Fairness in design : a tool for guidance for ethical artificial intelligence design

Shu, Ying

2021

Shu, Y. (2021). Fairness in design : a tool for guidance for ethical artificial intelligence design. Master's thesis, Nanyang Technological University, Singapore.
<https://hdl.handle.net/10356/154153>

<https://hdl.handle.net/10356/154153>

<https://doi.org/10.32657/10356/154153>

This work is licensed under a Creative Commons Attribution-NonCommercial 4.0 International License (CC BY-NC 4.0).

Downloaded on 26 Apr 2023 17:28:06 SGT

FAIRNESS IN DESIGN: A TOOL FOR GUIDANCE FOR ETHICAL ARTIFICIAL INTELLIGENCE DESIGN

Ying SHU

School of Computer Science and Engineering

A thesis submitted to the Nanyang Technological University
in partial fulfillment of the requirements for the degree of
Master of Engineering

2021

Statement of Originality

I hereby certify that the work embodied in this thesis is the result of original research, is free of plagiarised materials, and has not been submitted for a higher degree to any other University or Institution.

13/08/ 2021

Date

Ying SHU

Supervisor Declaration Statement

I have reviewed the content and presentation style of this thesis and declare it is free of plagiarism and of sufficient grammatical clarity to be examined. To the best of my knowledge, the research and writing are those of the candidate except as acknowledged in the Author Attribution Statement. I confirm that the investigations were conducted in accord with the ethics policies and integrity standards of Nanyang Technological University and that the research data are presented honestly and without prejudice.

13/08/2021

.....

Date

NTU NTU
NTU NT
NTU NTU
NTU NTU

J NTU
TU NT
U NT
J NTU

.....

Prof. Han YU

Authorship Attribution Statement

This thesis contains material from one paper published in the following peer-reviewed journal(s) / from papers accepted at conferences in which I am listed as an author.

Part of Chapter 1, Chapter 2 and Chapter 3 is published as

Shu, Y., Zhang, J. and Yu, H. (2021) Fairness in Design: A Tool for Guidance in Ethical Artificial Intelligence Design. In: Meiselwitz G. (eds) Social Computing and Social Media: Experience Design and Social Network Analysis. HCII 2021. Lecture Notes in Computer Science, vol 12774. Springer, Cham. https://doi.org/10.1007/978-3-030-77626-8_34

The contributions of the co-authors are as follows:

- A/Prof Han Yu provided the initial project direction and edited the manuscript drafts.
- I designed the cards and prepared the manuscript drafts.
- Jiehuang Zhang helped design the organisation of the paper and edited the final manuscript.

.....13/08/2021.....
Date

.....
Ying SHU

Acknowledgements

I wish to express my greatest gratitude to my supervisor Prof. Yu Han for providing me guidance, feedback, and support for my Masters study. I would like to thank him for introducing me to the field of ethical artificial intelligence. I am beyond grateful for his advice for my research and I appreciate the time he has given me to make my work better. I could not have imagined having a better supervisor for my study.

I would also like to thank all my collaborators and lab-mates for the support and help they have given me in this research process.

Lastly, I would like to thank my family, Songko, and friends for being there for me throughout this entire process. Their support and encouragement push me to be better in my research.

Abstract

As artificial intelligence (AI) becomes increasingly widely applied, societies have recognized the need for proper governance for its responsible usage. An important dimension of responsible AI is fairness. AI systems were once thought to be impartial and fair in their decisions, but studies have shown that biases and discrimination are able to creep into the data and model to affect outcomes even causing harm. Due to the multi-faceted nature of the notions of fairness, it is challenging for AI solution designers to envision potential fairness issues at the design stage. Furthermore, there are currently limited methodologies available for them to incorporate fairness values into their designs.

In this thesis, we present the Fairness in Design (FID) methodology and tool that aim to address the gap. It is available in both physical and online format. The tool provides AI solution designers with a workflow that allows them to surface fairness concerns, navigate complex ethical choices around fairness, and overcome blind spots and team biases. We have tested the methodology on 10 AI design teams ($n = 24$) and the results are supportive of our hypotheses. Not only 67% of the participants would recommend our physical methodology tool to their friend or colleague, but also 79% of the participants indicated that they are interested in using the tool in their future projects. This tool has the potential to add value to the ethical AI field and can be expanded to support other ethical AI dimensions such as privacy preservation and explainability.

Contents

Acknowledgements	ix
Abstract	xi
List of Figures	xvii
List of Tables	xix
Abbreviations	xxi
1 Introduction	1
1.1 Background	1
1.1.1 Current Influence of Artificial Intelligence on Society	1
1.1.2 Design Focus of Artificial Intelligence Systems	2
1.1.3 Repercussion of Artificial Intelligence on Society	2
1.2 Motivation	3
1.2.1 Need for Human Values in Artificial Intelligence Design	3
1.2.2 Importance of Fairness Value	3
1.3 Objectives	4
1.4 Major Contributions	4
1.5 Thesis Organisation	5
2 Literature Review	7
2.1 Overview of Literature Review	7
2.2 Ethical Artificial Intelligence	7
2.3 Value Sensitive Design	9
2.4 Fairness in Artificial Intelligence	10
2.4.1 Fairness Principles	11
2.4.2 Tools to Design Fair Systems	12
2.5 Conclusion	13
3 Fairness in Design Cards (FDC)	15
3.1 Framework and Overview	15
3.2 Defining Components	16

