer program called Deep Blue[4] over the world chess champion, Garry Kasparov (in 1997).

In 2011, IBM's AI-based computer system called Watson took a part in the teleseminar Jeopardy[5]. It answered questions using Natural Language Processing (NLP). It beat the previous champion of the game by a large margin. Watson has begun to be used to assist in the selection of lung cancer treatment therapies and the device's applicability among law corporations in the US is explored.

Another milestone in competition between human and AI was reached in 2016, when DeepMind's AlphaGo program defeated 18-time world champion Lee Sedol[6]. These and other examples show that recent years have been full of intensive development of technology as well as AI systems.

Speaking the language of numbers, the market research agency IDC investigated[7] that the worldwide market revenue for AI started to grow significantly in 2018, and this steady growth is forecasted to 2030. It projected that the global AI market will reach a size of over half a trillion U.S. dollars by 2024 and will triple by 2030, growing to over 1.5 trillion U.S. dollars. Although different studies suggest variations in how much the global market will grow in size, they agree on its growth trend. AI is anticipated to become one of the fastest growing industries in the near future. According to IDC, the following industries are expected to spend the most on AI in the coming years. The retail and banking sectors are expected to spend most, followed by discrete manufacturing[8], healthcare, and process automation. Other industries that will experience fast spending growth include public safety, emergency response, shopping advisors and

[1] https://hdsr.mitpress.mit.edu/pub/0aytgrau/release/3
[2] https://onlim.com/en/the-history-of-chatbots/
[3] https://www.techtarget.com/searchenterpriseai/definition/Turing-test
[4] https://www.kasparov.com/timeline-event/deep-blue/
[5] https://www.ibm.com/ibm/history/ibm100/us/en/icons/watson/
[6] https://www.deepmind.com/research/highlighted-research/alphago/the-challenge-match
[7] https://www.idc.com/getdoc.jsp?containerId=prEUR249536522
[8] https://www.techtarget.com/searcherp/definition/discrete-manufacturing)

978-1-6654-8045-1/22/$31.00 ©2022 IEEE

recommendations. Increasing interest and adoption of AI technology by Small and Medium Enterprises (SMEs) will also fuel this growth for many incoming years.

Speaking of governments, they are supposed to compete by growing their domestic AI industries. For instance, European Union, United Kingdom, Germany, France, and Canada have all committed intensifying growth of their domestic AI industries, after China published its plan to outdistance United States as the global leader in AI by 2030 [1]. As one can see, there is a rapidly growing interest and ever deepening need to use AI. A key aspect of using this technology is related to how AI is used and how to protect people affected by its biased results, from which certain actions and consequences may follow. In principle, ensuring the proper operation of AI systems concerns all stakeholders associated with these systems. For example, a bank that rejects a loan application due to a proposed decision of an AI-based banking system may also be de facto harmed. If that decision was based on improper considerations, the bank lost the opportunity to potentially profit from a loan it did not grant. Likewise, if the court makes a judgment based on the wrong decision proposals from the AI system, this can be a detriment not only to the wrongly judged person, but also to the court itself and society as a whole (incurring costs, etc.). One of the examples of harmful bias was seen during the gender bias lawsuit in university admissions against UC Berkeley[9]. After analyzing graduate school admissions data, it seemed like there was bias toward women, a smaller fraction of whom were being admitted to graduated programs compared to their male counterparts. However, when admissions data was separated and analyzed over the departments, women applicants had equality and, in some cases, even a small advantage over men. The paradox happened as women tended to apply to departments with lower admission rates for both genders.

A growing body of research emphasizes that the way to reach trustworthy system is to meet the principles of so-called Responsible AI.

Responsible AI is a concept that both, industry [10] and science [15], have undertaken to define. It is related to the general question, whether AI can be trustworthy and scalable?

Numerous researchers in academia, research centers, standardization and business organizations, and communities are working on principles for Responsible AI, however this list is still open.

For instance, some research proposes the following Responsible AI principles: ethics, transparency, regulation and control, socioeconomic impact, design, responsibility [2], while IBM considers explainability, fairness, robustness, transparency, privacy as its AI ethics principles [11] while Microsoft has formulated as its six principles[12] of Responsible AI: fairness, reliability and safety, privacy and security, inclusiveness, transparency, accountability. Although these proposed principles of Responsible AI differ to some extent, ethics and the principle of fairness (closely related to ethics) is mentioned in each case.

A general limitation to these principles is the lack of empirically proven and evaluated methods to effectively translate them into practice [3]. Thus, there is no guarantee that the principles of Responsible AI can be correctly implemented and effectively evaluated. Then, how can the correctness of their implementation and subsequent operation of an AI system be explicitly evaluated? **This is an identified by us research gap that should be filled.**

With the above in mind, we focus on addressing this gap regarding implementation of Responsible AI principles. Moreover, we do this by the example of one of the key principles of Responsible AI, that is the fairness (mentioned above), and bias which is related to fairness. We analyze how bias may affect AI systems, as well as what can play an important role in identifying and eliminating its negative impact (in order to do that it is necessary, among other things, to properly recognize possible source of bias).

In our conviction combined forces of science and industry can jointly fill this depicted gap. Besides industry and science, governments are playing a role too. The European Commission (EC) is addressing ethical principles for AI in Europe, developing the AI Act (AIA). In 2020, the United States, in close collaboration with the National Institute of Standards and Technology (NIST) set up the National AIA under the National AI Initiatives (NAII), (e.g. NIST Trustworthy AI principles[13]). China is also very active in the field of AI striving to be a global leader, planning to realize trustworthy systems by design (e.g. White Paper on Trustworthy AI, Ethical Norms for New Generation AI).

According to us, the contribution of international standards in this context could be important and could help stimulate the sustainable development of reliable AI systems. Based on the analysis of standardization works that we provide, we can constate that the gap between theoretical rules and their practical implementations could be significantly reduced.

II. BIAS AND FAIRNESS ANALYSIS

A. Bias definition and challenge

In general, bias known in science has many flavors. By examples, in different disciplines, a bias is a systematic error or simplification (e.g. mathematics, behavioral sciences). Commonly found in statistics or epidemiology, a bias is an approach or procedure that causes errors in the results of a study (overestimation or underestimation of the population parameter that can be measured). As for the above, bias comes from data and algorithms, methods, procedures, or research assumptions.

