Proceedings of the 55th Hawaii International Conference on System Sciences | 2022

A Value Sensitive Design Perspective on AI Biases

Isabel Gan
Researcher
isabelgan2001@gmail.com

Sara Moussawi
Carnegie Mellon University
smoussaw@andrew.cmu.edu

Abstract

Artificial Intelligence (AI) technology has made profound impacts in our society but concerns about AI biases are rising. This paper classifies AI-related biases and proposes strategies to tackle them. To inform our study, we review AI research on human values to identify three categories of AI biases: pre-existing, technical, and emergent. Informed by the value sensitive design (VSD) framework, we then map the AI biases to the three phases (conceptual, empirical, and technical) of VSD investigation. Our analysis shows that both conceptual and empirical investigations are helpful for addressing pre-existing bias, technical investigation for technical bias, and both technical and empirical investigations for emerging bias. The paper highlights that to effectively tackle AI-related biases, it is important for AI developers and the user community to understand human values in an AI context and to advocate for developing AI-specific value-oriented standards that are agreed upon and adopted by all stakeholders.

1. Introduction

Artificial intelligence is concerned with developing systems and algorithms capable of performing tasks that typically require human intelligence [1]. As Artificial Intelligence (AI) becomes more prevalent in our society, the profound impact of AI systems across various industries and societal domains has given rise to a debate surrounding the values and principles that should guide the development and use of AI systems [2, 3]. Values refer to what a person or group of people consider important in life [4]. According to Friedman et al [4], examples of universal human values include privacy, universal usability, trust, freedom from bias, and autonomy. Meanwhile, principles are abstract overarching actionable statements and directives that can be used to guide design and development of technologies. Principles can be value sensitive. The ethical implications related to AI development and use have not received wide attention, leading to many concerns surrounding these technologies and the biases in human decision making they can impose. Such ethical concerns are particularly relevant given the increasing adoption and use of AI in organizational decision making [3].

AI-related biases are destructive to users, including organizations, groups, and individuals. A technology is biased if it unfairly and / or systematically discriminates against certain individuals by denying them an opportunity or assigning them a different and undesirable outcome [5]. In her best-seller book “Weapons of math destruction: How big data increases inequality and threatens democracy,” O’Neil [3] refers to those math-powered applications as “Weapons of Math Destruction” (WMDs). Created with the best of intentions, “many of these models [actually] encoded human prejudice, misunderstanding, and bias into the software systems that increasingly managed our lives” (p.3). The author illustrates AI algorithms as WMDs and shows how they have become an integral part of organizational managerial and strategic decision making. In the end, the book demonstrates the importance of ethical considerations of AI system by illustrating how the destructive characteristics of those algorithms brought harm to employees and negatively affected work morals and productivity.

One major issue surrounding AI ethics is the extent of privacy sourced from users. The access of personal data of over 50 million of users given to Cambridge from Facebook is one such example where ethics play a deciding factor in the amount of user privacy that developers can breach and use to program AI [6]. Computer engineers and data scientists have or will encounter ethical issues such as biases in the building of machine learning models and stereotypes in the development of robots sourced from such data. Therefore, how to avoid the risk of constructing machine intelligence that mirrors a narrow and privileged vision of society has become an urgent and important question. If we put too much importance on the decision-making capabilities of AI, the lack of transparency becomes a severe problem.

Research has paid increasing attention to the issues surrounding AI ethics. While some research

URI: https://hdl.handle.net/10125/80014
978-0-9981331-5-7

focuses on the technical aspects of AI such as reshaping data processing and analysis [7], other research pays attention to organizational and societal impacts of AI such as the use and effects of AI applications in a variety of sectors including healthcare, transportation, and the production chain [8]. Recent work highlights the need to address emerging AI challenges through responsible development of AI and the need for participatory engagement in the creation of documents addressing the ethical implications of AI across organizations in all sectors [9, 10, 11]. Mapping system design onto principles for social good has also become the focus of recent explorations [12].

Although prior research offers useful insights into understanding AI development and impacts, we would argue that, to minimize the biases built in AI systems or emerging during the deployment of AI, we need to first understand the root causes of those biases and consider a systematic approach to cope with them. In this way, increasing dependence on AI will not cause unethical behaviors, such as discrimination against certain groups in favor of others [5]. Therefore, the objective of the paper is to classify AI-related biases and recommend strategies to minimize the biases. In particular, we aim to answer two research questions:

(1) What are the biases surrounding the development and use of AI?

(2) What are potential strategies to minimize the AI-related biases?