3.2.1	Application Types	16
3.2.2	Direct and Indirect Stakeholders	16
3.2.3	Fairness Principles	17
3.3	Implementation of Fairness in Design Cards	17
3.3.1	Step 1: Write the Application Scenario	17
3.3.2	Step 2: Choose an Application Card	18
3.3.3	Step 3: Identify the Stakeholders	19
3.3.4	Step 4: Draw a Fairness Card	20
3.3.5	Step 5: Write a Reflection	21
3.3.6	Step 6: Go to Step 3 OR Collect and Discuss OR Write Review	21
3.4	Conclusion	23
4	FAIR Application	25
4.1	Framework and Overview	25
4.2	Design of FAIR Application	26
4.2.1	Web Functionality Planning	26
4.2.2	User Flow	28
4.2.3	Information Architecture	29
4.3	Implementation of FAIR Application	30
4.3.1	Facilitator Account	30
4.3.2	Member Account	33
4.3.3	Both Facilitator and Member Account	37
4.4	Conclusion	38
5	User Study	39
5.1	Participants	39
5.2	Design of User Study and Questionnaires	40
5.3	User Study Procedure	42
5.4	Analysis of Results	43
5.4.1	Hypotheses Testing	43
5.4.2	Attitude towards Fairness in AI/ML	44
5.4.3	Feedback of Fairness in Design Workshop	45
5.5	Conclusion	46
6	Conclusion and Future Works	47
6.1	Conclusion	47
6.2	Future Work	48
A	Application Type Cards	49
B	Fairness Cards	51
B.1	Group Fairness Cards	51
B.2	Individual Fairness Cards	55

C Questionnaires	59
C.1 Pre-Study Questionnaire	60
C.2 Post-Study Questionnaire	71
List of Author's Publications	77
Bibliography	79

List of Figures

3.1	A visual guide of Fairness in Design.	18
3.2	Application card for industrial and commercial uses.	19
3.3	Direct and indirect stakeholder cards.	19
3.4	An example of a group fairness card (front and back).	20
3.5	An example of an individual fairness card (front and back).	21
3.6	Reflection card.	22
3.7	Review card.	23
4.1	User flow for FAIR Application.	28
4.2	Information Architecture for FAIR Application.	29
4.3	Main workflow in FAIR Application.	30
4.4	Facilitator Dashboard.	31
4.5	Application Scenario Page.	31
4.6	Application Type Page.	31
4.7	Invitation Page.	32
4.8	Settings for each project.	32
4.9	Member Dashboard.	33
4.10	In Progress Tab.	34
4.11	An example of a project background that members will first read.	34
4.12	An example of a stakeholder role.	34
4.13	An example of a randomised fairness card shown to members.	35
4.14	An example of a reflection done on a fairness card.	36
4.15	An example of a fairness review.	36
4.16	Results tab that is available in both facilitator and member account.	37
4.17	Results page of a project showing the computed fairness ranking.	38
5.1	Survey demographics of the application domains (left) and technology areas (right).	40
A.1	Life-Critical Systems.	49
A.2	Industrial and Commercial Uses.	49
A.3	Office, Home, Entertainment Applications.	50
A.4	Exploratory, Creative, Collaborative Applications.	50
A.5	Sociotechnical Systems.	50
B.1	Demographic Parity/ Statistical Parity (Front).	51

B.2 Demographic Parity/ Statistical Parity (Back).	51
B.3 Conditional Statistical Parity (Front).	52
B.4 Conditional Statistical Parity (Back).	52
B.5 Equalized Odds (Front).	52
B.6 Equalized Odds (Back).	53
B.7 Equal Opportunity (Front).	53
B.8 Equal Opportunity (Back).	53
B.9 Treatment Equality (Front).	54
B.10 Treatment Equality (Back).	54
B.11 Test Fairness (Front).	54
B.12 Test Fairness (Back).	55
B.13 Fairness Through Awareness (Front).	55
B.14 Fairness Through Awareness (Back).	55
B.15 Fairness Through Unawareness (Front).	56
B.16 Fairness Through Unawareness (Back).	56
B.17 Counterfactual Fairness (Front).	56
B.18 Counterfactual Fairness (Back).	57
B.19 Fairness in Relational Domain (Front).	57
B.20 Fairness in Relational Domain (Back).	57

List of Tables

5.1	Participant demographics: self-reported team role and participant ID.	41
-----	---	----

Abbreviations

AI	Artificial Intelligence
VSD	Value Sensitive Design
FID	Fairness In Design
FDC	Fairness in Design Cards

Chapter 1

Introduction

This chapter first presents the background information about the current influence of artificial intelligence (AI) on society. Motivation and objectives of the studies will be presented in Section 1.2 and Section 1.3 respectively. Section 1.4 highlights the contribution of the studies. Lastly, the organisation of the thesis is presented in Section 1.5.

1.1 Background

1.1.1 Current Influence of Artificial Intelligence on Society

With the speed of current breakthroughs, artificial intelligence has evolved to become a crucial backbone of society and a disruptor to almost every industry [1]. The influence is so huge that it is included as part of a technological revolution that will fundamentally change the way we live, work, and connect [2]. It is used to provide recommendation and personalisation, assist in prediction and decision making, and so much more. Due to its ability to learn and analyze a large amount of data in a short time, it is preferred as the more efficient and viable alternative to manual processing. More algorithm tools are also being used to supplement or replace human decision-making in many aspects of life [3], even in important and life-changing areas such as medicine and law. It is undeniable that with this level of power, there should be greater responsibility placed upon these AI systems.