In psychology, there is for instance a cognitive bias which is understood as the tendency to make decisions or take actions in an unknowingly irrational way. It can harm not only decision making, but also a judgment, values, and social

[9] https://www.brookings.edu/blog/social-mobility-memos/2015/07/29/when-average-isnt-good-enough-simpsons-paradox-in-education-and-earnings/

[10] https://www.accenture.com/us-en/services/applied-intelligence/ai-ethics-governance in addition to definition by scholars [15]

[11] https://www.ibm.com/artificial-intelligence/ethics

[12] https://www.microsoft.com/en-us/ai/our-approach?activetab=pivot1%3aprimaryr5

[13] https://www.nist.gov/speech-testimony/trustworthy-ai-managing-risks-artificial-intelligence

interactions. It seems somehow close to ethical aspects of making decisions. **However, technically in each case bias manifests as systematic error and oversimplification which may discriminate selections. It can occur intentionally or unintentionally (unwanted bias).**

Bias is a complex approach, as it can involve choices that are unfair (ethical issue), but it can also come from an error in data selection and insufficient generalization as its result (rather technical issue but can anyway harm people). At first glance, it can be difficult to distinguish between the causes of bias, especially since the effect can be similar or the same regardless the cause (e.g. unfair decision-making in any case: if bias was discriminatory, whether it was due to faulty assumptions or statistical errors).

When it comes to AI, there is an important need to explicitly work out how bias should be defined and understood. Considering European approach, the purpose of the AIA, one of the first sets of legal frameworks for AI is to frame the legal use of AI in proportion to the risks that AI poses to fundamental human rights (the requirements, mandatory for high-risky AI, will concern data, documentation and traceability, provision of information and transparency, human oversight, robustness and accuracy). A legal framework on AI and a Coordinated Plan with European Member States will address the risks of AI systems and new rules on Machinery will ensure the safe integration of the AI system into the overall machinery.

AI systems are categorized by their level of risks (low-, medium- and high-level) and the highest level of protection is set for high-risk systems. For high-risk system, without defining what is a bias, the AIA proposal requires the absence of bias in dataset used to train and test AI models (art 9.5) and allows for personal data processing in certain circumstances if needed to detect, correct and monitor biases (art 10.5). However, after analyzing the AIA proposal it seems to us that the general and brief consideration given to bias is not sufficient and does not indicate its relationship with fairness. Furthermore, no guidance on practical applications is given in the AIA proposal, however it makes the reference to harmonized standards (see part F-Standardization Specifics for more details) that would provide clarity on how to implement fair AI-based services.

B. Definitions of Bias by Standardization

Standardization supports business environment in providing unique understanding of concepts and terms. In particular, after an analysis from our part it becomes clear to us that standardization could help to define precisely bias for AI applications.

Some definitions of bias are proposed by stakeholders in international standardization bodies such as ISO and IEEE. A definition given by ISO/IEC JTC1 SC42 AI (Subcommittee 42 Artificial Intelligence) [4] identifies bias as "systematic difference in treatment of certain object, people or groups in comparison to others". The ISO intentionally does not provide explicit definition of fairness, instead indicating that it is a complex concept, highly contextualized (socially, ethically). As a consequence, it is difficult to define unambiguously its impact on AI systems, as it can vary for different applications. This comes from discussions inside this standardization organization and its experts. At the same time, the ISO has identified a number of ways in which AI systems can have unfair impact on individuals, groups of people, organizations, and societies (unfair allocation, unfair quality of service, stereotyping, denigration, over- and under-representation).

Ongoing work in IEEE [5] refers to an unequal treatment of certain cases compared to others (it also mentions a statistically significant correlation between certain outcomes and specific input variables).

We identified that, despite the very close definition of bias in AI given by standardization stakeholders and scholars (fairness in machine learning provided by the latter as "absence of any prejudice or favoritism towards an individual or a group based on their intrinsic or acquired traits in the context of decision-making"), there is no consensus of how to precisely measure bias. This also demonstrates to us the need for a discussion and collaboration that should involve both scholars and stakeholders **in standardization**, allowing to draw on different experiences.

C. Fairness definition

The concept of bias is associated with the notion of fairness as it could have positive or negative consequences.

Researchers define fairness in machine learning as "absence of any prejudice or favoritism towards an individual or a group based on their intrinsic or acquired traits in the context of decision-making) [6]. Standardization, in particular ISO, does not provide a definition of fairness, however, refers to it through a bias notion.

A lack of fairness can be described as discrimination if it relates to rights protected by law (such as sex, race, religion or belief, etc. as stated in the EU charter of fundamental rights art 21.1[14]).

Although no common consensus among AI stakeholders yet exists on how to define fairness, and moreover how to measure it, there seems to be awareness that metrics that will allow to measure fairness are important and should be defined. ISO has made an attempt to define first fairness metrics, which are associated with a bias approach [4].

D. Sources of Bias by Research and ISO Standards

Numerous studies, conducted by scholars but also standardization organizations have attempted to count and categorize biases that exist in the field of machine learning and there are more than twenty potential sources of bias AI based services. These biases come from existing biases in society and from humans facing cognitive biases, which influence all human decisions whether at the time of data collection or during the construction of the model or even its use, and from technical or statistical biases such as the use of non-representative training data or the use of an existing model (transfer learning) without checking that it is adapted to the new context.

We have provided our first comparative analysis of the approach to bias from an outlook of research and standards, seeing it as one of the elements of Responsible AI and in consequence trustworthy systems. Our observations show that in addition to the global view, each approach pays attention to some specific aspects of bias. From our perspective, it may be an asset that allows for a broader understanding of this issue. These differences can be seen, for example, in the

[14] https://fra.europa.eu/en/eu-charter/article/21-non-discrimination

understanding of bias sources by each of the stakeholders (and bias sources as stressed before, are one of the important elements to understand impact of AI functionalities on an AI system).

We have summed up bias sources proposed by one of the ISO standards, and the other ones proposed by researchers in the tables below. We have chosen to present details on biases from ISO works as they are most advance at this time.