To inform our study, we review AI research on human values associated with AI. We build on value sensitive design (VSD) [5], a commonly used framework in the human-computer interaction field, to make recommendations on coping with the different types of AI biases. We then explore some AI ethics cases reported in the media and industry reports and illustrate how VSD can be applied to prevent those biases from occurring in the future. AI ethics and principles are fields with a wealth of research. This paper does not intend to provide a comprehensive review of AI ethics research. Rather, the main objective of this paper is to illustrate different types of AI ethics cases and discuss strategies and considerations for AI design from human value perspective.

2. Theoretical background

2.1. Value sensitive design

Value Sensitive Design (VSD) is a theoretically grounded approach to the design of technology that accounts for human values in a principled and comprehensive manner throughout the design process [13]. “Value” is used as a broad term in VSD research; it refers to what a user considers important in life [4]. An example of human value is informed consent by users in website design. When a website collects personal identifying information of its users, users appreciate being informed about the website’s collection of such information as well as being asked for user consent [8]. Privacy, trust, freedom and autonomy are all referred to as human values in the VSD literature [1, 4]. Friedman and Kahn [14] have proposed a classification of values, referred to as a collection of 12 human values with ethical import, including human welfare; ownership and property; privacy; freedom from bias; universal usability; trust; autonomy; informed consent; accountability; identity; calmness; and environmental sustainability. Friedman et al. [13] indicate that this set of values is open to refinement. In other words, the collection of these human values should not be taken as fixed and universal. Rather, human values may be contextualized. For example, Dadgar and Joshi [15] study the role of information and communication technology in patient self-management and reveal twelve values shared by diabetes patients, including accessibility, accountability, autonomy, compliance, dignity, empathy, feedback, hope, joy, privacy, sense-making, and trust in the healthcare setting.

VSD is a design that prioritizes human values using a three-part methodology including conceptual, empirical, and technical investigations [13,14]. These investigations are applied iteratively. Conceptual investigations focus on understanding the various stakeholders of the technology, clarifying the stakeholders’ needs, and articulating their values and any values conflicts that might arise through the use of the technology. Empirical investigations rely on conducting design research studies to inform the technology designers' understanding of the users’ values and needs. During empirical investigations, the focus is to understand human responses to the actual technological products as well as the larger context of technological use. Technical investigations can involve either retrospective analysis - how people use existing, related technologies - or proactive system design - the design of systems to support stakeholder values identified in the conceptual and empirical investigations. Of the three investigations, empirical investigations focus on human responses to the technology itself [13], which is most relevant to our research objective in this study. Therefore, in this study we focus on the empirical investigation stage in the VSD paradigm, by understanding the responses of individuals and groups that are involved in or affected by AI systems. We also focus on identifying the biases in AI use to aid our empirical investigation.

To design and implement technologies for ethical use by organizations and individuals, the VSD framework emphasizes the process of conducting conceptual investigations of key values, implicating the values in technology design, and integrating value considerations in organizational structures [13]. However, research suggests that a practical use of the VSD approach starts from identifying a value, technology, and the context of use [13]. For developers to design or refine technical systems to account for human values, it is important to first identify the human values and any potential value conflicts. This is echoed by Le Dantec et al. [16], which highlight the importance of "inquiring about the values present in a given context and responding to those values - being sensitive to those values - through design" [p.1143].

Like web-based systems or organizational information systems studied in Friedman [13], human values should be implicated in AI systems as the technologies are designed to accomplish a goal (task) by applying different technical approaches [17]. Consistent with this suggestion, we review AI research to identify the key principles that have been considered by the AI development community. Such knowledge of AI principles will help us interpret the rising user concerns and AI biases in organizational use of AI.

2.2. AI values

Based on the content analysis of 84 sources on AI guidelines collected on a global scale, Jobin, Ienca, & Vayena [1] identify 11 overarching principles that are sensitive to ethical values in AI design, development, and use, including transparency, justice and fairness, non-maleficence, responsibility, privacy, beneficence, freedom and autonomy, trust, dignity, sustainability, and solidarity (p. 395). Among the values, five values, namely transparency, justice, non-maleficence, responsibility and privacy, emerged as the most promoted across stakeholders as they were referenced in more than half of all guidelines in the comprehensive document review by Jobin et al. [1].

In Table 1, the values are listed by descending order of importance, according to Jobin et al. [1]. Similarly, the definitions of the values below are modified from Jobin et al. [1] unless specified. Next, we discuss the AI values in more detail.

Transparency: It is the most popular value prevalent in the current literature of AI development, but thematic analysis of AI guidelines reveals significant variation in relation to the interpretation, justification, and domain of application. To achieve greater transparency, many sources suggest increased disclosure of information by those developing or deploying AI systems.

Fairness: Literature about this topic is extensive, particularly concerning the ethical need to understand the historical and social contexts into which these systems are being deployed [19]. If AI systems do not incorporate the value of fairness, the use of AI will likely cause discrimination among users. Fairness is one of four values highlighted in the context of medical use of AI [20].

