Bias – A Lurking Danger that Can Convert Algorithmic Systems into Discriminatory Entities

Towards a Framework for Bias Identification and Mitigation

Thea Gasser, Eduard Klein
Business Department
Bern University of Applied Sciences (BUAS)
Bern, Switzerland
email: thea.gasser@live.com; eduard.klein@bfh.ch

Lasse Seppänen
Business Information Technology Department
Häme University of Applied Sciences (HAMK)
Hämeenlinna, Finland
email: lasse.seppanen@hamk.fi

***Abstract*—Bias in algorithmic systems is a major cause of unfair and discriminatory decisions in the use of such systems. Cognitive bias is very likely to be reflected in algorithmic systems as humankind aims to map Human Intelligence (HI) to Artificial Intelligence (AI). An extensive literature review on the identification and mitigation of bias leads to precise measures for project teams building AI-systems. Aspects like AI-responsibility, AI-fairness and AI-safety are addressed by developing a framework that can be used as a guideline for project teams. It proposes measures in the form of checklists to identify and mitigate bias in algorithmic systems considering all steps during system design, implementation and application.**

Keywords – Bias; Algorithm; Artificial intelligence; AI-safety; Algorithmic system.

I. INTRODUCTION

Artificial intelligence is present in almost every area of our society, be it in medicine, finance, social media, education, human resource management and many more. This trend will take up a deeper part of people's lives, since according to the Accenture Trend Report [1], about 85% of the executives surveyed plan to invest widely in AI-related technologies over the next three years. Moreover, AI will play a central role in how customers perceive a company and define to a large extent how interactions with their employees and customers take place. AI will become a core competency and will reflect a large part of a company's character. In five years, more than 50% of the customers will no longer choose a service based on the brand but will focus on how much AI is offered for that service [1].

Recently, however, there has been growing concern about unfair decisions made with the help of algorithmic systems that have led to discrimination against social groups or individuals [2] [3]. As an example, Google's image search had been accused of bias indicating fewer women than the reality when searching for the term "CEO". Additionally, Google's advertising system displayed high-income jobs much less to women than to men [4]. The COMPAS algorithm was accused of predicting that "black defendants were far more likely than white defendants to be incorrectly judged to be at a higher risk of recidivism, while white defendants were more likely than black defendants to be incorrectly flagged as low risk" [5]. Microsoft's Tay robot held racist and inflammatory conversations with Twitter users which contained many political statements. It learned from the users' inputs and reflected it in its answers [6]. These and many more known examples show that methods to measure algorithms, recognize and mitigate bias and provide fair AI-software, especially in a highly data oriented machine learning context, are demanded [3] [7].

This article contributes to AI-safety by highlighting that bias in AI is very likely, illustrating possible sources of bias and proposing a framework which supports the identification and mitigation of bias during the design, implementation and application phases of AI-systems.

The following research questions from Gasser [8] are addressed to tackle the above-mentioned aspects: (1) What is expected from AI-systems in relation to how humans make decisions? (2) How is bias present in algorithmic systems that affect human behavior and decisions? (3) How can bias in algorithmic systems be identified? (4) What measures can be taken to mitigate bias in algorithmic systems? Questions (1) and (2) are discussed in Sections III and IV based on literature research, and the proposed framework in Section V gives advice for answering questions (3) and (4) in the context of machine learning based AI projects.

The rest of this paper is organized as follows. Section II describes the research design. In Section III, different types of bias are discussed, followed by related research in Section IV. Section V addresses the bias mitigation framework in finer detail. The conclusions in Section VI close the article.

II. RESEARCH DESIGN

Extensive and systematic literature research has been conducted and the results have been analyzed according to [9]. Systematic analysis was applied by researching specific AI and bias related topics and content, thereby identifying central sources. Based on backward search strategy, further literature was identified. In total, over 100 journal articles, collected works, reference works, books and websites were researched.

As starting points, plain web search and database searches in the scientific portals SAGE journals, ScienceDirect, Springer Link, Google Scholar and the

 ISBN: 978-1-61208-829-7

JSTOR Journal storage have been carried out. A set of search terms has been employed such as "expectations towards AI", "human intelligence", "algorithmic bias", "bias in software development", “mitigating algorithmic bias”, thereby filtering and selecting to 75 relevant sources.