Sources of bias proposed by ISO were synthesized by us in the form of a table from the ISO standard [4], and we supplemented them with few additional examples (referenced in the table), (see TABLE I).

ISO has identified several types of bias, such as:

- human cognitive bias (we mark it as "HC" in the table) - human bias that might impact the design and application of a system,
- data bias (we mark it as "D" in the table) - data properties that if unaddressed lead to AI systems that perform better or worse for different groups
- machine learning model architecture bias (we mark it as "MLMA" in the table)
- other biases (we mark it as "O" in the table) – not categorized elsewhere.

TABLE I. BIAS TAXONOMY BY ISO

Bias taxonomy identified by ISO		
Bias source	***Description***	***Example***
automation (**HC**)	Human cognitive bias due to over-reliance on the recommendations of an AI system	Human-decision favoring automated recommendations (even erroneous)
confirmation (type of implicit) (**HC**)	Human cognitive bias favoring predictions of AI systems that confirm pre-existing beliefs or hypotheses	Hypotheses, regardless of their veracity, more likely to be confirmed by the interpretation of information (intentional/ unintentional).
experimenter (**HC**)	A form of confirmation bias where an experimenter continues training models until a pre-existing hypothesis is confirmed	Rosenthal and Fode experiment [a] (1963) – misperception of groups of rats ("bright", "dull")
group attribution (**HC**)	A human assumption that what is true for an individual or object is also true for everyone, or all objects, in that group	In a non-representative sample used for data collection (a convenience sample), provided attributions might not reflect reality
implicit (**HC**)	A human association or assumption based on their mental models and memories	Attitudes or stereotypes affecting human's understanding, actions, and decisions in an unconscious way, making them difficult to control [b]
in-group (**HC**)	Showing partiality to one's own group or own characteristics	In the evaluation of others, allocation of resources, and many other ways.
out-group homogeneity (**HC**)	Seeing out-group members as more alike than in-group members	While comparing attitudes, values, personality traits, and other characteristics
"what you see is all there is" (WYSIATI) (**HC**)	Looking for information that confirms human's beliefs, overlooks contradicting information, and drawing conclusions based on what is familiar	Making human judgements and impressions according to the information available [c]
societal (**HC**)	When one of more similar cognitive biases (conscious or unconscious) being held by many individuals in society; originating from society at large, could be closely related to other cognitive or statistical biases; can be considered a type of data bias	In Machine Learning: when models learn or amplify pre-existing, historical patterns of bias in datasets (as data reflecting historical patterns). Also, when cultural assumptions about data are applied without regard to cross cultural variation
Bias source	***Description***	***Example***
selection bias (**D**)	When a dataset's samples are chosen in a way that is not reflective of their real-world distribution	Attributable to human cognitive biases in the data selection process
sampling bias (**D**)	A type of selection bias that occurs when data is not collected randomly from the intended population	If a dataset is biased in the number of samples, then the model might not accurately reflect the environment in which it will be deployed
coverage bias (**D**)	A type of selection bias that occurs when a population represented in a dataset does not match the population that the ML model is making predictions about.	Coverage bias in European telephone surveys (mobile and landline phones imbalance) [d]
non-response bias (**D**)	A type of selection bias, also called participation bias, occurs when people from certain groups opt-out of surveys at different rates than users from other groups	Poorly constructed surveys [e]
data labels and labelling process (**D**)	The labelling process may introduce the cognitive or societal biases to the data. Selected labels can be broadly interpreted. They may reduce a continuous spectrum to a binary variable	By deciding to classify people into male/ female, or old/ young, people are cast into discrete categories that do not necessarily represent the full reality being modelled
non-representative sampling (**D**)	In training data selection process bias might manifest in several ways: as result of the human cognitive biases, or due to sampling or coverage bias. Sometimes all available datasets have properties inherited from the human cognitive biases that produced them.	Biased training data selection. Facial recognition: non-representative data by attributes, e.g. skin tone (more images of people with a particular skin tone, the lighting conditions, the relative entropy of images)
missing features and labels (**D**)	In real data features are often missing from individual training samples. If the frequency of missing features is higher for one group than another (imbalance in data	The patient history for certain groups less complete than in other ones. The more fragmented care may lead to lower quality medical prediction

Bias taxonomy identified by ISO		
Bias source	***Description***	***Example***
	quality) then it is specific bias.	
data processing **(D)**	Bias might creep in due to pre-processing (or post-processing) of data, even though the original data would not have led to any bias. This might be caused by the human cognitive biases	Inputting missing values, correcting errors, removing outliers, or assuming specific data distribution models may lead to bias in the operation of an AI system
Simpson's paradox **(D)**	When a trend that is indicated in individual groups of data reverses when the groups of data are combined	Usually caused by the different weighting of the individual groups
confounding variables **(D)**	A confounding variable is a variable that influences both the dependent variable and independent variable causing a spurious association	A perceived relationship between two variables might be proven as partially or entirely false
non-normality **(D)**	If data in statistical methods is subject to a non-normal distribution, the results might be biased and misleading	Examples: Chi-Square, Beta, Lorentz, Cauchy, Weibull or Pareto
other sources of data bias **(D)**	The data and any labels can 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	Outliers - extreme data values; noise - distortion characterized by a statistically distributed variation of a physical quantity with a possible negative influence on the model in case of overfitting.
data aggregation **(D)**	Aggregating data covering different groups of objects with different statistical distributions can introduce bias into the data used to train AI systems	Human cognitive biases (out-group homogeneity bias) may cause it
distributed training **(D)**	This specific data bias related to non-participation of some data sources of the feature space in the training may occur, as the different sources of data might not have the same distribution of feature space	It can happen due to network issues, lower capability of computing devices for some data sources, or non-selection of the data source
informativeness **(MLMA)**	The mapping between inputs and outputs are especially difficult to learn for some groups. It applies both in training and evaluating a model. Model expressiveness is also a factor for informativeness	Case: some features highly informative about one group, and different set of features highly informative about another group. If a model has only one feature set available, might be biased against the group whose relationships are difficult to learn from available data
model **(MLMA)**	If there is data skew or under-representation in the data, a function used to determine parameters may amplify bias in the distribution. The impact may vary depending on a function (maximum likelihood estimator - amplifying any underlying bias, downstream activation functions like the sigmoid function may amplify small differences in features	If the distribution of males and females represented in the data is 60/40, a model might represent this skew at 80/20 by thresholds that do not consider the initial bias
model interaction **(MLMA)**	The structure of a model may create biased predictions. Any feature that impacts model expressiveness differently across groups has the potential to cause bias	For two variables relevant to predicting outcomes in two groups, independent in one group and interactional in the other, a model where both are present but cannot be isolated will potentially produce biased outcomes
bias in rule-based system design **(O)**	Different experiences within a team (developer experience, expert advice) might have a significant influence on rule-based system design while also potentially introducing various forms of human cognitive bias	A designer might provide an explicit rule based on an assumption on income causing separate models applied for people receiving a regular income versus those who do not. The split might embed a bias against persons self-employed versus those employed by a third party. The rule might also unfairly discriminate against different demographics as there may be links between type of employment and social demographics
requirements bias **(O)**	Human cognitive biases may manifest because of requirements creation. They will tend to draw the attention of AI system designers towards conditions similar to their own. However, that might not be representative of the overall target user base. The quantity being optimized during model training might also introduce requirements bias into the system. Naive translations of system requirements into utility equations might create bias	Implicit hardware capabilities assumptions made by developers of high socio-economic status are not necessarily true for all of the users of the product
statistical bias **(O)**	Type of consistent numerical offset in an estimate relative to the true underlying value, inherent to most estimates	It can include sampling, response, non-response, self-selection, measurement bias examples (e.g. answers in a survey to perform better – response bias, wrong or no calibration – measurement bias) [f]