Table 1. Values promoted in AI guidelines

Values	**Definitions**
(1) Transparency	Transparency refers to the efforts to increase explainability, interpretability or other acts of communication and disclosure.
(2) Justice and Fairness	Justice is expressed mainly in terms of fairness, and mitigation of unwanted bias and discrimination. Some sources focus on justice as referring to equality, inclusion, while others call for the right to redress and remedy.
(3) Non-Maleficence	Non-maleficence encompasses general calls for safety and security. It highlights the fact that AI should never cause foreseeable or unintentional harm. Harm is generally interpreted as discrimination, violation of privacy, or bodily harm.
(4) Responsibility	Responsibility is normally tied to acting with integrity and legal liability; it also focuses on the underlying reasons that may lead to potential harm.
(5) Privacy	Privacy refers to the right someone has to keep their personal life or personal information secret or known only to a small group of people.
(6) Beneficence	Beneficence means the augmentation of human senses, promotion of human well-being/flourishing, as well as economic prosperity and happiness.
(7) Freedom and Autonomy	Some specifically refer to the freedom of expression and self-determination while others refer to promoting freedom in a general sense. Autonomy is positive freedom, referred to as freedom to flourish.
(8) Trust	Trust is referred to as trustworthy AI research, technology, developers, organizations, and design principles. Trust also underlines the importance of customers' trust.
(9) Dignity	Dignity is referred to as avoiding harm, forced acceptance, automated classification, and unknown human-AI interaction.
(10) Sustainability	Sustainability refers to the development and deployment of AI to protect the environment, improve the planet's ecosystem and biodiversity, and creating more equal societies.
(11) Solidarity	Solidarity refers to sharing the prosperity created by AI and assessing the long-term implications before developing and deploying AI systems [18].

Non-maleficence: This value is related to safety and security concerns and calls for the avoidance of specific risks or potential harms, such as intentional misuse via cyberwarfare and malicious hacking.

Responsibility: It focuses on clarifying the attribution of responsibility and legal liability. Actors that are considered being responsible and accountable for AI actions and decisions include developers, designer, institutions, and industry.

Privacy: This value focuses on protecting personal information and personal life. However, it cannot be taken for granted since AI is largely developed by the private sector for deployment in public (i.e., criminal sentencing) and private contexts (i.e., insurance) [20]. It is one of the human values classified in the collection of 12 human values in the VSD literature [14].

Beneficence: It is often associated with promoting social benefits for human well-being. Private companies often tend to highlight the value of AI for customers.

Freedom and Autonomy: This value refers to freedom of expression, freedom to flourish, and self-determination [1]. The development of technical systems considers this an important human value in the VSD literature [14].

Trust is applied to a wide range of stakeholders in the AI community, from technology and developers to organizations and customers [1]. In research on human-machine collaboration, trust is also found to be an important success factor [21]. It is one of the human values highlighted in the VSD literature [14].

Dignity is intertwined with human rights: AI should not destroy but preserve/increase human rights.

Sustainability: It focuses on protecting our environment and ecosystems. To achieve this value, AI design, development and management should be concerned with energy efficiency [1]. This human value is emphasized in the technology design in the VSD literature [14].

Solidarity: Incorporating the value of solidarity means implementing mechanisms to redistribute the augmentation of productivity for all and sharing the burdens, making sure that AI does not increase inequality, and nobody is left behind [18]. According to Jobin, Ienca, & Vayena [1], just 6 out of 84 AI principles' guidelines mention the value of solidarity. However, other research present different views. For example, Luengo [18] finds that solidarity is one of the fundamental values at the heart of a peaceful society, present in more than 30% of world's constitutions and a foundational principle of institutions such as the European Charter.

Not all the values are highlighted in a use context. For example, in the context of medical use of AI, research highlights four principles that resemble the principles of medical ethics: respect for human autonomy; prevention of harm; fairness, and explicability - that are endorsed by the Organization for Economic Co-operation and Development (OECD) and European Commission's High-Level Expert Group on Artificial Intelligence [20].

3. Method

In this study, we searched relatively recent published books and articles (2016-2020) on AI ethics and performed content analysis. Then we applied the VSD framework to classify the categories of AI biases (i.e., [5]) and used this classification to recommend the type of investigation to account for human values [8, 9] in AI development and use.

To collect data on cases of AI biases, we used a snowball sampling approach. First, we collected all the cases of AI biases summarized in the book by O'Neil [3]. Then we searched news media (i.e., *New York Times)* and published academic journal articles in Google Scholar for the time period (2016-2020) for additional cases of AI biases in organizational decision making. Our search key terms included "AI ethics" and "AI bias." As a result of this search process, we collected a total 15 cases of AI-related biases (see Table 2).