Based on the findings of the literature research, sources of bias and methods for identifying and mitigating bias in algorithmic systems were identified and structured and are systematically presented in Sections III and IV. The findings led to a framework for use in project settings which is described in Section V, thereby identifying and mitigating bias through the use of a metamodel and a set of checklists.

III. FROM HI TO AI

With AI, terms like imitation, simulation or mimicking are repeatedly applied which implies copying something, respectively, someone as, e.g., acting, learning and reasoning like humans [10]. Therefore, if today’s AI-behavior such as Apple’s Siri is considered, it could be claimed that the voice assistant is not intelligent. Looking into details, Apple’s voice assistant is based on evaluated data and facts permitting to offer an appropriate answer [11]. An independently thinking and reasoning machine is not yet present since, amongst other things, an input is still needed. Even though AI acquires intelligence and learns through an autonomous process it lacks sentience and self-awareness and is still only a simulation of HI and nothing more [10].

Despite the expectations and efforts to map HI to AI, to date, there is no system that can be classified as "strong AI", since this would include machines that act completely autonomously and have their own intelligence and self-awareness like humans. However, "weak AI" systems working in a narrowly defined area are used successfully already [12]. Even in the case of self-learning machines, there is initial program code, a model and learning rules so that machine learning can be effective [13]. Because human traits like self-awareness or empathy are missing in today's AI-systems, there is still a gap between AI and HI. This, in turn, implies that partly intelligent systems are shaped by the influence of humankind and with it by cognitive bias which is naturally present in humans and subsequently reflected through individuals and societies in algorithmic systems [14]. Research questions (1) and (2) relate to the decision-making aspects with AI systems.

A. Lack of Transparency in AI Systems

Algorithms are penetrating more and more into people’s lives and will likely overtake even stronger parts of their daily routine so that they will depend heavily on how secure and efficient these algorithms are [15]. Considering that algorithms are becoming more and more complex, and systems may become opaque so that it becomes partly unclear even to the creators of such systems themselves how exactly the interactions in the system(s) take place [16], measures need to be taken in order to minimize undesirable ethical consequences that might arise though the use of such systems. Therefore, the focus must be on potential bias that might arise in the system design, implementation and application phases.

B. Bias and Fairness

Since the term bias is defined as “the action of supporting or opposing a particular person or a thing in an unfair way, because of allowing personal opinions to influence your judgement” (according to the Cambridge Dictionary) the topic of fairness plays a central role. A system might be viewed as fair in some circumstances and in other situations it might be considered unfair. In addition, the presence of bias in an AI system cannot be regarded as evidence of the classification of a system as unfair, which means that neutral or even desirable biases may be present in AI systems without producing undesirable results [17]. Therefore, classifying an AI-system as fair or unfair is subjective and may depend on the viewer, e.g., based on the application context's cultural setting.

Based on these factors, it is important to identify bias and consider whether there is a need for action for reducing it or whether bias should even be used specifically to prevent other in a different part of the system that would have more undesired consequences [17].

The question of whether recognized bias needs to be reduced at all should always be assessed in the individual system context since mitigating bias can be a major effort. On the one hand, several associations demonstrate differences in how and which values are put in the foreground and which seem less important. On the other hand, the situation can reach a level of complexity that no matter what perspective is adopted, some bias will always be identified from a certain point of view. In the end, technology cannot fully answer questions about social and individual values. It is therefore up to humans to make sure that the particular situation is always evaluated in a comprehensive context, meaning taking into account the whole ecosystem around the machine [17].

C. Sources of Bias

Different authors identified various sources of bias in AI-systems. Barfield & Bagallo [18] consider what we call *direct bias* whose sources are related to the core of AI systems: (1) *Input bias* where the source data is biased due to absence of specific information, nonrepresentativeness or reflecting historical biases; (2) *Training bias* which arises when the baseline data is categorized, or the output is assessed; (3) *Programming bias* which emerges in the design phase or when an algorithm modifies itself through a self-learning process.