[a] https://practicalpie.com/experimenter-bias/

[b] https://www.simplypsychology.org/implicit-bias.html

[c] https://facilethings.com/blog/en/what-you-see-is-all-there-is

[d] https://surveyinsights.org/?tag=coverage-bias

[e] https://www.statisticshowto.com/non-response-bias/

[f] https://helpfulprofessor.com/statistical-bias-examples/

We have summed up sources of bias collected by scholars [6] in the form of a table as well, as shown below (TABLE II).

TABLE II. BIAS TAXONOMY BY SCHOLARS

Bias taxonomy identified by scholars		
Bias source	***Description***	***Example***
measurement bias	It happens from the way we choose, utilize, and measure a particular feature	The recidivism risk prediction tool COMPAS (US), using prior arrests and friend/family arrests as proxy variables to measure level of "riskiness"/ "crime" (mismeasured proxies)
omitted variable bias	It occurs when one or more important variables are left out of the model	A model for predicting the annual percentage rate at which customers stop subscribing to a service, provides no warning. Reason: appearance of a new strong competitor in the market that the model didn't consider (omitted variable)
representation bias	It comes from the way definition and sampling from a population is done	Lacking geographical diversity in datasets like ImageNet as a bias towards Western countries
aggregation bias	When false conclusions are drawn for a subgroup based on observing other different subgroups or generally when false assumptions about a population affect the model's outcome and definition.	Clinical aid tools: diabetes patients may have apparent differences across ethnicities and genders, like HbA1c levels. A single model could be not best suited for all groups in a population.
Simpson's Paradox	It can bias the analysis of heterogeneous data composed of subgroups or individuals with different behaviors. Observation in underlying subgroups may be different from association or characteristic observed when these subgroups are aggregated.	The gender bias lawsuit in university admissions against UC Berkeley: a smaller fraction of women than men admitted to graduated programs (bias toward women), but there was the paradox: women applied to departments with lower admission rates for both genders
modifiable areal unit problem (MAUP)	A statistical bias in geospatial analysis, arising when modeling data at different levels of spatial aggregation (different trends learned from data aggregated at different spatial scales)	MAUP effect in the evaluation of others, allocation of resources, and many other ways.
sampling bias	Sampling bias arises due to non-random sampling of subgroups	Trends estimated for one population may not generalize to other.
longitudinal data fallacy	Observational studies often treat cross-sectional data as longitudinal, which may create biases due to Simpson's paradox	Bulk Reddit data: comment length decrease observed over time on average, in a cross-sectional snapshot of the population (different cohorts joining Reddit over years), but in disaggregation of cohorts, it was found to increase over time within each
linking bias	Network attributes from user connections, activities, or interactions may differ and misrepresent the true user behavior. It can be a result of many factors, e.g. network sampling, which can change its measures	Social networks can be biased toward low-degree nodes when only considering the links in the network and not considering the content and behavior of its users
algorithm to user	Algorithms modulate user behavior. Any biases in algorithms might introduce biases in user behavior	Some biases that as a result of algorithmic outcomes affect user behavior
algorithmic bias	Algorithmic bias is when the bias is not present in the input data and is added purely by the algorithm	predicting health care costs rather than illness (discrimination of pure patients)
user interaction bias	Observed on the Web and in other two sources: the user interface and through the user itself by imposing his/her self-selected biased behavior and interaction. It can be influenced by presentation and ranking biases.	Discovering patterns of language through interactions (e.g. Microsoft's chatbot called Tay)
ranking bias	Assuming top-ranked results are the most relevant and important will result in attraction of more clicks than by others. It affects search engines and crowdsourcing applications	Higher position in search results (e.g. Google search) is chosen more often.
popularity bias	Items that are more popular tend to be exposed more. Popularity metrics are subject to manipulation	Manipulation can be seen in recommendation systems, search engines, where popular objects would be presented more to the public, due to biased factors
emergent bias	A result of use and interaction with real users; a result of change in population, cultural values, or societal knowledge	User interfaces, as they tend to reflect the capacities, characteristics, and habits of prospective users by design
evaluation bias	It happens during model evaluation, including the use of inappropriate and disproportionate benchmarks for evaluation of applications	Adience and IJB-A benchmarks: composed of 79.6% and 86.2% light-skinned faces, used in the evaluation of facial recognition systems, biased toward skin color and gender
user to data	Many data sources used for training ML models are user generated. Any inherent biases in users might be reflected in the data they generate	When user behavior is affected/ modulated by an algorithm, any biases present in that algorithm might introduce bias in the data generation process
historical bias	It is the already existing bias and socio-technical issues in the world and can seep into from the data generation process even given a perfect sampling and feature selection	A 2018 image search for women CEOs ultimately resulted in fewer female CEO images due to the fact that only 5% of Fortune 500 CEOs were woman, which would cause the search results to be biased towards male CEOs
population bias	Statistics, demographics, representatives, and user characteristics different in the user population represented in the dataset or platform from the original target population	Originates from different user demographics on different social platforms, as women more likely using Pinterest, Facebook, Instagram, while men being more active in online forums like Reddit or Twitter
self-selection bias	A subtype of the selection or sampling bias in which subjects of the research select themselves	Ex. in a survey on successful students, some less successful ones taking the survey, which would bias the outcome. If more successful students would not fill out surveys that would even increase the risk of self-selection bias
social bias	It happens when other people's actions or content coming from them affect our judgment	By rating or reviewing an item with a low score, influenced by other high ratings, we may change our scoring thinking that it was too harsh
behavioral bias	It arises from different user behavior across platforms, contexts, or different datasets	Difference in emoji among platforms can result in different reactions and behavior from others and sometimes even leading to communication errors
temporal bias	It arises from differences in populations and behaviors over time	It arises from differences in populations and behaviors over time
content production bias	It arises from structural, lexical, semantic, and	Differences in use of language across different gender and age groups; also,