To inform our coding of the AI-related cases, we refer to Friedman [5], which defines three categories of bias, to guide our classification of biases of AI. According to Friedman [5], the first type of bias is *preexisting bias*. Pre-existing bias refers to bias that exists before the creation of technology. This sort of bias is rooted in social institutions and attitudes and reflected in personal biases. This bias emerges in technology implicitly or explicitly, consciously or unconsciously by individuals and institutions tasked with designing the technology designing the technology [5] Preexisting biases are the most frequent biases revealed in the AI literature. The second type of bias is *technical bias;* it pertains to issues in the technical design of a product, such as technical constraints or decisions. The last type of bias is *emergent bias*. Unlike preexisting and technical bias, emergent bias occurs during the real use of a design and happens because of a change in societal knowledge or cultural values. We present the three types of biases with examples from the AI literature in Table 2.

VSD presents a theoretical and methodological framework that allows for integrating values into the design process of technology. VSD uses an iterative tripartite methodology that integrates conceptual, empirical and technical investigations [4, 5]. The analysis of the cases of AI biases with the VSD framework allows us to provide recommendations to minimize these biases. Through our exploration, we also discuss how the values and principles identified in Table 1 have been undermined or promoted by the creation and use of AI.

Building on the analysis of three categories of biases and their roots, we evaluated how the roots of biases in each case can be traced to the three investigations of the VSD framework. For example, the AI application of IMPACT is an assessment tool used by school administrators to measure teacher performance and to make firing decisions (refer to Table 2). The AI bias in this case is rooted in the algorithm (the technical design) that did not promote the fairness value of the users (employees). To prevent this bias rooted in technical design, we found the technical investigation of the VSD framework relevant (refer to Table 4). Thus, this analysis led us to the mapping between technical bias (i.e., IMPACT tool) and the technical investigation of VSD. Further analysis of all the AI biases revealed four interrelated themes, which are presented in Tables 3-5 in sub-section 4.2.

4. Findings

4.1. Categorization of biases

In this paper, we refer to “bias” as “unfair,” “unwanted,” or “undesirable” systematic discrimination against certain individuals or groups of individuals based on the inappropriate use of certain traits or characteristics such as disabilities, race, gender, and sexual orientation, consistent with Wilson and Daughtery [22]. Table 2 lists the examples of the three categories of AI biases.

Preexisting biases are the most frequent biases revealed in the AI literature. Examples *of **preexisting biases*** can be found in real life and in AI stemming from the problem of racism. Take the Microsoft AI chatter bot for example: it learned from a “wrong” model and began to post inflammatory and offensive tweets through its Twitter account, causing Microsoft to shut down the service only 16 hours after its launch [23]. Another example is the criminal justice models that were found biased due to the oversampling of certain neighborhoods: such neighborhoods are overpoliced so the oversampling can result in higher rates of recorded crime, which results in more policing. Bias can also be introduced into the data in how it is collected or selected for use.

The majority of preexisting bias in the AI field can be linked to the understanding of certain concepts such as fairness, solidarity, and transparency. Because these ideas encompass such a large field, it is extremely hard for developers to agree on one definition for each category. This leads to preexisting bias within AI when system designers and developers have personal biases and have significant input into the design of the technology [5]. One origin of preexisting bias is gender bias. For example, software designers (who are mostly male) were found sometimes unknowingly designing software that is more aligned with males than with females [5].

Table 2. Classification of major AI biases

Bias	Examples and sources
Preexisting Bias	(1) Microsoft bot learned from a “wrong” model [23]. (2) Murder of Duane Buck (racism) [3] (3) Stance on concepts like fairness [3] (4) Criminal justice models (sampling certain neighborhoods over others) [19] (5) Understanding of concepts (i.e., perception of farness & solidarity) [18] (6) Reasonability of CO2 emissions (personal thinking) [18]
Technical bias	(7) IMPACT: an assessment tool to measure teacher performance for firing decisions [3] (8) Tech is opaque and only open to developers [3] (9) LSI-R: a questionnaire for prisoners [3] (10) Kyle’s Job Application (the technology/personality test was biased against him) [3] (11) Hiring Algorithm: Women’s colleges (development of tech was discontinued) [19]
Emergent bias	(12) PredPol (not really biased off data, the data skews the analysis) [3] (13) St. George’s model (gender bias for female applicants) [3] (14) COMPAS (incorrectly labeled African American defendants as “high-risk”) [19] (15) Facial Analysis Tech (ignoring discrepancies in experience of technology by users of different race and gender) [19]

Technical bias is associated with the design of the technology. For example, a technology company discontinued development of a hiring algorithm based on analyzing previous decisions after discovering that the algorithm penalized applicants from women’s colleges [19]. The problem occurs within the grasp of the technology itself, and so serves as a piece of technical bias. IMPACT, an assessment tool for teacher, and LSI-R, a questionnaire for prisoners, are both cases in which the technology simply does not take into account all the external factors it should, resulting in a feedback loop and incorrect analysis [3]. For example, in the case of IMPACT model, in 2007, the mayor of Washington DC, Adrian Fenty, thought students were not learning well enough (low graduation rates), and decided to get rid of the low-performing teachers. The assessment model called IMPACT, which was not well-developed, was created to measure students' academic improvement by paying attention to their scores but failed to account for the efforts made by teachers. A teacher, Sarah Wysocki, was known to be an excellent instructor but failed IMPACT, and so was fired [3].

