IEEE P7003™ Standard for Algorithmic Bias Considerations

Work in progress paper

Ansgar Koene
Chair of IEEE P7003 working group
Horizon Digital Economy Research institute, University of Nottingham
NG7 2TU
United Kingdom
ansgar.koene@nottingham.ac.uk

Liz Dowthwaite
IEEE P7003 working group secretary
Horizon Digital Economy Research institute, University of Nottingham
NG7 2TU
United Kingdom
liz.dowthwaite@nottingham.ac.uk

Suchana Seth
IEEE P7003 working group member
Berkman Klein Center for Internet & Society, Harvard University
MA 02138
USA
suchana.work@gmail.com

ABSTRACT*

The IEEE P7003 Standard for Algorithmic Bias Considerations is one of eleven IEEE ethics related standards currently under development as part of the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. The purpose of the IEEE P7003 standard is to provide individuals or organizations creating algorithmic systems with development framework to avoid unintended, unjustified and inappropriately differential outcomes for users. In this paper, we present the scope and structure of the IEEE P7003 draft standard, and the methodology of the development process.

CCS CONCEPTS

• **General and reference** → **Document types** → Computing standards, RFCs and guidelines

KEYWORDS

Algorithmic Bias, Standards, work-in-progress, methods

ACM Reference format:

A. Koene, L. Dowthwaite, and S. Seth. 2018. IEEE P7003 Standard for Algorithmic Bias Consideratoins. FairWare'18, May 29, 2018, Gothenburg, Sweden. 4 pages. https://doi.org/10.1145/3194770.3194773

1 INTRODUCTION

In recognition of the increasingly pervasive role of algorithmic decision making systems in corporate and government service, and growing public concerns regarding the 'black box' nature of many of these systems, the IEEE Standards Association (IEEE-SA) launched the IEEE Global Initiative on Ethics for Autonomous and Intelligence Systems [1] in April 2016. The 'Global Initiative' aims to provide "an incubation space for new standards and solutions, certifications and codes of conduct, and consensus building for ethical implementation of intelligent technologies". As of early 2018 the main pillars of the Global Initiative are:

- a public discussion document "Ethically Aligned Design: A vision for Prioritizing human Well-being with Autonomous and Intelligent Systems" [2], on establishing ethical and social implementations for intelligent and autonomous systems and technology aligned with values and ethical principles that prioritize human well-being in a given cultural context;
- a set of eleven working groups to create the IEEE P70xx series ethics standards, and associated certification programs, for Intelligent and Autonomous systems.

The IEEE P70xx series of ethics standards aims to translate the principles that are discussed in the Ethically Aligned Design document into actionable guidelines or frameworks that can be used as practical industry standards. The eleven IEEE P70xx standards that are currently under development are:

- **IEEE P7000**: Model Process for Addressing Ethical Concerns During System Design
- **IEEE P7001**: Transparency of Autonomous Systems
- **IEEE P7002**: Data Privacy Process
- **IEEE P7003**: Algorithmic Bias Considerations
- **IEEE P7004**: Standard on Child and Student Data Governance
- **IEEE P7005**: Standard on Employer Data Governance
- **IEEE P7006**: Standard on Personal Data AI Agent Working Group
- **IEEE P7007**: Ontological Standard for Ethically Driven Robotics and Automation Systems
- **IEEE P7008**: Standard for Ethically Driven Nudging for Robotic, Intelligent and Autonomous Systems
- **IEEE P7009**: Standard for Fail-Safe Design of Autonomous and Semi-Autonomous Systems
- **IEEE P7010**: Wellbeing Metrics Standard for Ethical Artificial Intelligence and Autonomous Systems

FairWare'18, May 29, 2018, Gothenburg, Sweden

ACM ISBN 978-1-4503-5746-3/18/05…$15.00
https://doi.org/10.1145/3194770.3194773

A brief paper outlining the aims of IEEE P7003 and its relationship to the other IEEE P700x series standards working groups was published in [3] and a tech-industry oriented summary of the eleven IEEE P70xx series standards appeared on the technology-industry blog TechEmergence [4].

In this paper we present a more detailed overview of the scope, structure and development process of the IEEE P7003 Standard for Algorithmic Bias Considerations [5].

