Validation of a Framework for Bias Identification and Mitigation in Algorithmic Systems

Thea Gasser, Rémy Bohler, Eduard Klein
Business Department
Bern University of Applied Sciences (BUAS)
Bern, Switzerland
email: thea.gasser@live.com
remy.bohler@sunrise.ch
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). We conducted an extensive literature review on the identification and mitigation of bias, leading to precise measures for project teams building AI systems. Moreover, we developed an awareness-raising framework for use as a guideline for project teams, addressing AI responsibility, AI fairness, and AI safety. The framework proposes measures in the form of checklists to identify and mitigate bias in algorithmic systems considering all steps during system design, implementation, and application. We validated the framework successfully in the context of industrial AI projects.**

Keywords – Bias Framework; Artificial intelligence; Algorithmic system; Validation.

I. INTRODUCTION

This paper is a long version and an extension of the Bias Identification and Mitigation Framework presented in [1]. The original material contains the framework definition and application. Since validating the approach is crucial and contributes to the improvement and optimization of the framework, the validation process was carried out in an industrial context and is discussed in detail below.

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. AI will become a greater part of people's lives since the Accenture Trend Report [2] states, that about 85% of the executives surveyed plan to invest widely in AI (artificial intelligence) 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 [2].

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 [3] [4] [5]. 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 [6]. The COMPAS algorithm has been accused of misclassifying black defendants as at risk of recidivism far more often than white defendants, while white defendants are misclassified as low risk far more often than black defendants [7].

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 [8]. These and many other well-known examples show the demand for methods to measure algorithms, recognize and mitigate bias and provide fair AI software, especially in a high data-oriented machine learning context [4] [9].

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

The following research questions from Gasser [10] and Bohler [11] are addressed to tackle the above-mentioned aspects:

(1) What can be expected from AI systems compared to human decision-making?
(2) In what form is bias present in algorithmic systems?
(3) How can bias in algorithmic systems be identified?
(4) What measures can be taken to mitigate bias in algorithmic systems?
(5) How could bias be identified and referenced?
(6) To what extent does the framework application contribute to AI project improvement?

Sections III and IV discuss questions (1) and (2) based on literature research, and the proposed framework in Section V advises answering questions (3) and (4) in the context of machine learning based AI projects. Section VI discusses questions (5) and (6) during framework validation.

The rest of this paper is organized as follows. Section II describes the research design. Section III discusses various types of bias, followed by related research in Section IV. Section V addresses the bias mitigation framework in finer detail. Section VI discusses the framework validation in the context of industrial projects. The conclusions in Section VI close the article.

II. Research Design

We conducted a literature search, mainly in SAGE Journals, ScienceDirect, Springer Link, Google Scholar, and the JSTOR Journal Storage. A range of search terms was used, such as "expectations towards AI", "human intelligence", "algorithmic bias", "bias in software development", “mitigating algorithmic bias”, resulting in the selection of about 125 sources. These were narrowed down to 75 relevant sources by cross-reading the abstracts and restricting them to articles from 2016 or later.

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 using a metamodel, a set of checklists, and a one-pager template including the bias assessment criteria visualizing the assessment result.

We conducted validation through a Delphi method [13] with members from industrial AI projects as experts. After an initial phase of introducing the framework, we applied it to suitable projects. We collected feedback in the form of responses to the framework checklists, and subsequently summarized and structured it. The results were presented to the project members and refined in subsequent iterations until sufficient stability was achieved.

In addition to project-related aspects, we also raised meta-questions about applicability, usefulness, size, comprehensibility, and coverage concerning framework improvements.

In addition to a three-part document (project description, framework application, recommendations for bias mitigation), the validation results are visualized as a one-pager, making the bias assessment visible at a glance.

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 [14]. 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 [15]. An independently thinking and reasoning machine is not yet present since, amongst other things, 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 (human intelligence) and nothing more [14].

Despite the expectations and efforts to map HI to AI, to date, no system can be classified as "strong AI", since this would include machines that act completely autonomously and have their intelligence and self-awareness like humans. However, "weak AI" systems working in a narrowly defined area are used successfully already [16]. 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 [17]. 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 [18]. Research questions (1) and (2) relate to the decision-making aspects of AI systems.

A. *Lack of Transparency in AI Systems*

Algorithms are penetrating more and more into people’s lives and are likely to play an increasingly important role in their everyday lives, so they will depend to a large extent on how secure and efficient these algorithms are [19]. Algorithms are becoming increasingly complex, and systems may become opaque so that it becomes partly unclear even to the creators of such systems how exactly the interactions in the system(s) take place [20]. Measures must therefore be taken to minimize undesirable ethical consequences that could arise from the use of such systems. The main focus must be on the potential bias that might occur 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 judgment” (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 [21]. 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 bias in a different part of the system that would have more undesired consequences [21].

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 [21].

C. *Sources of Bias*

Various sources of bias in AI systems have been identified. Barfield & Bagallo [22] consider what we call *direct bias* whose sources are related to the core of AI systems:

- *Input bias* where the source data is biased due to the absence of specific information, non-representativeness, or reflecting historical biases.
- *Training bias* arises when the baseline data is categorized, or the output is assessed.
- *Programming bias* emerges in the design phase or when an algorithm modifies itself through a self-learning process.