Another issue with certain technology is that it may be biased towards a group of people without them knowing. For example, in the case of the Hiring Algorithm model, the development of a hiring algorithm was based on analyzing previous decisions; the algorithm penalized applicants from women's colleges. The technology company discovered the bias and discontinued the development of the algorithm [19].

This is a problem highly relevant to those with disabilities as well. This bias is evidenced in the case of Kyle's Personality Test [3]. Kyle was a smart kid who attended Vanderbilt yet was not able to even secure a minimum wage job. He was bipolar and always failed the job-hiring personality tests. However, there was no way to challenge this process [3]. Many companies opted to use personality tests even when the Court ruled intelligence tests discriminatory in 1971. The feedback loop in this case is that red-lighting people with certain mental health issues prevents them from having a normal job and life, further isolating them, which goes against the Americans with Disabilities Act.

Preexisting and technical biases occur before and during the creation of a technology. Because it is the developers who mainly create and refine a technology product, this stakeholder group is mostly associated with preexisting and technical biases. However, emergent bias emerges during the real use of the technology (after it is developed). Therefore, emergent bias is mainly associated with users who test the technology product. As we can see, different types of biases are associated with different stakeholders, in this case the developers and users of a technology.

Emergent bias is often the most obvious type of bias. Many examples of emergent bias are linked to categories such as race and gender, such as in the examples of COMPAS and facial analysis technologies. COMPAS is a model used to predict recidivism in Broward County, Florida. It incorrectly labeled African American defendants as "high-risk" at nearly twice the rate it mislabeled white defendants [19]. In the example of Facial Analysis Tech, the development of the technology ignores "harms of representation," meaning discrepancies in how different groups experience technology. Error rates in facial analysis technologies were found differing by race and gender [19]. The St. George's Model is another example of emerging bias: Initially created to boost efficiency and establish fairness, the technology eventually learned how to discriminate from inputs, leading to the rejection of female applicants with the justification that their careers would be disrupted because of motherly duties [3]. Many examples of emergent bias, such as St. George's Model, pertain to categories that fixate on race and gender. This is because these are societal values that are constantly changing. There is no doubt that the way society views issues such as system racism have changed dramatically even over the last decade.

Among the three types of biases, not only does preexisting bias tend to be overlooked, but its importance is not acknowledged as well. Preexisting bias is linked to very vague and conceptual ideals, which differ from developer to developer. As a result, it is the hardest type of bias to fix. Preexisting biases may originate from society at large, in subcultures, or in formal or informal organizations and institutions [5]. It stems all the way down to the biased personal views of some developers. Furthermore, preexisting bias is the very first type of bias identified in the process of developing a product, and can serve as the root or base for the other two types of bias; this means many forms of technical and emergent bias may arise because of the effect that preexisting bias has on the technology. Relating to the field of AI, a great amount of issues happen to revolve around AI ethical principles such as transparency, fairness, and solidarity. These principles are essentially the foundation to all of AI development and are a major reason why preexisting bias plays a prominent role among the three types.

4.2. Strategies for minimizing AI biases

We classify solutions under three categories, depending on the type of investigations (referring to the VSD framework) in which we believe would help address different problems. These three types of investigations include conceptual, empirical, and technical investigations. A *conceptual* investigation is concerned primarily with breaking down or analyzing needs and values of various stakeholders involving in the technology design and use [13,14]. An *empirical* investigation is used to evaluate the success of a technological design, such as human responses to the actual technological products. It often involves observation and documentation; quantitative and qualitative methods used in research are applicable [13]. Finally, a *technical* investigation focuses on the technology itself rather than the people or social systems affected by the technology. It focuses on how properties of technologies and underlying mechanics support or hinder human values, as well as involves proactive design of systems to support values identified in conceptual investigations [13]. Tables 3-5 summarize the AI biases and relevant types of investigations.

First, strategies for reducing a pre-existing bias include both a conceptual and empirical investigation, as presented in Table 3. Conceptual investigations can often be used to target preexisting biases such as the understanding of concepts as well as personal biases of individuals. One solution to address this bias is to implement fairness constraints on the optimization process itself or use an adversary to minimize the system's ability to predict the sensitive attribute [19]. Another solution would be to create a unified framework and definitions to be used by everyone for these concepts [20]. However, this is not highly feasible as there are simply too many people to please. As an alternative, a VSD implementation could be put forth. While VSD is not an ideal solution, it presents fundamentals for determining common values across stakeholder groups and makes value conflicts functionally apparent and addressable [24].