IEEE P7003 is aimed to be used by people/organizations who are developing and/or deploying automated decision (support) systems (which may or may not involve AI/machine learning) that are part of products/services that affect people. Typical examples would include anything related to personalization or individual assessment, including any system that performs a filtering function by selecting to prioritize the ease with which people will find some items over others (e.g. search engines or recommendation systems). Any system that will produce different results for some people than for others is open to challenges of being biased. Examples could include:

- Security camera applications that detect theft or suspicious behaviour.
- Marketing automation applications that calibrate offers, prices, or content to an individual's preferences and behaviour.
- etc…

The requirements specification provided by the IEEE P7003 standard will allow creators to communicate to users, and regulatory authorities, that up-to-date best practices were used in the design, testing and evaluation of the algorithm to attempt to avoid unintended, unjustified and inappropriate differential impact on users.

Since the standard aims to allow for the legitimate ends of different users, such as businesses, it should assist them in assuring citizens that steps have been taken to ensure fairness, as appropriate to the stated aims and practices of the sector where the algorithmic system is applied. For example, it may help customers of insurance companies to feel more assured that they are not getting a worse deal because of the hidden operation of an algorithm.

As a practical example, an online retailer developing a new product recommendation system might use the IEEE P7003 standard as follows:

Early in the development cycle, after outlining the intended functions of the new system IEEE P7003 guides the developer through a process of considering the likely customer groups, in order to identify if there are subgroups that will need special consideration (e.g. people with visual impairments). In the next phase of the development, the developer is establishing a testing dataset to validate if the system is performing as desired. Referencing P7003 the developer is reminded of certain methods for checking if all customer groups are sufficiently represented in the testing data to avoid reduced quality of service for certain customer groups.

Throughout the development process IEEE P7003 challenges the developer to think explicitly about the criteria that are being used for the recommendation process and the rationale, i.e. justification, for why these criteria are relevant and why they are appropriate (legally and socially). Documenting these will help the business respond to possible future challenges from customers, competitors or regulators regarding the recommendations produced by this system. At the same time, this process of analysis will help the business to be aware of the context for which this recommendation system can confidently be used, and which uses would require additional testing (e.g. age ranges of customers, types of products).

2 SCOPE

The IEEE P7003 standard will provide a framework, which helps developers of algorithmic systems and those responsible for their deployment to identify and mitigate unintended, unjustified and/or inappropriate biases in the outcomes of the algorithmic system. Algorithmic systems in this context refers to the combination of algorithms, data and the output deployment process that together determine the outcomes that affect end users. Unjustified bias refers to differential treatment of individuals based on criteria for which no operational justification is given. Inappropriate bias refers to bias that is legally or morally unacceptable within the social context where the system is used, e.g. algorithmic systems that produce outcomes with differential impact strongly correlated with protected characteristics (such as race, gender, sexuality, etc).

The standard will describe specific methodologies that allow users of the standard to assert how they worked to address and eliminate issues of unintended, unjustified and inappropriate bias in the creation of their algorithmic system. This will help to design systems that are more easily auditable by external parties (such as regulatory bodies).

Elements include:

- a set of guidelines for what to do when designing or using such algorithmic systems following a principled methodology (process), engaging with stakeholders (people), determining and justifying the objectives of using the algorithm (purpose), and validating the principles that are actually embedded in the algorithmic system (product);
- a practical guideline for developers to identify when they should step back to evaluate possible bias issues in their systems, and pointing to methods they can use to do this;
- benchmarking procedures and criteria for the selection of validation data sets for bias quality control;
- methods for establishing and communicating the application boundaries for which the system has been designed and validated, to guard against unintended consequences arising from out-of-bound application of algorithms;
- methods for user expectation management to mitigate bias due to incorrect interpretation of systems outputs by users (e.g. correlation vs. causation), such as specific action points/guidelines on what to do if in doubt about how to interpret the algorithm outputs;

- a taxonomy of algorithmic bias
- … others yet to be determined

3 STRUCTURE

Discounting procedural sections, dealing with matters of Normative References, Definitions, Conformance etc, the standard document will consist of three main section categories: 1. Foundational sections covering issues related to the fundamentals of understanding algorithmic bias; 2. Algorithmic system design and implementation orientated sections addressing actionable recommendations for identifying and mitigating algorithmic bias; 3. Use cases providing examples of systems where the use of the P7003 standard could provide clear benefits.