Bias taxonomy identified by scholars		
Bias source	***Description***	***Example***
	syntactic differences in the contents generated by users	across and within countries and populations
presentation bias	It is a result of how information is presented	The Web users can only click on content that they see, so it gets clicks, while everything else gets no click. It could be the case that the user does not see all the information
cause-effect bias	A result of the fallacy that correlation implies causation. It can have serious consequences due to its nature and the roles it can play in sensitive decision-making policies	By analyzing a success level of a new loyalty program. Analyst sees that customers who signed up for the loyalty program spend more money than the other
observer bias	It happens when researchers subconsciously project their expectations onto the research	Researchers (unintentionally) influence participants (during interviews/surveys) or when they cherry pick participants/ statistics that will favor their research
funding bias	Biased results reported in order to support or satisfy the funding agency or financial supporter of the research study	Employees of a company report biased results in their data and statistics to keep the funding agencies or other parties satisfied

Based on the comparative analysis that we conducted (see TABLE I and TABLE II) on how sources of bias are classified by ISO and scholars, we can conclude that differences exist among these classifications. Each of the two (academics, standardization) capitalizes on its knowledge about bias, probably by adopting a scientific and industrial approach respectively. **In particular, standardizations bodies bring a more practical perspective than academics** (as they gather industry experts all over the world, together with national standardization organizations' representatives, academics, researchers from industrial innovation and research centers).

For instance, ISO provides such proposals related directly to bias or broader Responsible AI approach, to support industry and organizations, and these are just few examples.

As an example, in Table I, the data bias source due to distributed training highlights real fairness issues when one wants to use Federated Learning, a promising technique in terms of privacy. Some unique examples of proposals in Table II in this scope are: user interaction, observer bias or population bias (the two latter important while providing different experiments) We note that ISO focuses specifically on technical bias, while scholars on human aspects of bias.

E. Bias governance to involve all stakeholders

Societal and technical biases are present all along the AI life cycle, as described in some research [16] and standardization works [17]. Biases existing in the world are included in data, then data are transformed during a process, possibly biased. The data is then used by a biased learning mechanism to produce a model used by biased people that may use an inadequate fairness metric. Moreover, these biases are interconnected due to the existence of the feedback loop phenomenon (i.e. a situation in which the trained machine learning model makes decisions that produce outcomes affecting future data that will be collected for future training rounds) through which stereotypes, discrimination and inequality reinforce each other, thus contributing to perpetuating situations of inequality.

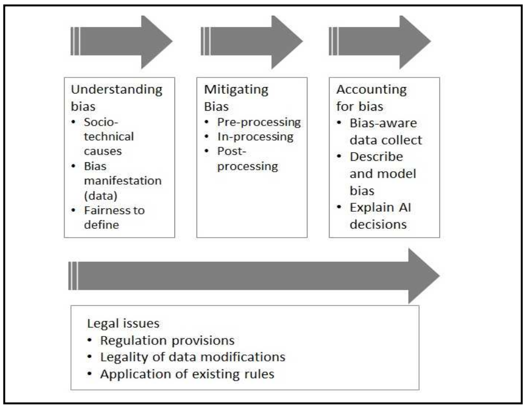

Fig. 1. Bias Treatment Process

It seems to us that all stakeholders should be involved and have a role to play in bias mitigation, thanks to **governance process and technical toolkits.**

A proposal of elements that can be a part of bias treatment process is provided in [7]. In fact, [7] proposes **four major elements** of the process: **understanding bias** (e.g. socio-technical causes of bias, like data generation, bias manifestation in data, like sensitive features and causal inferences), **mitigating bias** (pre-processing, in-processing, post-processing), **accounting for bias** (e.g. bias aware data collection, like bias elicitation), **legal issues** (e.g. regulation provisions, like data accuracy – GDPR). This is depicted in the figure (Fig. 1).

In our opinion a bias governance process should include all possible sources of bias at each step of the AI life cycle and all stakeholders have to be involved.

Women and AI Pledge from Cercle InterElles (https://www.interelles.com) or Arborus charter (https://charteia.arborus.org/en/) are good examples of the holistic approach to manage bias toward inclusive AI (i.e., AI that is focused on recognizing and eliminating biases, and involving as much diversity as possible, etc.).

F. Standardization Specifics

The EC intends to have standards that strengthen European competitiveness and benefits for society from AI. AIA proposal under elaboration by the EC sets out a comprehensive framework for AI governance and standards [8]. In particular, a conformance with so called "harmonized standards" is required to create a presumption of conformity for high-risk AI applications and services. More precisely, the AIA proposal gives **an important role to standards developed in Europe and** the AIA proposal aims to provide a strong incentive for industry to comply with **European standards** [9]. Creating harmonized standards in the area of AI is a great challenge, especially taking into account having harmonized standards produced and available by the time the AIA is enforced, that is at the end of 2024 or beginning of 2025. For speeding up the process, the EC issued in May 2022 a draft standardization request in support of safe and trustworthy AI.