In [23], sources are identified of what we call *indirect bias*, which is not located in the core of AI systems but in the surrounding ecosystem:

- *pre-existing bias*, which often emerges through social institutions, practices, and attitudes even before a system is designed.
- *Technical bias*, emerging from technical constraints, e.g., by favoring data (combinations) due to the order or size of screens and visual results presentation.
- *Emergent bias* arises when using a system outside its intended context of operation.

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). This presents methodologies to guide ethical research to promote public debate on how these intelligent and autonomous technologies can be aligned with moral values and ethical principles that prioritize human well-being [24].

The non-profit research organization AlgorithmWatch is developing an "AI Ethics Guideline Global Inventory" [25] 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 meta-studies presented the state of the art in human aspects of AI at the time of writing. In [26], an extended list is supplemented by a geographical distribution displayed on a world map. 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 [27], 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.

[28] presents a framework for algorithmic hygiene and employs best practices to identify and mitigate them. A set of questions is compiled analyzing bias impact by the insight of 40 thought leaders who participated in a roundtable.

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. [29] states, that 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, 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 [30], (2) applying rapid prototyping, formative evaluation, and field testing [23], (3) manipulating test data purposefully to determine whether the results are an indication of existing bias in the system [31], (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 [32].

The methods that aim to prevent bias in AI can be divided into technical approaches and awareness-raising approaches. Technical approaches attempt to integrate ethical principles into the design process of AI systems. Awareness-raising approaches aim to highlight the presence and the risks of bias through education and awareness initiatives so that members of AI projects could take care of bias problems and act responsibly in projects in this regard [33].

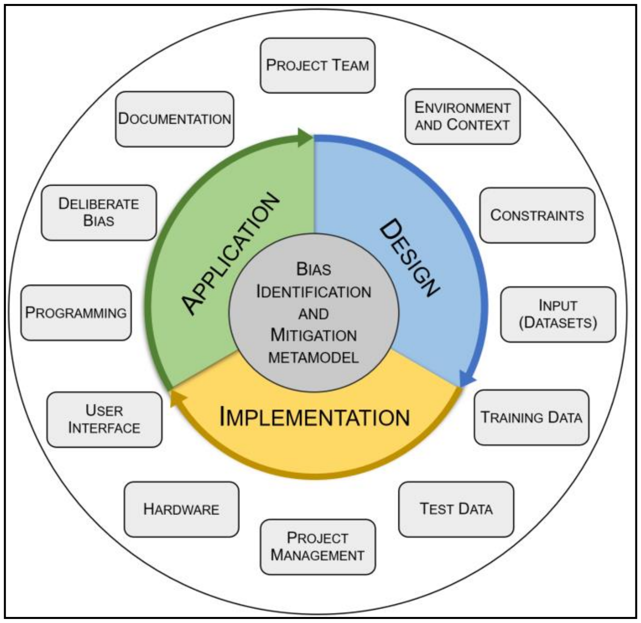


Figure 1. Metamodel for the Bias Mitigation Framework.

In contrast to the awareness-raising approach adopted in the presented framework, technical approaches such as IBM's "AI Fairness 360" offer metrics to check for unwanted bias in datasets and machine learning models [34]. 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 [35]. [36] tackles algorithmic bias, building models incorporating fair representation learning.

Many approaches have been suggested in the literature and tools 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.

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, Context) 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).

Project Team

(a) 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

(b) 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

(c) 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

(d) 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 reveal conflicts and misunderstandings
- Due to cultural and background dissimilarities members might (unconsciously) weigh attributes differently

(e) 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)

(f) 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 must be taken into consideration

(g) The project team has representatives from the public and private sector

-Exclusions need to be avoided

(h) 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* with commented elements (a), (b), …

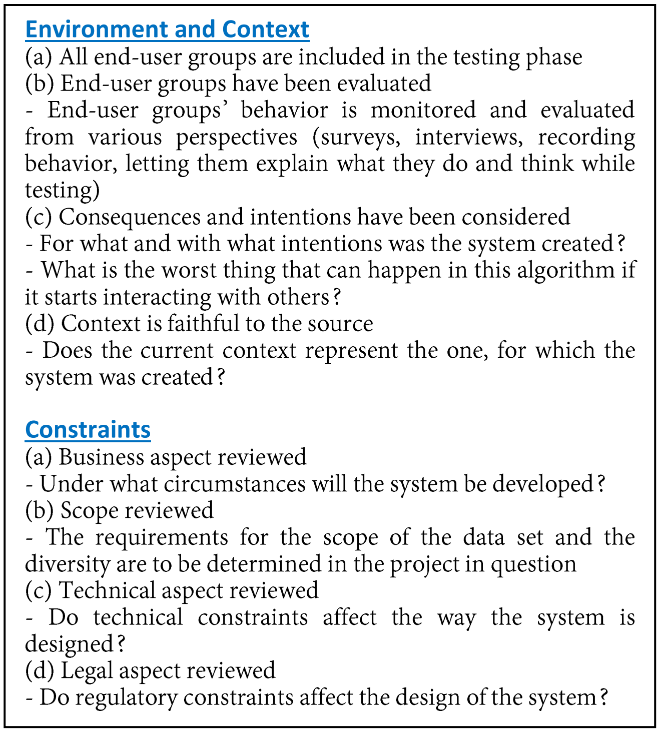
As an example, the area *Project Team* is subsequently described and detailed in Fig. 2. Knowledge, views, and attitudes of individual team members cannot be deleted or hidden, as these are usually unconscious factors, due to everyone's various backgrounds and experiences.