In [19], sources are identified of what we call *indirect bias*, which are not located in the core of AI systems but in the ecosystem around it: (1) *pre-existing bias* which often emerges through social institutions, practices and attitudes even before a system is designed; (2) *Technical bias*, emerging from technical constraints, e.g., by favoring data (combinations) due to the order or size of screens and visual results presentation; (3) *Emergent bias* arising when using a system outside its intended context of operation.

 ISBN: 978-1-61208-829-7

IV. Related Research

Recently, human aspects of AI have attracted a lot of attention. Not only private companies, research institutions and nonprofit organizations, but also public sector organizations and governments have issued policies and guidelines on human aspects of AI. Many recent publications cite or build on the IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems called "Ethically Aligned Design" (EAD), where methodologies to guide ethical research are presented with the aim of promoting a public debate on how these intelligent and autonomous technologies can be aligned with moral values and ethical principles that prioritize human well-being [20].

The non-profit research organization AlgorithmWatch is developing an "AI Ethics Guideline Global Inventory" [21] to address the question of how automated decision-making systems should be regulated. At the time of writing, more than 80 movements are listed, ranging from a few private companies (e.g. Google, Microsoft, IBM) to organizations (e.g. IEEE, ACM, Bitkom) and government-related organizations (e.g. China, European Commission, Canada, Singapore).

Several metastudies presented the state of the art in human aspects of AI at the time of writing. In [22], an extended list is supplemented by a geographical distribution displayed on a world map. A global convergence of ethical aspects is revealed, emerging around five ethical principles: transparency, fairness, nonmaleficence, responsibility and privacy. It highlights the importance of integrating efforts to develop guidelines and its implementation strategies.

In [23], a comprehensive literature review is presented based on key publications and proceedings complementing existing surveys of psychological, social and legal discussions on the subject with recent advances in technical solutions for AI governance. Based on the literature research, a taxonomy is proposed that divides the field into areas for each of which the most important techniques for the successful use of ethical AI systems are discussed.

All publications mentioned present principles and guidelines for the consideration of ethical aspects in AI systems, thereby addressing research questions (1) and (2). However, they are general and generic and could be used as high-level recommendations only, which are not sufficiently specific for AI projects. The framework presented in Section V further develops these ideas and therefore points the way to the next step in incorporating ethical aspects in a project-oriented environment. Based on a metamodel and a set of checklists, it allows to identify and mitigate bias in AI systems in a project-oriented setting, thereby addressing research questions (3) and (4). The integration of ethical aspects into all project phases during the conception, development and use of a system guarantees a high level of awareness among all project stakeholders.

V. The Bias Mitigation Framework

Awareness of the topic is the first step towards addressing bias in algorithmic systems. According to [24], 92% of AI-leaders make sure their technologists receive ethics training and 74% of the leaders assess AI-outcomes every week. However, it is not enough to just dispose ethics codes that prevent harm. Therefore, establishing usage and technical guidelines and an appropriate mindset among the stakeholders are suggested.

To address bias in algorithmic systems appropriately, an overarching and comprehensive governance must be in place in companies. Using the proposed framework, the project members should be committed to the framework, considering it as a binding standard.

In literature, many possibilities are described to identify bias such as (1) monitoring and auditing an AI system's creation process [25], (2) Applying rapid prototyping, formative evaluation and field testing [19], (3) manipulating test data purposefully in order to determine whether the results are an indication of existing bias in the system [26], (4) using the Socratic method promoting critical thinking and challenging assumptions through answering questions, where scrutiny and reformulation play a central role in the identification and reduction of bias [27].

Tool-based approaches such as IBM's "AI Fairness 360" offer metrics to check for unwanted bias in datasets and machine learning models [28]. Google's "What-If" tool enables visualization of inference results, e.g., for exploring the effect of a certain algorithmic feature and also testing algorithmic fairness constraints [29].

Despite the many approaches that have been suggested in literature and the tools that are available focusing on specific topics in ethical aspects, justification for the proposed framework is in incorporating aspects for all members involved in the process of creating an algorithmic system and all relevant aspects researched.