1.1.2 Design Focus of Artificial Intelligence Systems

For so long, key design metrics of AI systems were reliability, efficiency, and accuracy. The speed and capacity of these machines were the reason behind the rapid adoption and we aim to train them to be faster and better. Interestingly, as it starts to impact us more, we see a gradual change in the design process of it as well. Initially, focus was purely on the implementation and technical aspect of software systems [4]. Design process involves separating the project into different components that perform specific functions in the product itself. This was useful as the main audience were experts in specialized environments. When software systems proliferated and became more accessible to general users, the need to ensure that they are easy and efficient to use led to the focus on human computer interaction[5]. This caused a change in the design process whereby human factors are included in the product life cycle. Users are now able to be engaged and have their opinions heard throughout the design stages as opposed to just being presented with the finished product.

1.1.3 Repercussion of Artificial Intelligence on Society

However, the focus on these metrics above all else overlooked that these AI systems are not as impartial and reliable as we thought them to be. Even though they are machines that follow logic, biases and discrimination can creep into the data and models [6][7] to affect outcomes to the point of causing harm [8] [9]. The most notable example would be the Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) software used in state courts systems in the U.S. whereby the model predicted twice as many false positives for recidivism for African-Americans offenders than white Americans [10]. In 2019, a healthcare risk-prediction algorithm that was used on more than 200 million people in the U.S. was found to show racial bias as it was less likely to identify eligible black people for high-risk care management due to a faulty metric [11].

1.2 Motivation

1.2.1 Need for Human Values in Artificial Intelligence Design

Due to such serious negative consequences of algorithm bias, it is of utmost priority to develop responsible and ethical AI. This pushes us to re-evaluate the design process for creating such systems. One way is to start integrating human values early in the design process, especially when AI systems are capable of embodying political, social, and ethical values. Since product teams train AI with the data they choose, they have the power to ensure that machines hold up end-user values with a human-centric focus [12]. Therefore, methodologies for product teams to incorporate such values, as well as to avoid and minimize the negative impact of overlooking these values become crucial [13]. Mere awareness from product teams about the importance of human values in AI design is not enough. There needs to be tried and tested tools to help them identify the human values critical to the system, engage such values in their consideration, and discover any potential fatal problems, especially at the earliest stages of the design process when the specifications and prototypes are still fluid [14].

1.2.2 Importance of Fairness Value

The ethical value of fairness was chosen to be the focal point as the lack of consideration for it showed us the danger of algorithm bias. Furthermore, it is a complex human value that has different meanings for different people. In the machine learning literature, there are already more than twenty definitions and metrics for fairness [7][15]. Besides, there are also fundamental statistical incompatibilities between the fairness metrics such that there is no solution that can satisfy all metrics [16][17]. Thus, product teams are forced to consider the trade-offs between these metrics when considering fairness for their AI systems. It is therefore understandable that product teams will find it daunting to grasp all these in a short frame of time during their design process. A survey of industry practitioners in machine learning also indicated that they face problems in foreseeing trade-offs between

fairness definitions, anticipating and overcoming blind spots, and auditing when designing fairer AI systems [18].

1.3 Objectives

The main goal of the studies is to lower the barrier for AI practitioners to include fairness in their system design discussion in a methodological way through the following objectives.

- Objective 1: To enable practitioners to navigate ethical choices in fairness.
- Objective 2: To help practitioners surface fairness concerns.
- Objective 3: To highlight different stakeholder perspectives.
- Objective 4: Methodology is easy to learn and engaging to use without the need for any prior knowledge.

1.4 Major Contributions

The contributions of this thesis are stated as follows:

- In Chapter 3, I designed a Fairness in Design methodology in the form of playing cards for AI practitioners to incorporate fairness values into their system design. This methodology allows them to focus on the fairness aspects of their design and guide them in surfacing fairness issues from different perspectives.
- In Chapter 4, I designed a project management and collaboration platform that applies the Fairness in Design methodology to fairness projects. It enables users to hold fairness discussion online. In addition, details about the discussion and results are stored in the platform for easy reference.
- In Chapter 5, I conducted a user study to evaluate the usefulness of the Fairness in Design methodology in achieving its aims. The results were analysed.

1.5 Thesis Organisation

The thesis is organised as follows:

Chapter 1 provides the background of the influence of artificial intelligence on society, the motivation behind the research, the objectives of the research, and highlights the major contributions of the thesis.

Chapter 2 presents the literature review of the overall ethical artificial intelligence field, value sensitive design theory and methods, and a focus on fairness in artificial intelligence.

Chapter 3 introduces the Fairness in Design Cards and the Fairness in Design methodology. Components of the methodology are defined and the procedure is explained and illustrated step by step.

Chapter 4 introduces the FAIR application. The design process of the application is shared in three parts, (i) web functionality planning, (ii) user flow, and (iii) information architecture. The implementation of the application is presented through the users' perspectives.

Chapter 5 evaluates the Fairness in Design Methodology through a user study. The procedure and results are described and discussed.

Chapter 6 concludes the thesis and presents potential future research directions.

Chapter 2

Literature Review

2.1 Overview of Literature Review

In this chapter, the review is divided into three different areas. Section 2.2 details the current efforts in the ethical artificial intelligence field. Section 2.3 describes the various methods under Value Sensitive Design that have influenced the work in this thesis. Section 2.4 examines the interpretation of fairness in the artificial intelligence field and elaborates on the existing methods that have been created to develop fairer AI systems. Section 2.5 concludes this chapter.