The resulting bias is likely to be transferred into the algorithmic system. Therefore, measures must be taken to ensure system neutrality as far as appropriate. There must be 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 bias topic 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 provides an overview of the 11 remaining areas. For each area, the checklist is presented, and the corresponding literature references are explained.

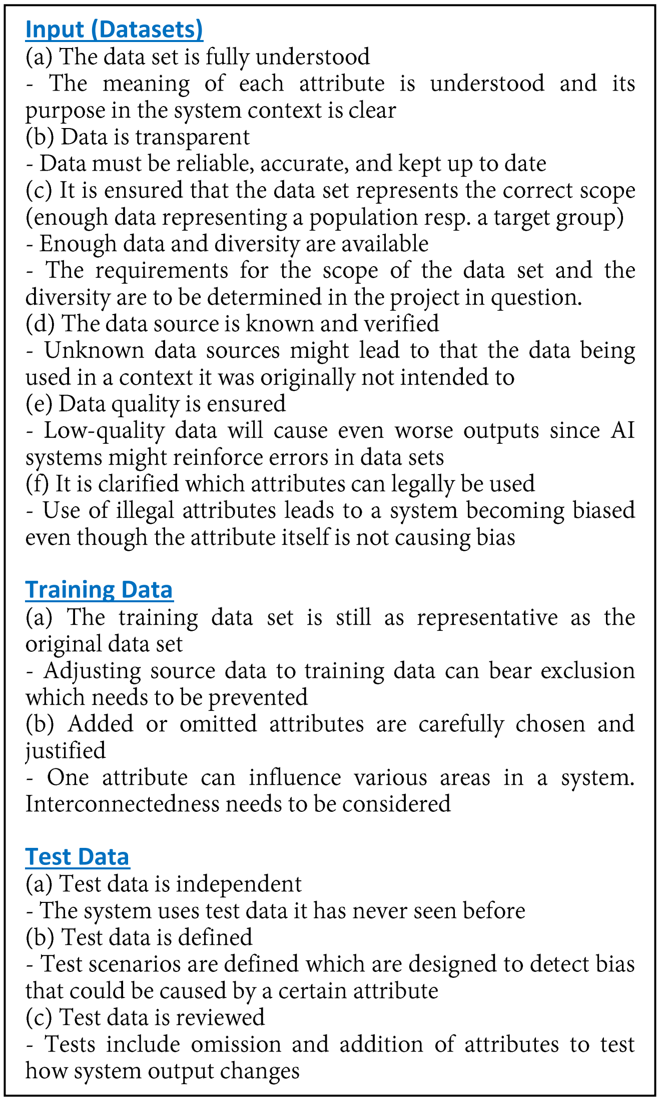


Environment and Context
(a) All end-user groups are included in the testing phase
(b) End-user groups have been evaluated
- End-user groups' behavior is monitored and evaluated from various perspectives (surveys, interviews, recording behavior, letting them explain what they do and think while testing)
(c) Consequences and intentions have been considered
- For what and with what intentions was the system created?
- What is the worst thing that can happen in this algorithm if it starts interacting with others?
(d) Context is faithful to the source
- Does the current context represent the one, for which the system was created?

Constraints
(a) Business aspect reviewed
- Under what circumstances will the system be developed?
(b) Scope reviewed
- The requirements for the scope of the data set and the diversity are to be determined in the project in question
(c) Technical aspect reviewed
- Do technical constraints affect the way the system is designed?
(d) Legal aspect reviewed
- Do regulatory constraints affect the design of the system?

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

In [23], the various cultural values and attitudes of individuals are emphasized 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. [17] [21] [31] [37] discuss the influence of direct bias (see "sources of bias" in Section III), leading to the areas in Fig. 4.



Input (Datasets)
(a) The data set is fully understood
- The meaning of each attribute is understood and its purpose in the system context is clear
(b) Data is transparent
- Data must be reliable, accurate, and kept up to date
(c) 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 project in question.
(d) The data source is known and verified
- Unknown data sources might lead to that the data being used in a context it was originally not intended to
(e) Data quality is ensured
- Low-quality data will cause even worse outputs since AI systems might reinforce errors in data sets
(f) It is clarified which attributes can legally be used
- Use of illegal attributes leads to a system becoming biased even though the attribute itself is not causing bias

Training Data
(a) 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
(b) Added or omitted attributes are carefully chosen and justified
- One attribute can influence various areas in a system. Interconnectedness needs to be considered

Test Data
(a) Test data is independent
- The system uses test data it has never seen before
(b) Test data is defined
- Test scenarios are defined which are designed to detect bias that could be caused by a certain attribute
(c) 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 favor bias and the effects recognized bias could have. Isele [32] suggests that critical questions should be asked, critical thinking adopted, assumptions challenged, and the systems' results evaluated. Project Management area aspects are gathered in Fig. 5.

Project Management
(a) The project management process includes methods that focus on bias issues
- Stakeholder analysis is adjusted for disadvantaged group identification in the worst case
(b) 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 bias
(c) 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 marginalized 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?
(d) Perspectives are changed continuously to challenge assumptions
- Various points of view ensure the identification of hidden assumptions
(e) Monitoring measures are defined, communicated, and applied
- End-user groups' behavior is monitored and evaluated from various perspectives (surveys, interviews, recording behavior, letting them explain what they do and think while testing)
(f) Auditing measures are defined, communicated, and applied
(g) Workshops/meetings are set frequently which address upcoming doubts of team members
- Critical thinking is continuously fostered in workshops and outside
(h) Scenario thinking is fostered
(i) 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 [23]. The design of visual representations of objects could also be a source of bias, requiring a careful design of the graphical user interface [38]. Checklists for hardware limitations and Graphical User Interface (GUI) design are detailed in Fig. 6.