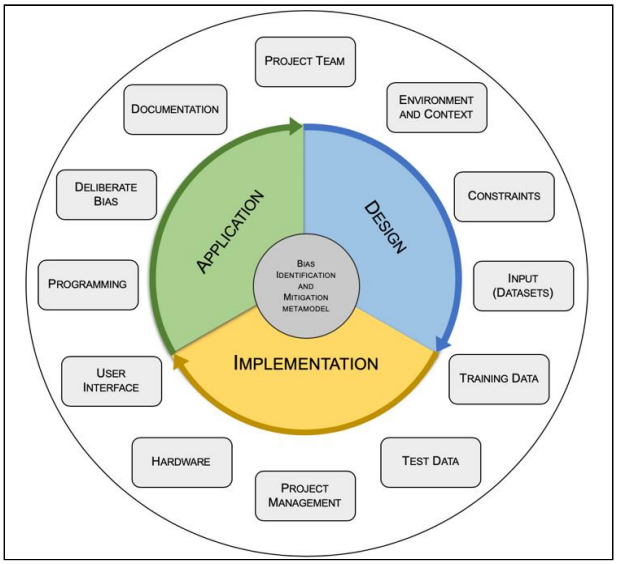

Figure 1. Metamodel for the Bias Mitigation Framework.

The framework consists of a metamodel (see Fig. 1) which is completed by checklists for areas covering the whole software life cycle around design, implementation and application. The areas (e.g. Project Team, Environment,

Content) are illustrated as rectangles in Fig. 1. The elements of each checklist consist of statements and questions that need to be addressed by the project team. The checklists are derived from the findings of the research described in Sections II, III and IV and relate to the research questions (3) and (4).

As an example, the area *Project Team* is subsequently described and detailed in Fig. 2. Knowledge, views and attitudes of individual team members cannot be deleted or hidden, as these are usually unconscious factors, due to everyone's different background and experiences.

Element	Description/Comments	Y/N
Project Team		
All project members have had ethical training	- Members have a confirmation that they have completed courses or workshops or similar - The minimum requirements to consider this element as fulfilled must be defined in the company	
All project members are aware of the topic of bias that exists in the human decision-making process	- Members took part in courses or workshops or similar - The minimum requirements to consider this element as fulfilled must be defined on a project or company level	
All project members know about the fact that human bias can be reflected in an algorithmic system	- Members took part in courses or workshops or similar - The minimum requirements to consider this element as fulfilled must be defined on a project or company level	
All project members consider the same attributes and factors as most relevant in the system context.	- A workshop is held where members share their views. Discrepancies are pointed out and a common understanding is developed. The workshops aim to share views, ideas and openly in order to reveal conflicts and misunderstandings - Due to cultural and background dissimilarities members might (unconsciously) weight attributes differently	
The project team represents stakeholders of all possible end user groups	- Stakeholder analysis comprehensively identifies end user groups with a focus on identifying users who might be disadvantaged through the system outcomes - Stakeholder analysis should be carried out with a change of perspective, where the worst scenario, i.e. if the system behaves discriminatory, identifies the groups that would be disadvantaged. (see area Project Management)	
The project team is a cross-functional team including diversity in ethnicity, gender, culture, education, age and socioeconomic status	- The inputs of all the diverse individuals have to be taken into consideration.	
The project team has representatives from the public and private sector	- Exclusions need to be avoided	
Independent consultants are included for comparison with competing products	- Pre-existing bias in the context of the company's culture, attitude and values can be revealed. - Independent consultants are needed because they are not biased by the companies' views	

Figure 2. Checklist for the metamodel area *Project Team*.

The resulting bias is likely to be transferred into the algorithmic system. Therefore, measures must be taken to ensure the neutrality of the system as far as appropriate. It is necessary that there is an exchange among project members where everyone shares their views and concerns openly, fully and transparently before creating the system. Misunderstandings, ideas of conflict, too much euphoria and unconscious assumptions or invisible aspects might get revealed this way. The checklist in Fig. 2 proposes the following concrete measures for addressing the above-mentioned issues: All project members (1) have had ethical training, (2) are aware of the topic of bias that exists in the human decision-making process, (3) know about the fact that bias can be reflected in an algorithmic system, and (4) consider the same attributes and factors as most relevant in the system context.

Ideally, the project team (1) represents stakeholders of all possible end user groups, (2) is a cross-functional team including diversity in ethnicity, gender, culture, education, age and socioeconomic status, (3) has representatives from the public as well as the private sector. Moreover, independent consultants are included for comparison with competing products.