2.2 Ethical Artificial Intelligence

As more incidents of algorithmic bias and harm surfaced, international organisations and private companies have started to announce their commitment in building responsible AI systems. This is especially so when AI systems are capable of embodying political, social, and ethical values. These organisations have published high-level tenets and proposals to guide AI practitioners in developing and implementing ethical AI systems. Global efforts include the Partnership on AI [19], Organisation for Economic Co-operation and Development [20], European Union’s High-Level Expert Group [21] and G20 [22]. These collaborations aim to research best practices in ethical AI, provide policy recommendations, and highlight ethical,

legal and societal issues. Private companies such as Microsoft [23] and Google [24] have also outlined specific principles to guide their own project teams.

With various guidelines by different organisations, a survey was done on 84 unique sets of AI ethics principles to analyse if there were any similarities and it was found that the AI principles were mostly focused on specific concepts such as accountability, transparency, privacy, and fairness [25]. In fact, these concepts were already established under professional and business ethics in other industries [26][27][28] and therefore, it is reasonable for them to be the guiding values for AI systems too. Out of the AI ethics principles, the leading value common to the codes analysed was fairness [28].

Despite efforts by organisations in outlining AI ethics, these concepts tend to be abstract and high level. Difficulty in operationalising the principles leaves them open for interpretation and inference [29]. Practitioners can have differing opinions about how the values should be implemented in the design of the systems leading to ineffectiveness and disputes. As a result, these guidelines and proposals are not suitable for actual AI system designs. This is exacerbated by the lack of guidance in resolving trade offs when stakeholder needs and expectations conflict, especially when not only practitioners but also other stakeholders come into play when evaluating AI ethics in the system [28].

Qualitative research on fairness design needs have exposed the disconnect between the focus of the researchers in the field and the needs of AI practitioners [30][18]. Practitioners have mentioned that the responsibility of achieving an ethical AI system often falls on them solely due to the multitude of mathematical solutions available to make the system fairer [30]. However, their own experience indicated that discovering challenges and solutions from sociocultural perspectives are effective in creating a fairer system as well. It is not surprising that AI ethics comprise of both sociocultural and technical aspects. Due to the emphasis on mathematical models, these social practices should be encoded with the mathematical solutions to provide a holistic viewpoint of AI such that they can be maintained and transferred across time, contexts and personnel. Furthermore, the creation of responsible AI systems should involve the entire product teams and not just the AI practitioners.

2.3 Value Sensitive Design

With the emphasis on building ethical AI systems, the ability to develop techniques for practitioners to incorporate ethical values, as well as to avoid and minimize the negative impact of overlooking these values become crucial. There have been various design methodologies in the literature that strive to achieve these aims. Methodologies allow researchers, designers, and engineers to follow a special process to probe and analyse their issues [13]. This gives them proper direction to address and include AI ethics principles in their research and discussion.

Value sensitive design (VSD) is the most famous theoretical approach in value design of technology that focuses on integrating human values in a principled and methodical way throughout the design process [31]. It encompasses a large literature of targeted methods to engage with values differently. Qualities of VSD methods include being committed to the theoretical constructs of VSD, being in descriptive form to help designers gain more insights, and staying open to integration with other methods, adaptation and changes [32]. At the end of this section, readers will have a better understanding of the methods that have influenced the work.

Direct and Indirect Stakeholder Analysis allows the identification and legitimization of stakeholders, as well as explores how they might be impacted, along with the potential benefits, harms, and tensions [13][33][34]. It is not only about individuals and groups but also about institutions and societies. Designers and engineers tend to focus on direct stakeholders, who are those that interact directly with the technology and often miss out on indirect stakeholders, those that do not interact directly with the technology but are still affected [35][36][37][38].

Co-evolution of Technology and Social Structure allows social structures to be integrated into the design space to yield new two-pronged solutions [39][40]. This is because designers often view technical design in an isolated space. However, in reality, it is always intertwined with social structures. Social structures can include policy, law, regulations, organizational practice, social norms.

Value Scenarios are used as narratives that include important humanistic and societal considerations of technology. The aim is to emphasise the implications

for both direct and indirect stakeholders, related key values, widespread use, indirect impacts, longer-term use, and systemic effects [13][41][36][37][42][43]. These scenarios can function for both value elicitation and value representation.

Value Dams and Flows are used when value conflicts happen often. It serves as an analytical method to reduce number of solutions and dissolve value tensions among design choices [40][37][44]. Value dams refer to solutions that even a small percentage of stakeholders strongly object to and so will be removed from design choices. Value flows refer to solutions that a good percentage of stakeholders find appealing and therefore are the focus in the design space.

Envisioning Cards enable designers to think about the long-term and systemic issues in technology design [45]. It consists of 32 physical cards and are built on four criteria: stakeholders, time, values, and pervasiveness. On one side is a title and an evocative image related to card theme. The other side has the envisioning criterion, card theme, and a focused design activity. It can be used for ideation, co-design, heuristic critique, evaluation, and other purposes.

Judgement Call is a game that allows the design team to identify ethical concerns in a specific system via self-generated feedback [46]. It is in the form of physical cards with four categories: the ethical value, the stakeholder, the number of stars, and wild cards. The team will choose a scenario to work with, and subsequently brainstorm on the stakeholders involved. Each player will draw a stakeholder card, a value card, and the number of stars card. They will write a fictional review of the system according to what they have drew from the deck. Everyone will share their reviews with the team so common topics of concerns can be identified. It is an engaging and creative process to think about the impact of values in technology.

2.4 Fairness in Artificial Intelligence

Despite the common belief that AI is logical and unbiased, multiple incidents involving unfair decisions have shown the harm these systems can cause [47][48]. Researchers have identified biases in the data used and unchecked algorithms as the two potential sources of unfairness in machine learning systems [7]. This shows how hidden unfairness can be in the systems and the need to incorporate fairness checks in the AI design process.