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 significantly reduced variance on small sample sizes, thereby greatly increasing reliability and robustness in future use [21].

Sources of bias in programming and documentation and discussion on deliberate bias [21] are given in Fig. 7.

Hardware
(a) Hardware limitations
- Do hardware limitations exist?
(b) Influence on the creation process
- Do these limitations influence the system creation process?
(c) Influence on production environment
- Do these limitations influence the system's functionality in the production environment?

User Interface
(a) Visual aspects are determined appropriately
- The font style, font size, font color, and placement of text are justified and reflect the intention of the system's functionality
- Color, size, and placement of forms and graphics are justified and reflect the intention of the system's functionality
(b) Visual result
- Does visual result representation (alphabetically or random) make any difference (user always chooses the results displayed first?)
(c) Navigation
- Does a change in navigation representation lead the user to favor different results?
(d) Graphical User Interface
- Is graphical UI limiting/favoring data over other data?
(e) Language Aspects
- How do the chosen language influence the user's perception and interpretation in various 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?
(f) Alternative GUI
- The system features are changed, and end-users are monitored once more to see how their behavior changes
- Several features may need to be changed various times to reveal hidden assumptions of end-users

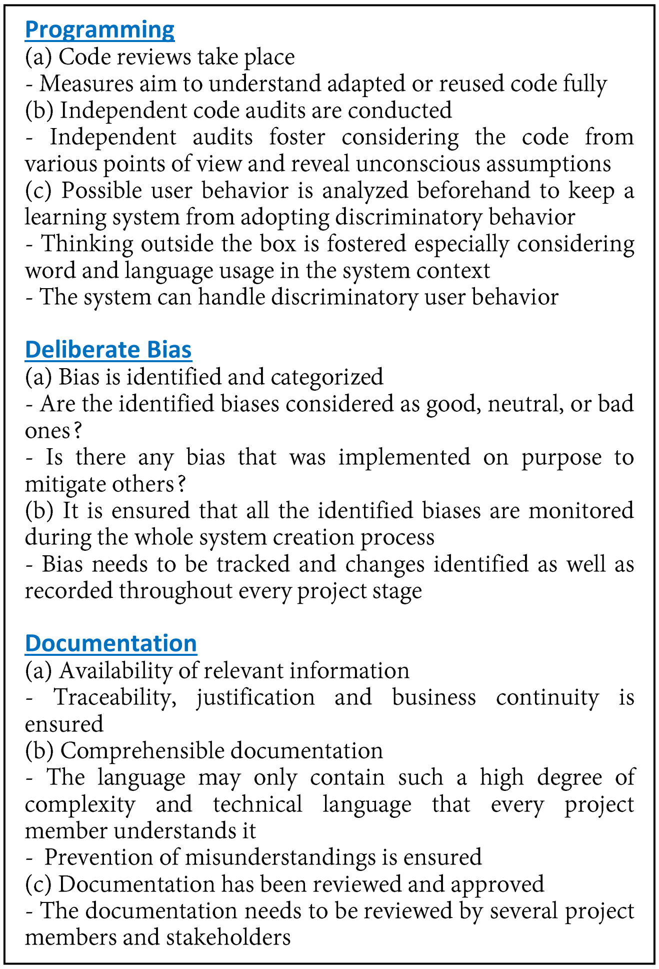
Figure 6. Checklist for areas *Hardware* and *UI.*

The checklists in table form for use in a bias assessment and supplementary material are available at https://instructor.github.io/bias/.

B. Framework Use

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 the 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 application or application domain in question. If the proposed framework is used mandatory within a project, it is very likely that the developed application reflects the neutrality defined by the project team or company.



Programming
(a) Code reviews take place
- Measures aim to understand adapted or reused code fully
(b) Independent code audits are conducted
- Independent audits foster considering the code from various points of view and reveal unconscious assumptions
(c) Possible user behavior is analyzed beforehand to keep a learning system from adopting discriminatory behavior
- Thinking outside the box is fostered especially considering word and language usage in the system context
- The system can handle discriminatory user behavior

Deliberate Bias
(a) Bias is identified and categorized
- Are the identified biases considered as good, neutral, or bad ones?
- Is there any bias that was implemented on purpose to mitigate others?
(b) 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 project stage

Documentation
(a) Availability of relevant information
- Traceability, justification and business continuity is ensured
(b) 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
(c) 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.*

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. Framework Validation

To reveal the advantages and disadvantages of the proposed framework, we conducted its application and evaluation in practical AI projects. Recommendations for the proper use and improvement of the approach are derived from the results.

First, we looked at a chatbot project of a Swiss insurance company to automate customer communication. Technically, it is based on NLP (natural language processing). The chatbot helps customers orientate and navigate their website to find the proper information based on customer queries. The learning mechanism enables the chatbot to improve customer help constantly. At the time of writing even non-specific customer input is being properly processed in many cases [39].

The second study object was the Smart Animal Health project from a Swiss government agency collecting data on farm animals to evaluate the effectiveness of measures in the farm animal sector. The system is designed to improve the well-being of animals, establishing an early-warning system for on-site problems. Another goal is the identification of trends.