A. *Checklists*

The metamodel in Fig. 1 illustrates 12 areas of interest, where the project team area was detailed already in Section V. This subsection gives an overview of the 11 remaining areas. For each area the checklist is presented, and the corresponding literature references are explained.

In [19], the different cultural values and attitudes of individuals are emphasised that could collide as they incorporate those into the project work. These aspects are covered by the areas *Environment and Context* and *Constraints* (Fig. 3) in the Framework. In [13] [17] [26] [30], the influence of direct bias is discussed (see "sources of bias" in Section III), leading to the basis for the areas in Fig. 4.

Element	Description/Comments	Y/N
Environment and Context		
All end user groups are included in the testing phase	- The behaviour of end users can only be reliably recorded if they test directly on the live system. Hidden behaviour can thus be detected	
End user groups have been evaluated	- End user groups' behaviour is monitored and evaluated from different perspectives (surveys, interviews, recording behaviour, letting them explain what they do and think while testing)	
Consequences and intentions have been considered	- For what and with what intentions was the system created for? - What is the worst thing that can happen in this algorithm if it starts interacting with others?	
Context is faithful to the original source	- Does the current context represent the one, for which the system was originally created?	

Element	Description/Comments	Y/N
Constraints		
Business aspect reviewed	- Under what circumstances will the system be developed?	
Scope reviewed	- The requirements for the scope of the data set and the diversity are to be determined in the respective project	
Technical aspect reviewed	- Do technical constraints affect the way the system is designed?	
Legal aspect reviewed	- Do regulatory/law constraints affect the way the system is designed?	

Figure 3. Checklist for areas *Environment and Context* and *Constraints*.

Element	Description/Comments	Y/N
Input (Datasets)		
The data set is fully understood	- The meaning of each attribute is understood and its purpose in the system context is clear	
Data is transparent	- Data must be reliable, accurate and kept up to date	
It is ensured that the data set represents the correct scope (enough data representing a population resp. a target group)	- Enough data and diversity are available - The requirements for the scope of the data set and the diversity are to be determined in the respective project.	
The source of the data is known and verified	- Unknown source of the data might lead to that the data is used in a context it was originally not intended to	
The quality of the data is ensured	- Data with low quality will cause even worse outputs since AI-systems might reinforce errors in data sets	
It is clarified which attributes can legally be used	- Use of illegal attribute leads to a system becoming biased even though the attribute itself is not cause for bias	

Element	Description/Comments	Y/N
Training Data		
The training data set is still as representative as the original data set	- Adjusting source data to training data can bear exclusion which needs to be prevented	
Added or omitted attributes are carefully chosen and justified	- One attribute can influence different areas in a system. Interconnectedness needs to be considered	

Element	Description/Comments	Y/N
Test Data		
Test data is independent	- The system uses test data it has never seen before	
Test data is defined	- Test scenarios are defined which are designed to detect bias which could be caused by a certain attribute	
Test data is reviewed	- Tests include omission and addition of attributes to test how system output changes	

Figure 4. Checklists for the areas concerning *direct bias*, derived from "sources of bias" in Section III.

It is suggested that the complete algorithmic system lifecycle is accompanied and controlled through all phases with a project management approach. The classical element "risk analysis" must be expanded with a focus on risk factors that could favour bias and the effects recognised bias could have. Isele [27] suggests that critical questions should be asked, critical thinking adopted, assumptions challenged, and the results of the system evaluated. Aspects on the Project Management area are gathered in Fig. 5.

Element	Description/Comments	Y/N
Project Management		
Project management process includes methods that focus on bias issues	- Stakeholder analysis is adjusted for disadvantaged group identification in worst case	
Risks concerning bias are assessed and known to each team member	- Risk analysis is adjusted for additional focus on bias and worst-case scenarios provoking to bias	
Critical thinking is promoted and demanded at every stage of the system creation process	- How would changes to a data point affect the model's prediction? - Does it perform differently for various groups? For example, historically marginalised people? - How diverse is the dataset I am testing my model on? - Is the system context the one the system was intended to? - Can the outcome/result/system recommendation be justified? - How diverse is the dataset I am testing my model on? - Does it perform differently for various groups–for example, historically marginalized people? - How would changes to a data point affect my model's prediction?	
Perspectives are changed continuously to challenge assumptions	- Different points of views ensure identification of hidden assumptions	
Monitoring measures are defined, communicated and applied	- End user groups' behaviour is monitored and evaluated from different perspectives (surveys, interviews, recording behaviour, letting them explain what they do and think while testing)	
Auditing measures are defined, communicated and applied	-	
Workshops / meetings are set frequently which address upcoming doubts of team members	- Critical thinking is continuously fostered in workshops and outside	
Scenario thinking is fostered	-	
Freedom of expression is guaranteed and desired	- Every input of any team member can reveal hidden bias	