Joined Committee of European Standardization, **CEN-CENELEC was designated by EC** to have a particular role and responsibility in the process of AI standards creation for Europe, and recognition of all stakeholders that should be invited for a cooperation on a development of these standards. For that, it has established a Standardization Request Ad Hoc

Group (SRAHG) on AI, for recognizing and cataloguing potential international standards that may help in speeding up the process. SRAHG requested experts within CEN/CENELEC to be nominated for its works and it had its first meeting on 10th June 2022, beginning to collect comments on the AIA and create a list of standards that are being developed under the AI.

An importance of **achieving a technical interoperability within the Trustworthy AI standards** is underpinned by [10] which considers that different stakeholders may present varying views of the relative importance of different proposed normative statements within the standardization works. Reference [10] stresses that the standardization of terms and a conceptual framework for Trustworthy AI should therefore enable clear, unambiguous communication between different stakeholders, so that these different viewpoints can be understood, made explicit and shared, enabling elaboration of relevant resolutions.

In general, international standards are aimed to **provide a reliable, consensus-based and voluntary basis for industries** (no obligation to use) in many aspects (incl. common understanding of technology, achieving interoperability). It also ensures that technical requirements are met enabling that the solutions work in different ecosystems and regions of the world. In general, international standards are being developed within standardization organizations, with active participation of all potential stakeholders like industry (both, SMEs and large enterprises), individuals, often government, and academics. A further conformance of technical solutions with particular standards is a guarantee of their quality.

International standardization is often US based and Chinese are also well represented there. At the same time, in Europe there is a tendency to produce European standards, reflecting European specifics, that may differ from international ones.

Standards are important not only from certification perspective but also to **give a guideline for companies** trying to develop fair AI models [11]. The objective for standards is to provide a reliable common basis for industries, through the same understanding of technological terms and expectations regarding products, services and processes.

Hereby, standards support business in facilitating trade and provide a framework for achieving economies, efficiencies, and interoperability.

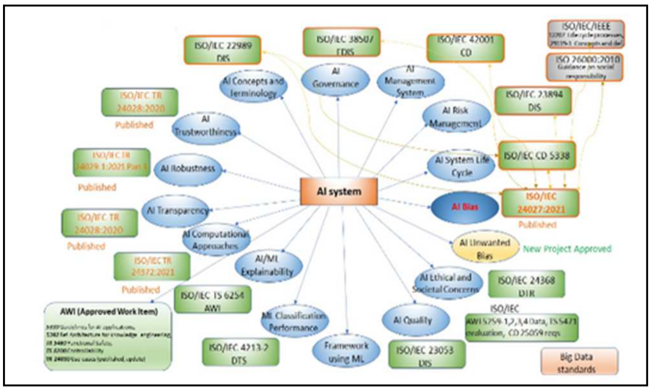

Fig. 2. ISO: AI Bias Standards Map

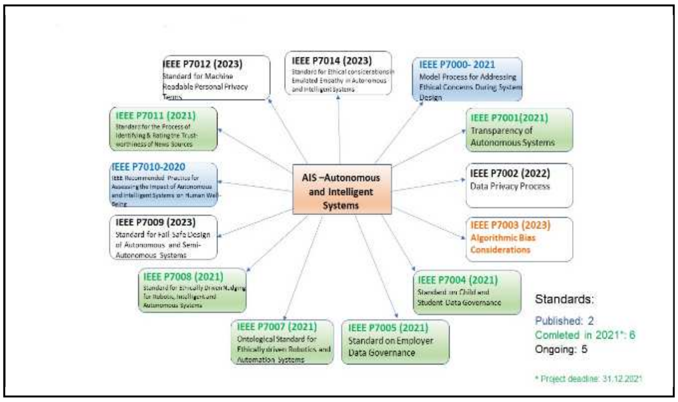
IEEE P7012 (2023)
Standard for Machine Readable Personal Privacy Terms
IEEE P7014 (2023)
Standard for Ethical considerations in Emulated Empathy in Autonomous and Intelligent Systems
IEEE P7000- 2021
Model Process for Addressing Ethical Concerns During System Design
IEEE P7011 (2021)
Standard for the Process of Identifying & Rating the Trust-worthiness of News Sources
IEEE P7001(2021)
Transparency of Autonomous Systems
IEEE P7010-2020
IEEE Recommended Practice for Assessing the Impact of Autonomous and Intelligent Systems on Human Well-Being
AIS –Autonomous and Intelligent Systems
IEEE P7002 (2022)
Data Privacy Process
IEEE P7009 (2023)
Standard for Fail-Safe Design of Autonomous and Semi-Autonomous Systems
IEEE P7003 (2023)
Algorithmic Bias Considerations
IEEE P7008 (2021)
Standard for Ethically Driven Nudging for Robotic, Intelligent and Autonomous Systems
IEEE P7004 (2021)
Standard on Child and Student Data Governance
IEEE P7007 (2021)
Ontological Standard for Ethically driven Robotics and Automation Systems
IEEE P7005 (2021)
Standard on Employer Data Governance
Standards:
Published: 2
Comleted in 2021*: 6
Ongoing: 5
* Project deadline: 31.12.2021

Fig. 3. IEEE: AI Bias Standards Map

Ensuring that a technical solution conforms to standards helps guarantee its quality and compliance with the requirements of its business ecosystem. A consensus approach to standards, achieved through various expert groups representing countries, regions, and industries is essential to enable the use of "standardized" solutions on a global scale.

G. Standardization Works Related to Bias

Bias management is a subject of work of different international standardization organizations: ISO/IEC, IEEE, NIST. ISO/IEC JTC1 is very active in the field, as three recent major standardization works on ethical aspects, bias in AI systems and its relations with fairness have been launched, articulated with existing relevant standards as it is a factor of success of fair AI in production (ISO/IEC TR 240027 and 24368 - in a form of technical reports - TR, ISO/IEC New Project TS 12791 is a technical specification – TS, close to normative document). Bias is here defined and addressed, especially where AI aids humans in decision making.

Landscape of standards concerned with bias considerations for ISO, synthesized in form of diagram during our analysis, is shown in Fig. 2.