Source data are extracted from governmental open data and private data sources deducing key indicators such as disease symptoms, animal behavior, and husbandry conditions. By linking the data, a more complete picture of livestock farms emerges. Statements on animal health can be derived for farms or the farm animal population under consideration as a whole. Technically, the project uses machine learning algorithms, specifically supervised learning [40].

The Delphi-based project member involvement resulted in a textual summary for each of the 12 framework areas of interest (detailed in [11]) and a visual presentation in the form of a one-pager (see Fig. 8) reflecting each area of interest through a rectangular shape. The areas of interest are placed in three columns, arranged from left to right and top to bottom. Each rectangular area contains the elements from the corresponding checklist.

Based on the project members' feedbacks, the elements are highlighted in green (darker background in black and white print) if the element has successfully been addressed, in yellow (lighter background in black and white print) if the element has partially been addressed, or in white resp. not highlighted if it is not applicable in the project context in question. The distinction between successful and partially successful is open to interpretation. We determined the final classifications during the iterations in the Delphi process.

The one-pager template, the results of the two projects studied, and supplementary material can be found at https://instructor.github.io/bias/.

A. Framework Application

In both projects, a sequence of activities has been carried out:

a) explanation of the framework
b) application to the project
c) evaluation of the results
d) bias identification
e) derivation of recommendations for mitigating bias.

Steps b) to d) were repeated until sufficient stability was achieved.

In the one-pager for the chatbot project (see Fig. 8) the yellow regions show potential for improvement concerning bias aspects in the project, while the green regions satisfy the bias criteria in question to a sufficient degree. Since all elements in the one-pager are highlighted in green or yellow, they could be well reflected and were considered relevant by the project members involved.

Elements from the following areas of interest were identified in the yellow regions: project team, project management, programming, and deliberate bias. After analyzing the comments on the yellow elements, the following recommendations were derived to mitigate bias:

- Interaction bias [41] can arise as the learning chatbot adopts biased customer statements into its interaction pattern. It is recommended to monitor the interactions and proactively adjust them if necessary.
- In learning systems, bias can change dynamically based on the machine learning process. Deliberate bias should be monitored carefully to prevent its spillover to unwanted bias.
- To raise awareness and establish know-how, it is recommended to conduct training and workshops on the topic of bias.
- It is recommended that the bias assessment process be repeated at appropriate intervals to achieve ethical sustainability in AI projects.

Based on the project partners' answers to the framework application, we used color-coding to mark not only the areas of compliance with the framework but also the areas of deviation. Subsequently, we further broke down the areas marked in yellow by documenting in a project-specific way how the potential for bias in these areas could have a negative impact on the project.

Finally, we discussed the color-coding and documentation with the project partners and made recommendations to address the yellow highlighted areas, thereby answering the research questions from Section I. Through the documentation and recommendations, we were able to both raise awareness of bias in the respective projects and, with the help of the framework, provide a guide for dealing with and avoiding potential bias.

The one-pager for the smart animal health project revealed similar bias issues (and recommendations) with an additional yellow region for the input data sets (which are still too few). Some elements are not highlighted in green or yellow because not all areas and elements could be meaningfully applied due to the early project stage. Bohler [11] documents both one-pagers for the above-mentioned projects in detail.

B. *Framework Validation*

After framework application, the projects replied to a set of meta-questions ([11], p.37) about applicability (strength, weaknesses), usefulness (relevance, increased awareness for bias), size, comprehensibility, and coverage:

- *Applicability:* The framework is perceived as applicable to the project in question. Specific strengths are the sensitization to the topic of bias and the holistic perspective on AI projects nudging for reflection on the framework areas of interest and their set of elements. Applicability increases significantly when training in bias or ethical topics had taken place beforehand.
- *Usefulness:* The framework addresses a relevant problem in the context of AI projects, especially when decision-making algorithms are involved. It raises awareness of ethical issues in the design of AI systems. Bias mitigation can only be achieved after referencing identified bias in the project context. Moreover, the possible weighting of elements and the use of metrics could enable more targeted referencing and recommendations for improvement in many project settings.
- *Size:* The framework is considered extensive with its mixture of organizational, social, and technical aspects. Its application in an AI project is therefore complex and may require several people with appropriate expertise. Especially for smaller projects, a holistic application is challenging, since e.g., it cannot be assumed to establish a cross-functional and diverse project team. Its deployment can be reflected at most imaginary instead.
- *Coverage:* The scope was considered appropriate, as no elements were felt to be superfluous or missing.
- *Comprehensibility:* The areas and checklist elements are largely found to be understandable. Examples or sample comments could further increase comprehensibility and give orientation filling in the one-pager.