Figure 5. Checklist for the area *Project Management.*

Hardware limitations, such as screen size or performance bottlenecks, could influence system output [19]. The design of visual representations of objects could also be a source of bias, requiring a careful design of the graphical user interface [31]. Checklists for hardware limitations and Graphical User Interface (GUI) design are detailed in Fig. 6.

Element	Description/Comments	Y/N
Hardware		
Hardware limitations	- Do hardware limitations exist?	
Influence on creation process	- Do these limitations influence the system creation process?	
Influence on production environment	- Do these limitations influence the system's functionality in the production environment?	

Element	Description/Comments	Y/N
User Interface		
Visual aspects are determined appropriately	- The font-style, font-size, font-colour and placement of text are justified and reflect the intention of the system's functionality - Colour, size and placement of forms and graphics are justified and reflect the intention of the system's functionality	
Visual result	- Does visual result representation (alphabetically or random) make any difference (user always choses the results displayed first?)	
Navigation	- Does a change in navigation representation lead the user to favour different results?	
Graphical User Interface	- Is graphical UI limiting/favouring data over other data?	
Language Aspects	- How does the chosen language influence the user's perception and interpretation in different contexts and circumstances? - Is a translation of data/information necessary? - Do the information and results become distorted through the application of translation? - How is the translation interpreted by the end users?	
Alternative GUI	- The system features are changed, and end users are monitored once more in order to see how their behaviour changes - Several features may need to be changed various times in order to reveal hidden assumptions of end users	

Figure 6. Checklist for areas *Hardware* and *UI.*

Sources of bias in programming and documentation and discussion on deliberate bias [17] are given in Fig. 7.

Element	Description/Comments	Y/N
Programming		
Code reviews take place	- Measures aim to understand adapted or reused code fully	
Independent code audits are conducted	- Independent audits foster considering the code from a different point of view and reveal unconscious assumptions	
Possible user behaviour is analysed beforehand to keep a learning system from adopting discriminatory behaviour	- Thinking outside the box is fostered especially considering word and language usage in the system context - The system can handle discriminatory user behaviour	

Element	Description/Comments	Y/N
Deliberate Bias		
Bias is identified and categorized	- Are the identified biases considered as good, neutral or bad ones? - Is there any bias which was implemented on purpose in order to mitigate other?	
It is ensured that all the identified biases are monitored during the whole system creation process	- Bias needs to be tracked and changes identified as well as recorded throughout every stage of the project	

Element	Description/Comments	Y/N
Documentation		
Availability of relevant information	- Traceability, justification and business continuity is ensured	
Comprehensible documentation	- The language may only contain such a high degree of complexity and technical language that every project member understands it - Prevention of misunderstandings is ensured	
Documentation has been reviewed and approved	- The documentation needs to be reviewed by several project members and stakeholders	

Figure 7. Checklist for areas *Programming, Documentation and Deliberate Bias.*

The presence of deliberate bias might be surprising at first, however, is applied in some cases to prevent bias from arising in another, more important area of a system. As an example, a statistically biased estimator in an algorithm might exhibit significant reduced variance on small sample sizes, thereby greatly increasing reliability and robustness in future use [17].

B. Application of the Framework

Based on the outcome of the above-mentioned literature research, the approach presented is intended to be an initial framework that can be adapted to specific needs within a given project context. It comes in shape of a guideline complemented with checklists, e.g., for the members of a project team.

The adjustments could be made based on an adapted understanding of system neutrality which may be specific for the respective application or application domain. If the proposed framework is used in a mandatory manor within a project, it is very likely that the developed application reflects the neutrality defined by the project team or company.