However, fairness is an abstract value that is hard to understand and implement. The concept of fairness itself has been a topic of interest for philosophers and psychologists for centuries. There are many different school of thought in regards to what actually constitute fair. The fact that there is no universal definition of fairness signifies the difficulty of a common understanding of the concept [49]. This is inevitable as different cultures and countries have their own interpretation and preference of fairness. As algorithmic fairness derives from moral philosophy, the same situation occurs whereby there are many definitions operationalised to be used in system design.

2.4.1 Fairness Principles

In algorithmic fairness, the aim of fairness is to prevent any prejudice or favoritism toward an individual or group based on their hereditary characteristics [50]. There have been considerable efforts in the field of AI to determine how to operationalize fairness in a way algorithms can understand [51][52][15][28]. As the philosophical theories of fairness are discussed in words, the attempt is to construct similar concepts at a mathematical level.

So far, there exists more than twenty principles of fairness in the literature [15][53][7]. Researchers have found ways to categorised the fairness principles. In one classification, two different categories have been distinguished for fairness definitions: individual and group fairness [54][55]. Individual fairness requires similarly situated individuals to be treated similarly [54]. It is fulfilled when the algorithm gives similar outputs to individuals who have similar attributes that determine the results. Group fairness on the other hand requires different groups to be treated equally [54]. It is fulfilled when the distribution of outcomes is the same for each group.

Another way of classification is according to notions of distributive justice in social science. Fairness definitions are separated into four categories according to their goals: to achieve parity in treatment, to achieve parity in impact, to satisfy preferred treatment, and to satisfy preferred impact [56].

The last way of classification is to proximate the fairness principles to normative principles of fair distribution. This means that the fairness principles are categorised according to if they are consequentialist or deontological in outcome [53]. A consequentialist approach will evaluate the effects on the parties after the distribution whereas the deontological approach will evaluate the distribution based on rights.

2.4.2 Tools to Design Fair Systems

Despite the number of fairness principles available to make an AI system fairer, there is no instruction as to which metrics should be used in each condition. Moreover, some fairness principles are similar to each other and it is hard for anyone with no prior knowledge to grasp. The difficulty is further amplified when there is no one solution that can fulfil all the metrics [16][17]. Researchers have been creating checklists and toolkits to address these issues.

Google has created a What-If tool that allows practitioners to input their data-set into the tool [57]. They can then analyse and make changes to their machine learning model to visualise the impact on accuracy and the end users. Similarly, Aequitas is a bias and fairness audit toolkit that helps to look for biased actions or effects that are based on untrue assumption and stereotype about the affected population sub-groups in your data [58]. However, these toolkits require the actual data-set to undergo fairness analysis, which is often only available at the later part of the design process.

Fairness checklists have also been created to track the progress of task completion and ensure fairness is included in the objectives [59]. They are usually separated into sections in terms of the product life-cycle and contain binary questions answered in "yes" or "no". DrivenData created a data science ethics checklist to remind developers about ethical actions and this checklist can be included as a command line tool to data science projects [60]. Microsoft Research co-designed a AI fairness checklist with AI practitioners that contains actionable steps and tasks under each development stage [59]. Although these fairness checklists provide clear questions and tasks to follow, AI practitioners using such checklists have to spend time and effort to modify the checklists to fit their organisation contexts.

Furthermore, checklists often provide a false sense of security that completing the checklists mean a fairer system is created.

2.5 Conclusion

In this chapter, the current gaps in ethical artificial intelligence are presented. The theory of VSD and methods that fall under it are also explained in details. Lastly, classification of fairness principles and the existing tools to design fair AI systems are elaborated.

Chapter 3

Fairness in Design Cards (FDC)

This chapter introduces Fairness in Design Cards (FDC)¹, a physical methodological tool that aims to help product teams be more aware of fairness criteria for systems, and also enable them to surface potential fairness concerns for their application scenario by stimulating perspective taking from different stakeholders. Section 3.1 presents the framework and overview of the Fairness in Design Cards. Section 3.2 describes how certain components of the cards were defined. The implementation and workflow of the cards are demonstrated in Section 3.3. Lastly, Section 3.4 concludes this work.

3.1 Framework and Overview

Fairness in Design methodology is in the form of physical cards. Cards are used as the medium to facilitate the methodology as they are tangible, easy to learn and use, and inexpensive to produce. In addition, cards can be easily customisable with the flexibility to add or change the cards in the deck. The size of cards also ensure that the content on it is succinct and easy to digest.

The methodology consists of five to six steps. It can be played in groups of minimum two and only requires pens or pencils. The cards can be used by a diverse

¹The work in this chapter has been published in Shu, Y., Zhang, J. and Yu, H. (2021) Fairness in Design: A Tool for Guidance in Ethical Artificial Intelligence Design. In: Meiselwitz G. (eds) Social Computing and Social Media: Experience Design and Social Network Analysis. HCII 2021. Lecture Notes in Computer Science, vol 12774. Springer, Cham. https://doi.org/10.1007/978-3-030-77626-8_34

range of users with varying expertise in AI from industry product teams to university students. It can also be used for different application areas involving AI systems. Users do not need any prior knowledge of fairness before using this tool.

3.2 Defining Components

Three components in the methodology are defined by literature.