C. *Reflection on Framework Application*

During framework application and based on the feedback on the above meta-questions, it showed to be helpful to follow specific restrictions and to establish several preconditions:

- Before application, awareness of the bias problem in general and for a specific project must be created. The risks of bias potentials should be illustrated. Otherwise, there is a risk that project participants will underestimate the importance of the framework criteria or not consider them important.
- There should be an introduction to the framework in terms of goals, content, and procedures. In particular, it should be conveyed that the framework is to be understood as a guideline rather than a prescription. The strengths of the framework should be emphasized, in part because it is based on a sound literature review. It is beneficial to answer the questions as given.
- The framework can be suitably adapted for the application (e.g., textual formulations, specification of examples).
- The application should be accompanied and supported by trained persons. If this person is outside the project, a regular exchange between the trained persons and the project experts should be established.
- To establish a sustainable awareness effect, the bias assessment should be repeated after a reasonable period.
- The results of a bias assessment should be documented in the form of the one-pager, the summarized expert feedback (e.g., per area of interest), and the set of project-specific recommendations for bias mitigation. By highlighting the risks, awareness of the relevance of measures for mitigating bias is promoted. [11]

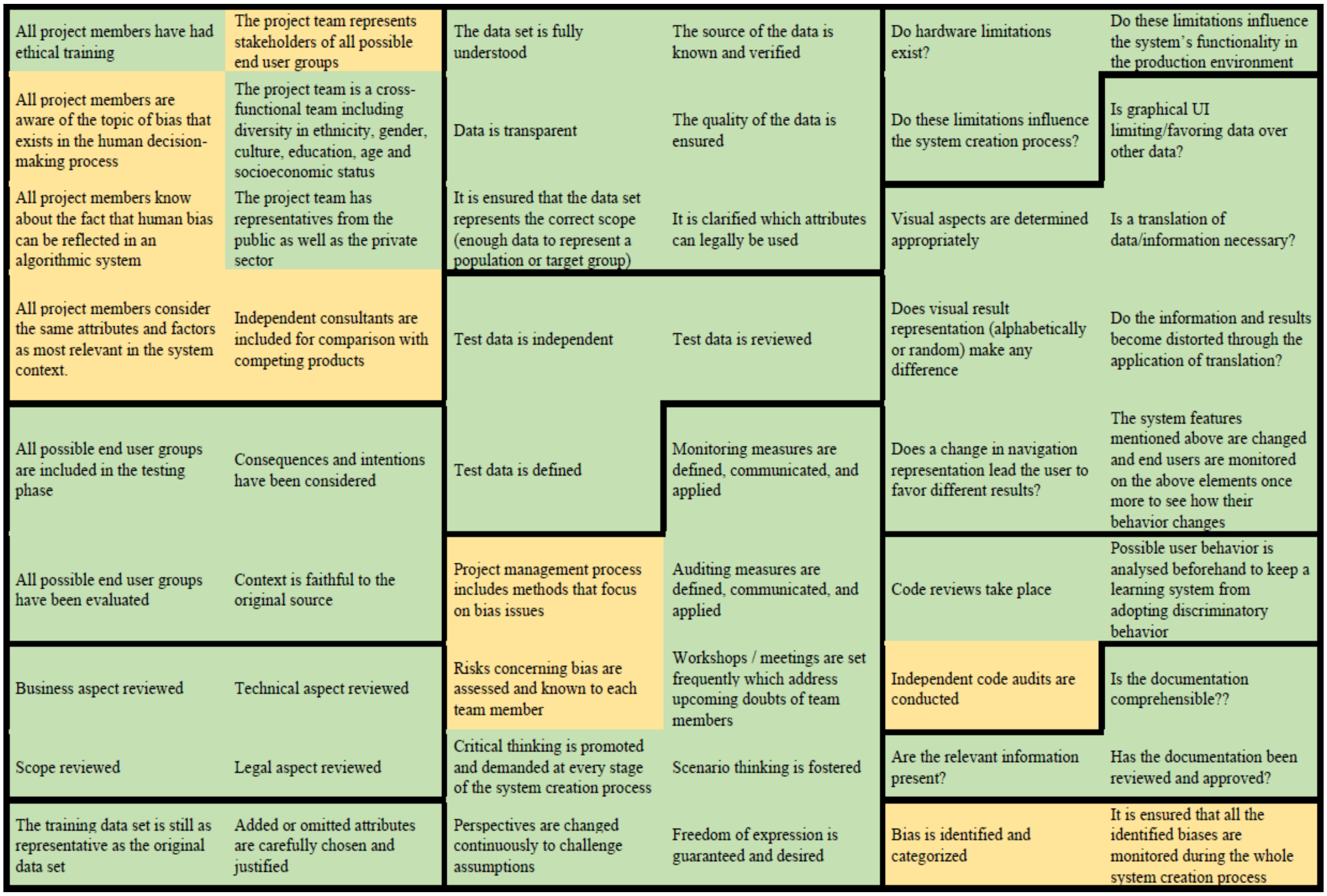
contains sample documentation for both above-mentioned projects.

- Identified bias potentials do not necessarily confirm the existence of bias. Nevertheless, bias potentials should be considered in-depth and appropriate measures taken. This prevents bias potentials from developing into bias.
- After completion of a bias assessment, users are to be surveyed regarding their application experience. The resulting insights can be used to improve the framework and to increase its practical validity.

D. *Framework Improvements*

Although the framework was successfully applied to practical AI projects, we could identify several aspects for framework improvement and its future application:

- To focus on particularly relevant areas of interest or checklist elements, a project-specific weighting scheme could be helpful, by which the checklists are instrumented in advance.
- Because the framework contains text-based checklists, there is always room for interpretation. Although the iterative bias assessment process can settle the scope for interpretation in certain cases, the introduction of metrics and the use of a glossary could reduce diffusivity.
- The presence of sample answers or examples for checklist criteria would, on the one hand, speed-up orientation. On the other hand, it could act as a time-saving cheat sheet that prevents users from creatively developing their ideas.
- To add more stability to the framework, more bias assessments should be performed in AI projects from various application domains. This could also identify, for example, areas where bias potential is particularly common (or rare) and thereby give valuable hints for future projects. Moreover, valuable indications of the general situation around bias in AI could be derived.
- From time to time, the framework should be reflected based on user feedback regarding the meta-questions and formulations within the checklists.
- With the rapid development of the AI field, new framework conditions and facts may arise about bias in AI. The framework should therefore be reviewed to ensure that it is up to date and adapted if necessary.