Verifying that the framework has been applied and the requirements have been met will help to determine the extent to which the system is neutral and the need for appropriate action.

VI. CONCLUSION

Since currently there are only weak AI-systems which lack self-awareness and depend on human advice in shape of created models and selected training data, human bias is naturally and unintentionally reflected in crafted algorithmic systems. A framework has been proposed which helps to

identify and mitigate bias in algorithmic systems, covering aspects of the complete life cycle of such software systems.

Since the framework in its current state is a synthesis of desk research, future research should implement the approach in realistic software project situations such that its added value could be observed, evaluated, validated and subsequently adapted based on the project experiences. During validation, each metamodel area would require separately assessing the priority of the questions and requirements in the checklists and ensuring useful answers.

In addition, it would be useful to investigate to what extent automation of the use of the framework could mitigate subjective opinions and views of the stakeholders involved. As an example, the following scenario could be realized: Information about the adapted framework (metamodel areas and checklist elements), which is considered standard for ensuring system neutrality up to a certain point in the respective project, could be supported by a software system. During the project, the checklists are continuously filled with data by the project team, thereby enabling analysis of the process, comparison of different implementations of the framework and revealing indications where the recommendations were complied with and where it was not.

On the one hand, a specific project team would always be aware when creating an algorithmic system, which of the specified areas would not be adhered to and could exhibit potential bias. On the other hand, this mechanism could also be used for end users. They could more easily assess how reliable the results of the AI-system are and which areas need more attention regarding bias. The impact of decisions taken through the AI-system's suggestions can be better analyzed by knowing which areas do not comply with the elaborated standard.

However, to reach this point, there are several aspects that need to be considered. Elements from the checklists would have to be detailed at micro level to define, for example, what a stakeholder is or how it can be verified that the test user belongs to a respective gender. Instead of a yes/no check mark in the check lists, there could be more detailed measures, e.g., indication of the level to which a team member has received ethical training. In addition, mechanisms could be integrated to take account of the truthfulness of the answers in the checklists.

REFERENCES

[1] Accenture, "AI is the new UI – Experience Above All," Accenture Technology Vision, 2017. https://www.accenture.com/_acnmedia/accenture/next-gen-4/tech-vision-2017/pdf/accenture-tv17-full.pdf. [retrieved: 18-Aug-2020].

[2] A. Koene, "Algorithmic Bias: Addressing Growing Concerns," IEEE Technol. Soc. Mag., vol. 36, no. 2, pp. 31–32, Jun. 2017.

[3] M. Veale and R. Binns, "Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data," Big Data Soc., vol. 4, no. 2, pp. 1-17, Dec. 2017.

[4] D. Cossins, "Discriminating algorithms: 5 times AI showed prejudice," New Scientist, 2018. https://www.newscientist.com/article/2166207-discriminating-algorithms-5-times-ai-showed-prejudice/. [retrieved: 18-Aug-2020].

[5] S. Feuerriegel, M. Dolata, and G. Schwabe, "Fair AI," Bus. Inf. Syst. Eng., pp. 379-384, 2020.

[6] E. Hunt, "Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter," The Guardian, 24-Mar-2016. https://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter. [retrieved: 18-Aug-2020].

[7] M. Hardt, E. Price, and N. Srebro, "Equality of opportunity in supervised learning," in Advances in Neural Information Processing Systems, pp. 1-9, 2016.

[8] T. Gasser, "Bias – A lurking danger that can convert algorithmic systems into discriminatory entities," HAMK University, Finland, 2019.

[9] M. Kornmeier, Wissenschaftlich schreiben leicht gemacht (academic writing made easy), 8th ed. Haupt Verlag, 2018.

[10] S. Holder, "What is AI, really? And what does it mean to my business?," 2018. https://www.sas.com/en_ca/insights/articles/analytics/local/what-is-artificial-intelligence-business.html. [retrieved: 18-Aug-2020].

[11] A. Goel, "How Does Siri Work? The Science Behind Siri," Magoosh Data Science Blog, 2018. https://magoosh.com/data-science/siri-work-science-behind-siri/. [retrieved: 18-Aug-2020].