Table 3. Preexisting bias and investigation strategy

Bias in AI	**Recommendations**
Understanding of concepts / developers from different backgrounds (ideology) [20] [a]	-To implement fairness constraints on the optimization process itself or use an adversary to minimize the system's ability to predict the sensitive attribute [18] - To create a unified framework and definitions used by everyone for these concepts [19] -VSD is not an ideal solution but presents fundamentals for determining common values across stakeholder groups and makes value conflicts functionally apparent and addressable [24] Committees and similar groups like the UK Select Committee can acknowledge a common set of values amongst select stakeholders, extend conceptual and empirical investigations to other groups not considered and determine overlaps [24]
Criminal Justice models [18] [b]	pre-processing the data to maintain as much accuracy as possible while reducing any relationship between outcomes and protected characteristics, or to produce representations of the data that do not contain info about sensitive attributes, this includes "counterfactual fairness" approaches and post-processing techniques [18]

[a] merits a conceptual investigation
[b] merits an empirical investigation

Another main issue relating to AI is transparency and privacy concerns. Technology is opaque and often only open to developers. To address this, we need to shift our perspective and focus on the privacy concerns of users [22]. We could look at the four principles that resemble those of medical ethics for transparency: respect for human autonomy, prevention of harm, fairness, and explicability. Lastly, we could maybe consider moving towards the European model as it has more user input. There are many issues of bias existing within the field of AI, but these are some approaches under the three types of investigations that could be taken to help address them.

Second, technical biases should most often be fixed in technical investigations. Problems with the technology itself should be changed in this step to address certain issues. In cases more complex such as discrimination against applicants, we can use empirical investigations to see how the discrimination occurs as well as execute pre-processing and post-processing training techniques to minimize feedback loops [18]. During empirical investigations, we may also implement innovative training techniques like transfer learning or decoupled classifiers for different groups [18]. Another solution may be to have more people managing and training the technology [22].

Table 4. Technical bias and investigation strategy

Bias in AI	**Recommendations**
Tech is opaque and only open to developers [3] [a]	-Needs to address more privacy concerns of users [22] -To address 4 principles (transparency) that resemble those of medical ethics [19]: 1) Respect for human autonomy; 2) Prevention of harm; 3) Fairness; 4) Explicability -To consider moving toward European model – more user input [3]
Women's colleges [18] [c]	-Consider changing the hiring algorithm in itself by using empirical investigation to see how it discriminates based on applicant (pre & post-processing [18] -Have more people managing & training tech [16]
IMPACT [3] [c]	-Outputs must be audited for fairness, -To have the model designed and tested carefully by humans before its deployment for automating managerial decisions [3]
LSI-R [3] [c]	-To manage the feedback loops on main issues [3] -To Adopt post-processing techniques [18]
Kyle's Personality Test [3] [c]	-To change tech to accommodate to special conditions like bipolar & feedback loop. -To consider remedying [5]. For example, handedness problem, allowing to toggle between both configurations; Archimedes Project at Stanford are developing approach to designing for ppl w disabilities

[a] merits a conceptual investigation
[c] merits a technical investigation

From these solutions discussed above, we can see that these types of bias can be solved by overlapping the different types of investigations. In the context of technical biases relating to conceptual terms, we must train the device accurately to interpret this type of data. Furthermore, outputs must be designed and carefully tested by humans before allowing AI to automatically make decisions [3]. Remedying during empirical investigations is another much needed step for accommodating special conditions and those with disabilities to prevent discrimination against certain groups of people [5].

Finally, strategies for reducing an emerging bias include both an empirical and technical investigation, as presented in Table 5. An empirical investigation is used to evaluate the success of a technical design and involves observation and documentation from the application to activities [13]. The last type of investigation, technical investigation, focuses of the technology itself rather than the people or social systems affected by the technology. It focuses on how properties of technologies and underlying mechanics support or hinder human values, as well as involves proactive design of systems to support values identified in conceptual investigations [13].

Table 5. Emergent bias and investigation strategy

Bias in AI	Recommendations
COMPAS [18] [b]	-Use empirical investigation to see how it discriminates based on applicant (pre & post-processing [18]) OR have more ppl managing & training tech [22]
Facial analysis [18] [b]	-Adopt innovative training techniques like transfer learning or decoupled classifiers for different groups [18] – use techniques after conducting empirical investigation of results
PredPol [3] [c]	-Need to change tech: Somehow manage the feedback loops – main issue [3] -Adopt post-processing techniques [18]
St. George's Model [3] [c]	-Consider changing the hiring algorithm in itself by using empirical investigation to see how it discriminates based on applicant (pre and post-processing) [18] -Have more people managing & training tech [22]

[b] merits an empirical investigation
[c] merits a technical investigation

5. Discussion

In this paper, we have attempted to map the three categories (pre-existing, technical, and emergent) of AI biases to the three phases (conceptual, empirical, and technical) of the investigation in the VSD framework. However, effective practices in minimizing AI biases require three key considerations.