All project members have had ethical training
The project team represents stakeholders of all possible end user groups
All project members are aware of the topic of bias that exists in the human decision-making process
The project team is a cross-functional team including diversity in ethnicity, gender, culture, education, age and socioeconomic status
All project members know about the fact that human bias can be reflected in an algorithmic system
The project team has representatives from the public as well as the private sector
All project members consider the same attributes and factors as most relevant in the system context.
Independent consultants are included for comparison with competing products

All possible end user groups are included in the testing phase
Consequences and intentions have been considered
All possible end user groups have been evaluated
Context is faithful to the original source

Business aspect reviewed
Technical aspect reviewed
Scope reviewed
Legal aspect reviewed

The training data set is still as representative as the original data set
Added or omitted attributes are carefully chosen and justified

The data set is fully understood
The source of the data is known and verified
Data is transparent
The quality of the data is ensured
It is ensured that the data set represents the correct scope (enough data to represent a population or target group)
It is clarified which attributes can legally be used

Test data is independent
Test data is reviewed
Test data is defined

Monitoring measures are defined, communicated, and applied
Project management process includes methods that focus on bias issues
Auditing measures are defined, communicated, and applied
Risks concerning bias are assessed and known to each team member
Workshops / meetings are set frequently which address upcoming doubts of team members
Critical thinking is promoted and demanded at every stage of the system creation process
Scenario thinking is fostered
Perspectives are changed continuously to challenge assumptions
Freedom of expression is guaranteed and desired

Do hardware limitations exist?
Do these limitations influence the system's functionality in the production environment
Do these limitations influence the system creation process?

Is graphical UI limiting/favoring data over other data?
Visual aspects are determined appropriately
Is a translation of data/information necessary?
Does visual result representation (alphabetically or random) make any difference
Do the information and results become distorted through the application of translation?
Does a change in navigation representation lead the user to favor different results?
The system features mentioned above are changed and end users are monitored on the above elements once more to see how their behavior changes

Code reviews take place
Possible user behavior is analysed beforehand to keep a learning system from adopting discriminatory behavior

Independent code audits are conducted
Is the documentation comprehensible??
Are the relevant information present?
Has the documentation been reviewed and approved?

Bias is identified and categorized
It is ensured that all the identified biases are monitored during the whole system creation process

Figure 8. One-pager for the chatbot project presenting 12 areas of interest (rectangular shapes) containing its corresponding checklist elements. Green (resp. yellow) elements have successfully (resp. partially) been addressed. In black and white print, green means darker background and yellow means lighter background.

VII. Conclusion

Since currently there are only weak AI systems that lack self-awareness and depend on human advice in the shape of created models and selected training data, human bias is naturally and unintentionally reflected in crafted algorithmic systems. A framework has been proposed and validated, which helps to identify and mitigate bias in algorithmic systems, covering aspects of the complete life cycle of such software systems.

The framework was developed based on desk research, and validation was conducted through field research. The approach was implemented in realistic software project situations and its added value could be observed, evaluated, and validated. During validation, each area of interest of the metamodel was evaluated against the criteria and questions in the checklists, and user feedback was summarized and structured along with the areas of interest. Subsequent reflection on the bias assessments led to slight framework improvements.

As a future improvement, it would be useful to investigate to what extent automation of the framework use 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 project in question, could be supported by a software system. During the project, the checklists are continuously filled with data by the project team, thereby enabling process analysis, comparison of different framework implementations, and revealing indications if and where the recommendations were complied with.

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 the reliability of the results of an 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, several aspects need to be considered. Elements from the checklists would have to be detailed at the micro level to define, for example, what a stakeholder is or how it can be verified that the test user belongs to a specific gender. Instead of a yes/no checkmark in the checklists, there could be more detailed measures, e.g., an indication of the level to which a team member has received ethical training. Furthermore, the integration of mechanisms that consider the plausibility of the answers in the checklists would be helpful.

References

[1] T. Gasser, E. Klein, and L. Seppänen, "Bias – A Lurking Danger that Can Convert Algorithmic Systems into Discriminatory Entities," in *Centric2020 - The 13th Int. Conf. on Advances in Human-oriented and Personalized Mechanisms, Technologies, and Services*, 2020, pp. 1–7.

[2] Accenture AG, "AI is the new UI – Experience Above All," *Accenture Technology Vision*, 2017. [Online]. Available: https://www.accenture.com/_acnmedia/Accenture/next-gen-4/tech-vision-2017/pdf/Accenture-TV17-Full.pdf. [Accessed: 24-Nov-2021].

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

[4] 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.

[5] S. Feuerriegel, M. Dolata, and G. Schwabe, "Fair AI," *Bus. Inf. Syst. Eng.*, vol. 62, no. 4, pp. 379–384, Aug. 2020.

[6] D. Cossins, "Discriminating algorithms: 5 times AI showed prejudice," *New Scientist*, 2018. [Online]. Available: https://www.newscientist.com/article/2166207-discriminating-algorithms-5-times-ai-showed-prejudice/. [Accessed: 24-Nov-2021].