[12] J. R. Searle, "Minds, brains, and programs," Behav. Brain Sci., vol. 3, no. 3, pp. 417–457, 1980.

[13] E. Alpaydın, Introduction to Machine Learning, 2nd ed. 2012.

[14] J. Burrell, "How the machine 'thinks': Understanding opacity in machine learning algorithms," Big Data Soc., vol. 3, no. 1, pp. 1-12, 2016.

[15] A. Smith, "Franken-algorithms: the deadly consequences of unpredictable code," The Guardian, 2018. https://www.theguardian.com/technology/2018/aug/29/coding-algorithms-frankenalgos-program-danger. [retrieved: 18-Aug-2020].

[16] C. Smith, B. McGuire, T. Huang, and G. Yang, "The History of Artificial Intelligence," Univ. of Washington, 2006.

[17] D. Danks and A. J. London, "Algorithmic Bias in Autonomous Systems," in Proceedings of the Twenty-Sixth International Joint Conference on Artificial Intelligence, 2017, pp. 4691–4697.

[18] W. Barfield and U. Pagallo, "Research Handbook on the Law of Artificial Intelligence. Edited by Woodrow Barfield and Ugo Pagallo. Cheltenham, UK," Int. J. Leg. Inf., vol. 47, no. 02, pp. 122–123, Sep. 2019.

[19] B. Friedman and H. Nissenbaum, "Bias in computer systems," ACM Trans. Inf. Syst., vol. 14, no. 3, pp. 330–347, Jul. 1996.

[20] IEEE Global Initiative on Ethics of Autonomous and Intelligent Systems, "Ethically Aligned Design," 2019.

[21] AlgorithmWatch, "AI Ethics Guidelines Global Inventory," 2019. https://inventory.algorithmwatch.org/about. [retrieved: 18-Aug-2020].

[22] A. Jobin, M. Ienca, and E. Vayena, "The global landscape of AI ethics guidelines," Nat. Mach. Intell., vol. 1, no. 9, pp. 389–399, Sep. 2019.

[23] H. Yu, Z. Shen, C. Miao, C. Leung, V. R. Lesser, and Q. Yang, "Building Ethics into Artificial Intelligence," in Proceedings of the Twenty-Seventh International Joint Conference on Artificial Intelligence, 2018, pp. 5527–5533.

[24] SAS, "Organizations Are Gearing Up for More Ethical and Responsible Use of Artificial Intelligence, Finds Study," 2018. https://www.sas.com/en_id/news/press-releases/2018/september/artificial-intelligence-survey-ax-san-diego.html. [retrieved: 18-Aug-2020].

 ISBN: 978-1-61208-829-7

[25] A. Raymond, “The Dilemma of Private Justice Systems: Big Data Sources, the Cloud and Predictive Analytics,” Northwest. J. Int. Law Bus., vol. 22, pp. 1-44, 2014.

[26] I. Žliobaitė, “Measuring discrimination in algorithmic decision making,” Data Min. Knowl. Discov., vol. 31, no. 4, pp. 1060–1089, 2017.

[27] E. Isele, “The Human Factor Is Essential to Eliminating Bias in Artificial Intelligence,” Chatham House, 2018. https://www.chathamhouse.org/expert/comment/human-factor-essential-eliminating-bias-artificial-intelligence. [retrieved: 18-Aug-2020].

[28] R. K. E. Bellamy et al., “AI Fairness 360: An extensible toolkit for detecting and mitigating algorithmic bias,” IBM J. Res. Dev., vol. 63, no. 4–5, pp. 1–15, 2019.

[29] J. Wexler, M. Pushkarna, T. Bolukbasi, M. Wattenberg, F. Viegas, and J. Wilson, “The What-If Tool: Interactive Probing of Machine Learning Models,” IEEE Trans. Vis. Comput. Graph., vol. 26, no. 1, pp. 56-65, 2020.

[30] S. Barocas and A. D. Selbst, “Big Data’s Disparate Impact,” 104 Calif. Law Rev. pp. 671-732, 2016.

[31] S. J. Russell and P. Norvig, Artificial Intelligence: A Modern Approach, 3rd ed. Upper Saddle River: Prentice Hall, 2010.

 ISBN: 978-1-61208-829-7