3.1 Foundational sections

Foundational sections are currently envisioned to include sections on 'Taxonomy of Bias', 'Legal frameworks related to Bias', 'Psychology of Bias' and 'Cultural context of Bias'. Each of these sections will outline the associated socio-technical aspect of algorithmic bias, providing a background understanding of the reasons for, and importance of, the design/implementation recommendations that are provided in the subsequent sections. Even though the presence of these foundational sections may appear unusual for an industry standard, we believe that they play an important part in an 'ethics' standard such as IEEE P7003. The foundational sections provide a framework of understanding that should allow the designers of algorithmic systems to go beyond a mechanistic 'tick-box' compliance exercise towards a deeper engagement with the underlying ethical issues of algorithmic bias.

3.2 System Design and Implementation sections

The 'algorithmic system design and implementation' orientated sections are currently envisaged to include sections on 'Algorithmic system design stages', 'Person categorizations and identifying of affected groups', 'Representativeness and balance of testing/training/validation data', 'System outcomes evaluation', 'Evaluation of algorithmic processing', Assessment of resilience against external biasing manipulation', 'Assessment of scope limits for safe system usage' and 'Transparent documentation', though it is anticipated that further sections will be added as work progresses.

The intent of these sections is to provide a clear framework of guidance including challenge questions to help designers identify unintended bias issues that would go unnoticed unless specifically looked for. A possible comparison would be the way in which explicit questioning of everyday behavior is required in order to identify and mitigate unconscious bias in management practices.

Proposed solutions to identified causes of algorithmic bias will likely primarily take the form of listing classes of solution methods, with links to relevant work being published at venues such as FairWare, FAT*, KDD and similar publications, in order to reflect the context dependent nature of optimal solutions and the dynamic development in the research on improved methods.

3.3 Use Cases

The Use Cases form an annex to the IEEE P7003 standard document listing a number of illustrative examples of algorithmic systems that resulted in unintended bias, or that highlight specific types of concerns about bias that could be addressed by following the framework provided by IEEE P7003. The inclusion of the Use Cases, and their standardized presentation format, were proposed by a working group participant with experience of industry engagement with standards. They form an important element for 'making the case' for using ethics standards within a corporate context.

Some examples of the use cases that have been gathered so far include:

- "Tay the Nazi chatbot", an example of deliberate system behavior corruption through biased manipulation of inputs by an external 'adversary';
- "The use of facial expression recognition to support diagnostic assessment for patient prioritization", an example of a sensitive application context where differences in operational capability of the system for different population groups can easily result in reputation damaging claims of unjustified bias;
- "Beauty contest judging algorithm that appeared biased to favor lighter skin tones", an example of bias in the training data resulting in biased outcomes that undermined the credibility of the statement purpose of the algorithm (to produce objective beauty contest judgements);
- …

4 METHODOLOGY

Methodologically, the content of the P70xx standards are developed by the working group members through an open deliberation process in which each participant is encourage to suggest content or amendments for the standard document. In order to reflect the broad socio-technical nature of the AI ethics issues addressed by the P70xx standards, the working group members are drawn from a broad range of stakeholders including civil-society organizations, industry and a wide range of academic disciplines. Participation in the working groups is on an individual basis. Even through the participants are affiliated with particular stakeholder organizations, all voices in the standard development process are treated as equals. With the exception of the working group chair and vice-chair, IEEE membership is not required and does not change the status of the participant within the working group.

For the P7003 Standard for Algorithmic Bias Considerations the working group currently consists of 78 participants identifying as having expertise in: Computer Science (18), Engineering (8), Law (6), Business/Entrepreneurship (6), Policy (6), Humanities

(4), Social Sciences (3), Arts (2) and Natural Sciences (1)[1]. In light of the nature of the topic of the P7003 standard, dealing with bias/discrimination, the working group also expressed special concerns about establishing sufficient cultural diversity in its participants. As of early 2018 the participants who chose to indicate their geographic location were from: USA (11), UK (6), Canada (3), Germany (3), Brazil (2), India (2), Japan (2), the Netherlands (2), Australia (1), Belgium (1), Israel (1), Pakistan (1), Peru (1), Philippines (1), S. Korea (1) and Uganda (1); clearly indicating a strong N. America / W. Europe bias that has not yet been resolved. With respect to types of employers, the participants are roughly separated into 1/3 academics, 1/3 industry and 1/3 civil-society affiliations.