[7] J. Larson, S. Mattu, L. Kirchner, and J. Angwin, "How We Analyzed the COMPAS Recidivism Algorithm," *ProPublica*, 2016. [Online]. Available: https://www.propublica.org/article/how-we-analyzed-the-compas-recidivism-algorithm. [Accessed: 24-Nov-2021].

[8] E. Hunt, "Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter," *The Guardian*, 24-Mar-2016. [Online]. Available: https://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter. [Accessed: 24-Nov-2021].

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

[10] T. Gasser, "Bias – A lurking danger that can convert algorithmic systems into discriminatory entities," 2019. [Online]. Available: https://www.theseus.fi/bitstream/handle/10024/167429/Gasser_Thea.pdf. [Accessed: 24-Nov-2021].

[11] R. Bohler, "Validation of the Framework for Bias Identification and Mitigation," BUAS - Bern University of Applied Sciences, Master Thesis (in German), 2021.

[12] H. Snyder, "Literature review as a research methodology: An overview and guidelines," *J. Bus. Res.*, vol. 104, pp. 333–339, Nov. 2019.

[13] L. Harold, M. Turoff, and O. Helmer, *The Delphi Method - Techniques and Applications*. Addison-Wesley, 1975.

[14] S. Holder, "What is AI, really? And what does it

mean to my business?," 2018. [Online]. Available: https://www.sas.com/en_ca/insights/articles/analytics/local/what-is-artificial-intelligence-business.html. [Accessed: 24-Nov-2021].

[15] A. Goel, "How Does Siri Work? The Science Behind Siri," *Magoosh Data Science Blog*, 2018. [Online]. Available: https://magoosh.com/data-science/siri-work-science-behind-siri/. [Accessed: 24-Nov-2021].

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

[17] E. Alpaydın, *Introduction to Machine Learning*, 4th ed. MIT Press, 2020.

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

[19] A. Smith, "Franken-algorithms: the deadly consequences of unpredictable code," *The Guardian*, 2018. [Online]. Available: https://www.theguardian.com/technology/2018/aug/29/coding-algorithms-frankenalgos-program-danger. [Accessed: 24-Nov-2021].

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

[21] 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.

[22] 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.

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

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

[25] AlgorithmWatch, "AI Ethics Guidelines Global Inventory," 2019. [Online]. Available: https://algorithmwatch.org/en/project/ai-ethics-guidelines-global-inventory/. [Accessed: 24-Nov-2021].

[26] 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.

[27] 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.

[28] N. Turner Lee, P. Resnick, and G. Barton, "Algorithmic bias detection and mitigation: Best practices and policies to reduce consumer harms," 22-May-2019. [Online]. Available: https://www.brookings.edu/research/algorithmic-bias-detection-and-mitigation-best-practices-and-policies-to-reduce-consumer-harms/. [Accessed: 24-Nov-2021].

[29] SAS, "Organizations Are Gearing Up for More Ethical and Responsible Use of Artificial Intelligence, Finds Study," 2018. [Online]. Available: https://www.sas.com/en_id/news/press-releases/2018/september/artificial-intelligence-survey-ax-san-diego.html. [Accessed: 24-Nov-2021].

[30] A. Raymond, "The Dilemma of Private Justice Systems: Big Data Sources, the Cloud and Predictive Analytics," 2014.

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

[32] E. Isele, "The Human Factor Is Essential to Eliminating Bias in Artificial Intelligence," *Chatham House*, 2018. [Online]. Available: https://www.chathamhouse.org/expert/comment/human-factor-essential-eliminating-bias-artificial-intelligence. [Accessed: 24-Nov-2021].

[33] S. Beck, A. Grunwald, K. Jacob, and T. Matzner, "Künstliche Intelligenz und Diskriminierung," 2019. [Online]. Available: https://www.plattform-lernende-systeme.de/publikationen-details/kuenstliche-intelligenz-und-diskriminierung-herausforderungen-und-loesungsansaetze.html. [Accessed: 24-Nov-2021].

[34] K. R. Varshney, "Introducing AI Fairness 360, A Step Towards Trusted AI," *IBM Research Blog*, 2018. [Online]. Available: https://www.ibm.com/blogs/research/2018/09/ai-fairness-360/. [Accessed: 24-Nov-2021].

[35] 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.

[36] E. Adeli *et al.*, "Representation Learning with Statistical Independence to Mitigate Bias," in *2021 IEEE Winter Conference on Applications of Computer Vision (WACV)*, 2021, pp. 2512–2522.

[37] S. Barocas and A. D. Selbst, "Big Data's Disparate Impact," *SSRN Electron. J.*, vol. 104, pp. 671–732, 2016.

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

[39] Sozialversicherungsanstalt des Kantons St. Gallen, "Der IPV-Chatbot," 2019. [Online]. Available: https://www.svasg.ch/news/meldungen/ipv-chatbot.php. [Accessed: 24-Nov-2021].

[40] Swiss Federal Food Safety and Veterinary Office, “Research project «Smart Animal Health»,” 2020. [Online]. Available: https://www.blv.admin.ch/blv/en/home/tiere/forschungsprojekte-tiere/forschungsprojekt-smart-animal-health.html. [Accessed: 24-Nov-2021].

[41] K. LLoyd, “Bias Amplification in Artificial Intelligence Systems,” in *AAAI FSS-18: Artificial Intelligence in Government and Public Sector*, 2018.