Few examples of ISO works that may have importance in scope of Responsible AI are listed in the table below (see Table III).

TABLE III. ISO EXAMPLES FOR RESPONSIBLE AI

ISO documents for Responsible AI - Examples		
Document	***Description***	***Published***
TR 24027:2021	Bias in AI systems and AI aided decision making	2021
AWI 12792	Treatment of unwanted bias in classification and regression machine learning tasks (new project registered	Ongoing work
ISO/IEC WD TS 6254	Explainability of ML models and AI systems	Ongoing work
ISO/IEC AWI 5339	Guidelines for AI applications	Ongoing work
ISO/IEC 22989:2022	Artificial intelligence – Concepts and terminology	2022
ISO/IEC 23053:2022	Framework for Artificial Intelligence (AI) Systems Using Machine Learning (ML)	2022
ISO/IEC 42001	Artificial intelligence - Management System	Ongoing work
ISO/IEC 8183	AI - Data life cycle framework	Ongoing work

IEEE, as another international standardization organization, also provides work on bias in AI systems. Reference [12] underpins a role of P7003 activities in recognition of the increasingly pervasive role of algorithmic decision-making systems in corporate and government services, and growing public concerns regarding the ‘black box' nature of AI. It presents a whole ecosystem of ethics standards (P7000 series) aiming to translate the principles in the Ethically Aligned Design document (first edition of IEEE comprehensive report concerned with ethical implementation of autonomous and intelligent systems, A/IS) into actionable guidelines or frameworks to be used as practical industry standards. This is a set of eleven IEEE P70xx standards (some of them already accomplished). The IEEE P7003 objective is to provide a framework to help developers of algorithmic systems and people responsible for their deployment to identify and mitigate unintended, unjustified and inappropriate biases in the outcomes of the algorithmic systems. The standard will describe specific methodologies that will allow users to assert their way of work on addressing and eliminating issues of above-mentioned bias in the creation of their algorithmic system. This will help to design systems auditable by external parties (such as regulatory bodies).It is important to emphasize that these standards are part of broader ecosystems within their standardization organizations which enables to mutually use achievement of different normalizations and also build on the top of that.

Landscape of standards concerned with bias considerations for IEEE is shown in Fig. 3.

Few important elements of the standards achievements are depicted before (especially for ISO, as this particular standard is already published).

ETSI's work is also worth mentioning here. White Paper produced by ETSI [13], exploring key issues of AI and presenting both, huge potential benefits and risks, and new challenges for information as well. It considers bias from the perspective of interoperability and interchangeability as intrinsic benefits of a standardization process. In this context AI interchangeability in complex systems could be a pragmatic way to (partially) manage the risk of bias. It was also stressed that the ETSI community should identify and analyze the best practices recommended by other standardization organizations to tackle these problems in a complementary and collaborative manner.

H. The Importance of Certification and Conformance

We should emphasize that, from our industrial point of view, the ability to certify AI systems (as it is concerned with the high-risk ones) is fundamental. This means that the industry is keen to see a coherent practical approach to implementing and monitoring trustworthy AI systems developed as soon as possible. This is particularly true for the certification of AI systems as bias-free (or more accurately, unwanted bias-free). Creation of AI systems free of unwanted bias, may be a primary goal for industry as implementing systems fully free of bias may not be feasible.

At this point, it should be mentioned that. this approach is closely related to quality as it is concerned with satisfying needs and requirements (e.g. ISO 9000 series). It can result in certification of AI systems and assigning possibly a certification label.

As such, standards are required for checking a compliance of a product or a service (showing, that a product, a service or a system meets the requirements of a standard).

Certification is the provision by an independent body of written assurance (a certificate) that the product, service or system in question meets specific requirements.

A certification label is a label or symbol indicating that compliance with standards has been verified.

While the certificate is a form of communication between seller and buyer, the label is a form of communication with the end customer.

An interesting achievement towards AI certification is **IEEE CertifAIEd**, at this point a high-level proposal of a suite of criteria for evaluation, conformity assessment, and certification of properties of products and services based on Autonomous Decision Making and Algorithmic Learning Systems (ADM/ALS). They involve decision-making related to accountability, transparency, freedom from unacceptable algorithmic bias/fairness, privacy, and responsible governance [14].

Some promising works concerned with conformity assessment are just beginning in European standards, i.e., in the activities of the CEN-CENELEC JTC21 subcommittee related to AI. However, it is still an initial stage.

III. DISCUSSION

Providing standards for AI is a highly challenging task, as it concerns activities on the borderline between the technical and the non-technical approaches (and stakeholders representing two kinds of objectives which may contradict). Recent particular concern for different organizations is also that the technology is changing faster than the regulator is able to take it into account. This results in the situation that the development of rules and legislation to control AI systems is a step behind increasing implementation of AI in various types of systems and environments [10]. In this situation, standards could play a significant role in recognizing different data-related biases, identifying ethical issues and providing the necessary framework to address all the identified issues. They could serve as guidance to assist regulators in their work and also increase a possibility of adoption and use of new systems.

IV. CONCLUSIONS

In this paper, we survey different aspects of bias in Responsible AI. **First**, we sum-up the existing considerations of a top-level approaches of Responsible AI and consider the practical implementations of this concept. **Then**, we depict some examples of challenges for the bias issue (dual nature of bias – technical and ethical, bias sources). We emphasize importance of proper identification of bias sources in AI systems risk impact assessment. We also note existence of various provided works on this subject, namely by industry and science. **Finally**, we identify a research gap that exists between theorical and practical approaches of Responsible AI and propose a remedy for it.

Faced with the gaps identified in the research, i.e., the gap between the theoretical foundations and the practical implementations of the notion of Responsible AI, it is necessary to identify and involve all the potential stakeholders. Beyond researchers and regulators, all economic stakeholders must be involved. The challenge is to develop solutions for practical implementation of Responsible AI systems from design and development to operationalization. Being aware of expectations and requirements from industry, seeing EU perspective expressed in AIA proposal, we have looked at existing standardization activities.

In our opinion, a close cooperation between European and international standardization, acting as a consensus-based body of international experts is needed. In particular, this close cooperation is fundamental when it comes to bias in AI systems, which is a complex and multidimensional issue (ethical, legal and technical).