During the first eight months, the work of developing the standard focused on growing the participant membership and on exploratory discussions during the monthly conference calls to identify possible factors and sections that could be of relevance for including in the standard. Much of this centered on the foundational sections, which were mostly proposed by working group members as a result of these discussions. In the time between the monthly meetings, working group members are encouraged to develop the document content. During this initial exploratory phase detailed document development was initiated primarily for two of the foundational sections, 'Taxonomy of Bias' and 'Legal frameworks related to Bias'.

As of January 2018, the standard development process has transitioned into the next phase, moving from the initial exploration of the problem space towards consolidation and specification of the standard document content. All P7003 working group members are asked to identify document sections that they will take primary responsibility for, with the aim of having teams of at least two participants for each section. The monthly conference calls will focus on providing updates from each of the teams to the complete working group regarding their progress during the intervening month and any issues that might require input from other teams. This will also be the primary opportunity for all other working group members to raise questions, make suggestions and/or volunteer to (temporarily) contribute to the work of another team.

Once the IEEE P7003 draft document is completed and approved by the IEEE P7003 working group, it will be submitted for balloting approval to the IEEE-SA. The IEEE-SA will send out an invitation-to-ballot to all IEEE-SA members who have expressed an in interest in the subject, i.e. Algorithmic Bias. If the draft receives at least 75% approval, the draft is submitted to the IEEE-SA Standards Board Review Committee, which checks that the proposed standard is compliant with the IEEE-SA Standards Board Bylaws and Operations Manual. The Standards Board then votes to approve the standard, which requires a simple majority. At that point, about 2.5 to 3 years after the proposal for developing the standard was first submitted, the standard is published for use.

[1] Number in brackets indicate number of participants who identified as having this expertise as part of an informal internal survey. Many participants chose not to respond while some chose to indicate multiple expertise.

5 CONCLUSION

As part of the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems a series of eleven ethics standards are under development, designated IEEE P7000 through IEEE P7010. As outlined in this paper, the IEEE P7003 Standard for Algorithmic Bias Considerations aims to provide an actionable framework for improving fairness of algorithmic decision-making systems that are increasingly being developed and deployed by industry, government and other organizations. The IEEE P7003 standard is currently transitioning from an initial exploratory phase into a consolidation and specification phase. Participation in the IEEE P7003 working group is open to all who are interested in contributing towards reducing and mitigating unintended, unjustified and societally unacceptable bias in algorithmic decisions.

Minutes of recent IEEE P7003 working groups meetings are available at [3].

ACKNOWLEDGMENTS

The participation of Ansgar Koene (working group chair) and Liz Dowthwaite (secretary) in the IEEE P7003 Standards development process forms part of the UnBias project supported by EPSRC grant EP/N02785X/1. UnBias is an interdisciplinary research project led by the University of Nottingham in collaboration with the Universities of Oxford and Edinburgh. For more information about the UnBias project, see http://unbias.wp.horizon.ac.uk/

REFERENCES

[1] The IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems. https://ethicsinaction.ieee.org/

[2] *Ethically Aligned Design: A Vision for Prioritizing Human Well-being with Autonomous and Intelligent Systems*, Version 2. IEEE, 2017. http://standards.ieee.org/develop/indconn/ec/autonomous_systems.html

[3] Ansgar Koene. Algorithmic Bias: Addressing Growing Concerns. *IEEE Technology and Society Magazine,* 26, 2, (June 2017), 31-32. DOI: http://dx.doi.org/10.1109/MTS.2017.2697080

[4] Daniel Fagella, The Ethics of Artificial Intelligence for Business Leaders – Should Anyone Care? *TechEmergence*, December 9, 2017. https://www.techemergence.com/ethics-artificial-intelligence-business-leaders/

[5] IEEE P7003 Working Group http://sites.ieee.org/sagroups-7003/