First, to achieve an effective strategy, it is important for AI developers and the user community to understand the human values in the context of AI use and to advocate for developing AI-specific value-oriented metrics and standards that are agreed on and adopted by all stakeholders. For example, defining and measuring fairness is expected but not an easy endeavor, so different metrics and standards will likely be required, depending on the case and circumstances [13]. One recommendation to minimize bias and to realize the fairness value is to pre-process the data to maintain as much accuracy as possible while reducing any relationship between outcomes and protected characteristics, or to produce representations of the data that do not contain information about sensitive attributes [13]. In another example, to achieve greater transparency, many sources suggest increased disclosure of information by those developing or deploying AI systems, although specifications regarding what should be communicated vary greatly: use of AI, source code, data use, evidence base for AI use, limitations, laws, responsibility for AI, investments in AI and possible impact [1].

Second, humans need to reevaluate the relationship between humans and machines. Humans need to perform three crucial roles. They must train machines to perform certain tasks; explain the outcomes of those tasks, especially when the results are counterintuitive or controversial; and sustain the responsible use by machines, for example, preventing robots from harming humans [22]. With increasing adoption of AI in a variety of fields ranging from healthcare to supply chain, humans find themselves interacting with AI systems or "robots" more. As AI systems increasingly reach conclusions through processes that are opaque (the so-called black-box problem), they require human experts in the field to explain their behavior to non-expert users. These "explainers" are particularly important in evidence-based industries, such as law and medicine, where a practitioner needs to understand how an AI weighed inputs into, say, a sentencing or medical recommendation [22]. This is easier said than done. The explainers may not be sufficient or sustainable. Perhaps, making the technology more communicative and transparent is the solution to communicate to the users what it (the technology/ AI) is doing and how. Customizations also come into play, which can vary by user level and background.

Finally, addressing AI biases and promoting ethical AI requires concerted effort from multiple stakeholders and communities. The increasing concerns and debates on AI ethics have attracted the attention from both private and public sectors across the globe. As shown in Jobin et al. [1], national and international organizations have responded to these concerns by developing ad hoc expert committees on AI, often mandated to draft policy documents. These committees include the High-Level Expert Group on Artificial Intelligence appointed by the European Commission, the expert group on AI in Society of the Organization for Economic Co-operation and Development, the Advisory Council on the Ethical Use of Artificial Intelligence and Data in Singapore, and the Select Committee on Artificial Intelligence of the UK House of Lords.

Our findings on the mapping between the VSD investigations and AI biases also offer implications for Design Science Research. The patterns revealed in our data analysis through a VSD lens help prioritize the types of AI biases and the focus on a particular investigation. For example, empirical investigation in VSD is found being associated with both pre-existing and emergent biases, highlighting the importance of accounting for the values of the stakeholders during the empirical investigation. For example, in the case of the COMPAS model that incorrectly labeled African American defendants as “high-risk” [18], we can conduct design studies to see how the AI system discriminates applications based on applicants in order to inform AI designers and developers’ understanding of potential users’ values.

6. Conclusion

Our paper has made an initial attempt to map the three types of AI bias to the three phases of VSD investigations in an effort to recommend effective strategies in minimizing AI biases in AI development and use. Fifteen AI-related cases were used for illustration. One limitation of the study is the sample size. Although the 15 cases were collected from multiple sources with rich details, future studies will provide additional insights by expanding the data set of real-life AI-related biases. For example, one promising study is to extend the collection period from 5 years (i.e., 2016-2020) to 10 years (i.e., 2010- 2020) to increase the data sample of AI cases in future research.

One big remaining challenge in human battle against AI-related biases is the lack of a unified regulatory framework for AI on a global scale, a framework that establishes clear fiduciary duties towards data subjects and users. Should such a framework emerge from AI Ethics, a principled approach could be deemed successful. But without a strong regulation that establishes fiduciary duties for the key interests of data subjects and users, we cannot conclude that a comparable degree of value alignment exists for AI [20]. This challenge is closely related to characteristics of the AI field, including AI developers, AI development method, and accountability mechanisms. According to Mittelstadt [20], coming from varied disciplines and professional backgrounds, AI developers bring with them incongruous histories, cultures, incentive structures, and moral obligations. Under such circumstances, it would be an oversimplification to reduce the field to a single vocation or type of expertise. Moreover, AI development does not have comparable empirically proven methods to translate principles into practice in real-world development contexts. Neither does it have comparable professionally or legally endorsed accountability mechanisms, excluding certain types of risks such as privacy violations [1]. Therefore, the AI field has an urgent need for additional regulations to promote human values and protect fairness and human rights in the development and use of AI.