3.2.1 Application Types

Application types refer to the sector the application is used for. It is separated into five categories, 1) life-critical systems, 2) industrial and commercial uses, 3) office, home, and entertainment, 4) exploratory, creative, collaborative applications, and 5) sociotechnical applications. They are adapted from Shneiderman's classification for usability motivation in the literature of Human-Computer Interaction (HCI)[61]. It is a simple yet powerful classification system to differentiate the needs and characteristics of different sectors.

3.2.2 Direct and Indirect Stakeholders

We adopted direct and indirect stakeholder analysis [13] from VSD in our methodology to enable the teams to identify the stakeholders for their application scenario. Stakeholders are not only individuals and groups but can also be institutions and societies. We used the definitions described in Judgment Call [46] as our definitions. Direct stakeholders refer to those who interact directly with the technology and can include end-users, designers, engineers, hackers, and administrators whereas indirect stakeholders do not interact with the technology but are affected by its use and can include advocacy groups, families of end-users, regulators, and society at large [46].

3.2.3 Fairness Principles

As mentioned in Section 2.4.1, there are different interpretations to fairness. We chose to focus on mathematical operations of fairness as they are the most usable and understandable to AI practitioners. From all the mathematical operations available in the field, we streamlined our fairness principles to a total of ten fairness metrics as they are the most widely used definitions identified in the machine learning literature [7]. These ten metrics are separated into individual fairness (fairness through awareness[54], fairness through unawareness[62], and counterfactual fairness[55]) and group fairness (demographic/statistical parity[15], conditional statistical parity [63], equalized odds [64], equal opportunity[64], test fairness[16], treatment equality[65], and fairness in relational domain[66]).

3.3 Implementation of Fairness in Design Cards

This section will outline the steps to use the Fairness in Design Cards and it is illustrated in Figure 3.1. I will also elaborate more about the cards used in each step below.

3.3.1 Step 1: Write the Application Scenario

The first step involves the whole team identifying the application scenario they will use the tool for. It can be the actual product that they are working on or a fictional product. To have the most optimal effect, members should also be specific and elaborate on the component of the product that they want to assess fairness on. This will help to aid the discussion later by making the fairness criteria more relatable to the application scenario. As the tool can be used in any stage of the design process, product teams may choose to focus on an area that they have found or suspected to have fairness issues. This allows them to think about fairness concerns in a more open environment away from the technical details. Product teams developing complex AI systems can also consider breaking down the system into individual modules before starting so that the discussion will be focused and effective.

FIGURE 3.1: A visual guide of Fairness in Design.

3.3.2 Step 2: Choose an Application Card

After indicating their application scenario, the team has to choose an application card that is the most relevant to their application scenario. As defined in Section 3.2.1, there are five types to choose from.

Similarly in ethical AI design, different AI applications will have different motivations and requirements for their design that will affect how they view fairness. Hence on each card, it will include thought-provoking statements and questions to provide a guide to the team in their evaluation of fairness criteria later on. It also includes application examples so that the team can more easily identify which application category their application scenario belongs to. Figure 3.2 shows an example of an application card.

FIGURE 3.2: Application card for industrial and commercial uses.

FIGURE 3.3: Direct and indirect stakeholder cards.

3.3.3 Step 3: Identify the Stakeholders

Using the application scenario and application card as the background, the team will be given the definitions of direct and indirect stakeholders as seen in Figure 3.3 to read and understand. Each member will take a reflection card and write a stakeholder role on the space provided for it. The members have the flexibility to decide if they want to write a direct or indirect stakeholder role. This will be their identity for the remaining steps.

Identifying both categories is crucial as designers tend to focus on direct stakeholders and often miss out on indirect stakeholders in examining the impact on them.

FIGURE 3.4: An example of a group fairness card (front and back).

3.3.4 Step 4: Draw a Fairness Card

Once every member has written their stakeholder role, they will draw one fairness card each and write it down on the reflection card as well. This will be their fairness metric to reflect on later. One side of the card contains the fairness metric name, definition, and the mathematical formula of it. On the flip side contains an illustration of how the fairness metric works under a certain scenario. All cards have the same scenario. As shown in an example of a card in Figure 3.4, the scenario is a bank that has a predictive model to determine if the loan application will be approved or denied. The cards are divided into two color schemes: orange and blue. This is so to help differentiate the cards more easily into group fairness (in orange) and individual fairness (in blue). An example of an individual fairness card is shown in Figure 3.5. It is also labeled at the top right-hand corner to highlight to the members. We included the mathematical formulae as some engineers might understand the metrics easier by referring to them. As the cards are from the existing literature of fairness in AI, if new and better fairness metrics are operationalised, the tool is flexible enough to include them.

FIGURE 3.5: An example of an individual fairness card (front and back).

3.3.5 Step 5: Write a Reflection

After every member gets a fairness card, they will need to write their thoughts on the reflection card shown in Figure 3.6. With the application card as a guide, they are required to think about how the fairness metric will impact them from the perspective of their stakeholder role. Two questions are on the card to prompt the team in their thought process, namely asking them what could go right and what could go wrong with the fairness metric they have gotten. The team has to remember to think from the shoes of their stakeholder role and not as themselves.

3.3.6 Step 6: Go to Step 3 OR Collect and Discuss OR Write Review

When everyone has written their reviews, the team now has three different options to proceed.

First option is to go back to step 3 to write another round of stakeholder roles and start the process again. This is recommended for teams who want to cover

REFLECTION CARD

Stakeholder role: _____ **Fairness principle:** _____

What can go right?

What can go wrong?

FIGURE 3.6: Reflection card.

more stakeholder roles as one round might not be enough to cover both direct and indirect stakeholders.