Beyond the needed cooperation within standardization organizations, it seems particularly important that various communities work together on methods of identifying and preventing bias (from a regulatory and legal as well as technical point of view, incidentally, with international standardization organizations as a part). Currently, this communication and collaboration looks insufficient to provide coherent solutions to assess the degree of trustworthiness of AI systems, and how to build trustworthy AI systems that will work in practice (taking into account as example the bias issue). This is, based on our observations, a strong rationale for close cooperation between the regulators and standardization organizations (that represent a technical approach and associate experts like analysts, developers, data scientists).

A. Authors and Affiliations

First A. Ewelina Szczekocka current employment Orange Polska S.A., Orange Innovation Poland, Warsaw, Poland, the M.S. degree in Applied Mathematics at the Warsaw University of Technology. Ms. Szczekocka is a research engineer focused on new technologies in data science, with a specific interest in Responsible AI for Telco, international expert in ISO (SC38, SC42 – Artificial Intelligence), IEEE member (P7003), ENISA (WG on AI Security). Contact her at ewelina.szczekocka@orange.com.

Second B. Christèle Tarnec, is with Orange S.A., Orange Innovation, France. Ms. Tarnec is a research engineer focused on Ethical AI with a specific interest on fairness issues for Telco. Contact her at christele.tarnec@orange.com

Third C. Janusz Pieczerak, is with Orange Polska S.A., Orange Innovation Poland, Warsaw, Poland. Mr. Pieczerak is a research expert focused on new networking technologies including virtualization, international expert in ETSI, ITU. Contact him at janusz.pieczerak@orange.com.

Acknowledgment

We would like to acknowledge and give our warmest thanks to our program manager (Emilie Sirvent-Hien) who made this work possible. Her guidance and advice carried us through all the stages of our project.

References

[1] J. Popper et al., “Artificial intelligence across industries - IEC Whitepaper,” IEC, 2018. [Online]. Available: https://www.researchgate.net/publication/329191549_Artificial_intelligence_across_industries_-_IEC_Whitepaper (IEC Whitepaper)

[2] “Responsible AI - Key Themes, Concerns and Recommendations for European Research and Innovation, Summary of Consultation with Multidisciplinary Experts,” “A Collaborative Platform to Unlock the Value of Next Generation Internet Experimentation” (HUB4NGI) project under EC grant agreement 732569, 2018. [Online].Available: https://futurium.ec.europa.eu/en/european-ai-alliance/open-library/responsible-ai-key-themes-concerns-recommendations-european-research-and-innovation?language=it/responsible_ai_consultation_-_public_recommendations_v1.0.pdf (report).

[3] C. Sanderson, Q. Lu, D. Douglas, X. Xu, L. Zhu, J. Whittle, “Towards Operationalising Responsible AI:,” extended and revised ver. to: “Software Engineering for Responsible AI: An Empirical Study and Operationalised Patterns,” Proc. International Conference on Software Engineering: The Software Engineering in Practice, ICSE-SEIP, 2022. (conference proceedings).

[4] “ISO/IEC TR 24027:2021 - Information Technology - Artificial Intelligence (AI) - Bias in AI systems and AI-aided decision making,” published by ISO, 2021. (published standard document).

[5] “IEEE P7003 - Draft Standard for Algorithmic Bias Considerations,” IEEE. (unpublished standard document).

[6] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, A. Galstyan, “A Survey on Bias and Fairness in Machine Learning,” ACM Computing Surveys, Article No.: 115, pp 1–35, 2022. (book).

[7] E. Ntoutsi, P. Fafalios, U. Gadiraju et al., “Bias in data - driven artificial intelligence systems - An introductory survey,” WIREs Data Mining and Knowledge Discovery published by Wiley Periodicals, Inc., 2020. (journal).

[8] The EU AI Act, [Online]. Available: https://artificialintelligenceact.eu/the-act/ (publicly available draft).

[9] M. McFadden, K. Jones, E. Taylor and G. Osborn, “Harmonising Artificial Intelligence: The role of standards in the EU AI Regulation,” Oxford Information Labs, 2021. [Online].Available: https://oxil.uk/publications/2021-12-02-oxford-internet-institute-oxil-harmonising-ai/ (report)

[10] D. Lewis, L. Hogan, D. Filip and P. J. Wall, “Global Challenges in the Standardization of Ethics for Trustworthy AI,” Journal of ICT, Vol. 8_2, 123–150. River Publishers, 2020. (journal).

[11] W. Ziegler, “A Landscape Analysis of Standardisation in the Field of Artificial Intelligence,” Journal of ICT, Vol. 8_2, 151–184. River Publishers, 2020. (journal).

[12] A. Koene, L. Dowthwaite and S. Seth, “IEEE P7003TM Standard for Algorithmic Bias Considerations, work in progress paper,” Proc. ACM/IEEE International Workshop on Software Fairness, FairWare'18, 2018. (conference proceedings)

[13] L. Frost,et al., Artificial Intelligence and future directions for ETSI, ETSI White Paper No. #34, 2020. (white paper)

[14] “IEEE CertifAIEd™ – Ontological Specification for Ethical Algorithmic Bias,” IEEE Standards Association. [Online].Available: https://engagestandards.ieee.org/ieeecertifaied.html (standard specification).

[15] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, et al., "Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI", Inf. Fusion 58, C (Jun 2020), 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

[16] Suresh H, Guttag J., Understanding Potential Sources of Harm throughout the Machine Learning Life Cycle. MIT Case Studies in Social and Ethical Responsibilities of Computing [Internet], 2021. Available from: https://mit-serc.pubpub.org/pub/potential-sources-of-harm-throughout-the-machine-learning-life-cycle

[17] Schwartz, R. , Vassilev, A. , Greene, K. , Perine, L. , Burt, A. and Hall, P. , Towards a Standard for Identifying and Managing Bias in Artificial Intelligence, Special Publication (NIST SP), National Institute of Standards and Technology, Gaithersburg, MD, 2022. [online]. Available: https://doi.org/10.6028/NIST.SP.1270,

https://tsapps.nist.gov/publication/get_pdf.cfm?pub_id=934464