7. References

[1] A. Jobin, E. Vayena, and M. Ienca, “The global landscape of AI ethics guidelines,” *Nature Machine Intelligence*, 1(9), 2019, pp. 389-399.

[2] E. Awad, S. Dsouza, R. Kim, J. Schulz, J. Henrich, A. Shariff, J-F. Bonnefon, and I. Rahwan, “The Moral Machine experiment,” *Nature*, 563, 2018, pp. 59–64.

[3] C. O’Neil, *Weapons of math destruction: How big data increases inequality and threatens democracy*, Crown, New York, 2016.

[4] B. Friedman, P. H. Kahn, A. Borning, and A. Huldtgren, *Value sensitive design and information systems. In Early engagement and new technologies: Opening up the laboratory* (pp. 55-95). Springer, Dordrecht, 2013.

[5] B. Friedman, “Value-sensitive design”, *Interactions*, 3(6), 1996, pp. 16-23.

[6] R. Blackman, “A Practical Guide to Building Ethical AI”, *Harvard Business Review*, 2020. Retrieved on 6/12/2021: https://hbr.org/2020/10/a-practical-guide-to-building-ethical-ai

[7] M. I. Jordan, and T. M. Mitchell, “Machine learning: trends, perspectives, and prospects”, *Science*, 349, 2015, pp. 255–260.

[8] W. W. Stead, “Clinical implications and challenges of artificial intelligence and deep learning”, *JAMA*, 320, 2018, pp. 1107–1108.

[9] J. Borenstein, and A. Howard, “Emerging challenges in AI and the need for AI ethics education,” *AI and Ethics*, 1(1), 2021, pp. 61-65.

[10] M. Hickok, “Lessons learned from AI ethics principles for future actions,” *AI and Ethics*, 1(1), 2021, pp. 41-47.

[11] D. Schiff, J. Borenstein, J. Biddle, and K. Laas, “AI ethics in the public, private, and NGO sectors: a review of a global document collection,” *IEEE Transactions on Technology and Society*, 2021.

[12] S. Umbrello, and I. van de Poel, “Mapping value sensitive design onto AI for social good principles,” *AI and Ethics*, 2021, pp. 1-14.

[13] A. Borning, B. Friedman, and P. H. Kahn Jr., "Value sensitive design and information systems”, *The Handbook of Information and Computer Ethics*, Wiley (Editors: K. Himma and H. Tavani), New Jersey, 2008, pp. 69-102.

[14] B. Friedman, and P. H. Kahn Jr. “Human values, ethics, and design,” in *The Human–Computer Interaction Handbook: Fundamentals, Evolving Technologies and Emerging Applications*, J. A. Jacko and A. Sears (eds.), Mahwah, NJ: Lawrence Erlbaum Associates, Inc., 2003, pp. 1177-1201.

[15] M. Dadgar, and K. D. Joshi, “The role of information and communication technology in self-management of chronic diseases: an empirical investigation through value sensitive design,” *Journal of the Association for Information Systems*, 19(2), 2018, pp. 86-112.

[16] C. A. Le Dantec, E. S. Poole, and S. P. Wyche, “Values as lived experience: Evolving value sensitive design in support of value discovery,” in *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems*, 2009, New York: ACM, pp. 1141-1150.

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

[18] M. Luengo-Oroz, “Solidarity should be a core ethical principle of AI,” *Nature Machine Intelligence*, 1(11), 2019, pp. 494-494.

[19] J. Manyika, and J. Silberg, “Notes from the AI frontier: Tackling bias in AI (and in humans),” *McKinsey Global Institute*, 2019.

[20] B. Mittelstadt, “Principles alone cannot guarantee ethical AI,” *Nature Machine Intelligence*, 2019, pp. 1-7.

[21] X. Deng, “Artificial Intelligence and Human-Robot Teaming: Challenges and Design Considerationsk,” *Knowledge Management, Innovation* (Editor: M. Jennex). IGI Global, Hershey, Pennsylvania, USA, 2020.

[22] J. H. Wilson, and P. R. Daughtery, “Collaborative Intelligence: Humans and AI are Joining Forces,” *Harvard Business Review*, July-August 2018.

[23] D. Victor, “Microsoft Created a Twitter Bot to Learn From Users. It Quickly Became a Racist Jerk,” *New York Times*, March 24, 2016.

[24] S. Umbrello, “Beneficial artificial intelligence coordination by means of a value sensitive design approach,” *Big Data and Cognitive Computing*, 3(1), 2019, p. 5.