Second option is to collect the reflection cards, randomize them, and take turns reading them out loud. The cards will not contain any names for anonymity and will include the stakeholder role, the fairness metric that is chosen, and the reflection. Discussing the cards can surface potential benefits, harms, and tensions of fairness definitions. It can also help to discover potential blind spots. As everyone takes on a role, it differs from a normal brainstorming meeting whereby the only perspectives are usually of engineers and designers. This gives a more comprehensive overview without having to conduct extensive stakeholder studies and mitigates the potential team biases.

Third option is for each member to review another member's reflection card. Reflection cards will be collected, randomized, and given out. On the review card, members can give a rating of importance over a Likert scale of 1 to 5, 1 representing not important at all and 5 representing extremely important. This is shown in Figure 3.7. The importance scale is to measure how important the reviewer thinks the fairness metric is to the product that they are creating. They can also add a comment to elaborate on why they give this rating. After everybody is done

REVIEW CARD

How important is this fairness principle to the application scenario?
From 1 - not important at all to 5 - extremely important, state your score for this fairness principle:

1
2
3
4
5

Not important at all
Extremely important

Score: _____

Please state the reason for your score:

FIGURE 3.7: Review card.

with the reviews, the cards will be collected, randomized, and shared out loud. This means that in addition to reflection cards in the previous option, the team can discuss the reviews as well. This is recommended for teams who want to work with a ranking of the fairness metrics that will likely be most important to their stakeholders and products.

3.4 Conclusion

Fairness in Design Cards embodied the FID methodology in cards format. We engaged the existing fairness literature in machine learning and the theory of VSD to create a tool that aims to help users to surface fairness concerns, navigate complex ethical choices around fairness, and overcome blind spots and team biases.

With this methodological tool, product teams that view fairness as an important value in their AI system can now be focused on it in an engaging and fun manner. They do not need to start understanding about fairness only when they have their data-set or algorithm models ready. We understand that product teams often face barriers to improve fairness in their products even when they are motivated to do so. Since our tool takes a short amount of time and is easy to follow in bite-size

details, even non-technical product team members can be included in the fairness discussion to give more diverse opinions. We hope that with this methodological tool, the bar to include fairness into the design discussion will be lowered and product teams can make better and more informed decisions for fairness in their application scenarios.

Chapter 4

FAIR Application

This chapter introduces FAIR, a project management and collaborative platform to incorporate fairness in AI design. Due to the Covid-19 situation, there is an increased uptake in working from home arrangements. Product teams are mostly working together online instead of in the office. This makes Fairness in Design cards impractical since they would not be physically together. As a result of this new norm, we extended efforts to digitise Fairness in Design into a web application. Product team members can now collaborate and go through the methodology online.

This chapter is structured as follows. Section 4.1 presents the framework and overview of the FAIR application. Section 4.2 will go into detail regarding the design and functions of the web platform. The implementation and workflow of the application are demonstrated in Section 4.3. Lastly, Section 4.4 concludes this work.

4.1 Framework and Overview

The main objective of FAIR application remains unchanged from Fairness in Design Cards. It strives to provide users a workflow that allows them to surface fairness concerns, navigate complex ethical choices around fairness, and overcome blind spots and team biases. Beyond that, it also serves as a project management platform that has functions for project leads to manage their teams, handle multiple

projects, and control the process. The platform is designed to be asynchronous in interaction meaning that the interaction between users can occur at different times and not in real time. This gives users individual time to complete the activities at a timing, place, and manner that they prefer and are comfortable with. As a result, they can contribute to their full potential.

The intended users for this application are a wide range, from product teams working on commercial products to students who want to learn more about machine learning fairness. They do not need to have any existing knowledge regarding AI fairness to use the system. The application is also flexible to be used in any stage of an AI solution design process, even when no training data has been collected yet.

Lastly, the application is hosted on a web-page domain. Users are required to create an account before using the system.

4.2 Design of FAIR Application

Even though the methodology for FAIR application is similar to Fairness in Design cards, the design considerations are different and more complicated due to the medium (web application), the interaction (asynchronous) and the purpose (project management and collaboration). This section elaborates on the functions of the FAIR application as well as illustrates the design process.

4.2.1 Web Functionality Planning

Web functionality planning is needed to decide on the important features required to fulfil the functions of the web application. This usually takes place before the design phase. We split our features into two categories: application purpose and audience needs.

Application purpose features are minimum functionality requirements to achieve the goals of the application. In this case, the goals of FAIR application are to guide fairness workflow, enable team collaboration, and manage multiple projects.

Therefore, key features include account creation, user rights management, project settings, team management, multiple tabs, progress bar, and past projects library.

Features to guide fairness workflow:

- **User rights management:** There are two types of accounts that one can create: facilitator and member account. The accounts have different dashboards and authority in the system. Facilitator accounts are able to create new projects and invite members to join in the discussion. Member accounts are only able to join in projects to contribute their opinions.
- **Project settings:** Only facilitators have access to this feature. This allows them to personalise the workflow according to team sizes. They can change the default settings of the project.

Features to enable team collaboration:

- **Account creation:** Users have to create an account with their emails. The email address must be unique and not found in the database. This gives them a unique identity in the application and makes it easier for facilitators to find their team members.
- **Team management:** Only facilitators have access to this feature. This allows them to invite and communicate with team members of different projects with ease.

Features to manage multiple projects:

- **Multiple tabs:** There are multiple tabs on the member dashboard, namely Project Requests, In Progress, Needs Review, and Results. The tabs indicate the specific actions needed to perform for the projects under them. This provides a clearer picture of the inputs needed and a cleaner organisation of all the projects.
- **Progress bar:** It is indicated on every project's display card. This helps facilitators to keep track of their project progress at a glance.
- **Past projects library:** Completed projects will be stored in this separate library so that they do not interfere with the current projects in the interface.

FIGURE 4.1: User flow for FAIR Application.

Beside these functions, audience needs features are important as well. They are features that enable the end users to operate the application efficiently. These include messaging, in-app and email notification systems.

- **Messaging:** Chat function allows users to communicate with each other within the application. This reduces the need to bring project discussion outside of the application.
- **In-app and email notification systems:** As users are able to do their activities on their own preferred timing, a notification system is important to update them about the status of the projects and notify them of the next step. A dual notification system is used as users might not log into the application to check in-app notification.

4.2.2 User Flow

After we identified the essential features to include in the application, we need to determine how users can complete their tasks in the application. User flow is a diagram that gives a bird's eye view of the path users will take through the application to achieve certain goals. Figure 4.1 illustrates the user flow for FAIR application. It is in the form of a flow chart and has nodes for each major navigational path.

FIGURE 4.2: Information Architecture for FAIR Application.

As indicated in the legend, the blue rectangles represent pages or screens in the application, yellow circles represent actions that users have to take, white triangles represent decisions that users have to make, and green ovals represent terminators of the paths. This gives a clear indication of the steps required to get from the entry point of registration to eventually viewing the results of their discussion. This information will then be used to design the individual pages of the application, as well as to determine the content and navigation links.

4.2.3 Information Architecture

As a project management and collaborative platform, there are many functions required. If the functions are not organised properly, it can cause frustration and confusion to the users. Therefore, it is important to provide a structure to them so it is easy for the users to learn and use the application. Information architecture is a method to organise and structure the content of an application.

Since we know the features needed and the user flow, we are able to create an information architecture for the application as shown in Figure 4.2. It is in the form of a sitemap which visually indicates how different pages, content and functions are related to one another. The sitemap also illustrates the hierarchy and flow of the content. From Figure 4.2, we can see the login page at the top which signals the start of the application and drops down to email login before separating into

FIGURE 4.3: Main workflow in FAIR Application.

facilitator and member dashboards. It is clear from the diagram which contents are overlapping between the two dashboards and which contents are different. Therefore, the information architecture is able to be used to guide the creation of the application.

4.3 Implementation of FAIR Application

This section will focus on the procedure of a project discussion on the FAIR application from the user perspective. Figure 4.3 illustrates the main workflow of both types of account in the application.

4.3.1 Facilitator Account

Figure 4.4 shows the dashboard for facilitators. There are two tabs: My Projects and Results. My projects tab display all the projects created by the facilitator in cards style. Results tab display all the completed projects that were created by the facilitator in cards style as well.

When facilitators click on the "New Project" button to create a new project, they are required to input the application scenario shown in Figure 4.5. It can be an actual product they are working on or a fictional one. They are advised by the prompt given to be specific about the areas for which there are concern about fairness issues.

FIGURE 4.4: Facilitator Dashboard.

FIGURE 4.5: Application Scenario Page.

FIGURE 4.6: Application Type Page.

FIGURE 4.7: Invitation Page.

FIGURE 4.8: Settings for each project.

Next, they choose an application-type card that is most relevant to their application scenario shown in Figure 4.6. There are five categories to choose from: 1) life-critical systems; 2) industrial and commercial uses; 3) office, home, and entertainment; 4) exploratory, creative, collaborative applications; and 5) sociotechnical applications. They are adapted from Shneiderman’s classification of usability motivations [61] in the field of Human-Computer Interaction (HCI). The cards also come with statements that detail the unique pointers about fairness for each application type to serve as a guide to prompt deeper thinking.

Facilitators can then key in the names or emails of the members they wish to invite and send the invitations through the platform in Figure 4.7.

Once all the members have responded, the system will notify facilitators to change the setting of the project. This is shown in Figure 4.8. The default minimum

FIGURE 4.9: Member Dashboard.

number of fairness cards given to each member is ten and the default minimum number of reviews for each fairness card is $(n-1)$, whereby n represents the number of members in the project. They can also set a deadline for the project. The facilitator role in the discussion ends here.

4.3.2 Member Account

Figure 4.9 shows the dashboard for members. There are four tabs: Project Requests, In Progress, Needs Review, and Results. All the tabs use the same style to display the projects.

Upon receiving the project request, members can decide whether to accept or reject the project. If members accept the invitation, they can choose to start on the project immediately or save it for later. The project card will be moved to In Progress tab. If they reject the invitation, the project card will be removed from their dashboard.

In Progress tab will show the project cards in two colour as seen in Figure 4.10. Orange indicates that members have not given their input while green indicates that input is received and complete.

Members start the fairness discussion process by viewing the application scenario the facilitator has written as well as the application type chosen. An example is shown in Figure 4.11.

Next, members are shown definitions of direct and indirect stakeholder so that they are made aware about the differences between the two. This also highlights the

FIGURE 4.10: In Progress Tab.

FIGURE 4.11: An example of a project background that members will first read.

FIGURE 4.12: An example of a stakeholder role.

FIGURE 4.13: An example of a randomised fairness card shown to members.

role of indirect stakeholder as most users tend to think of only direct stakeholders. These definitions are taken from Judgment call [46]. Direct stakeholders refer to those who interact with the AI system, whereas indirect stakeholders refer to those who do not interact with the AI system but are still affected by it.

Based on the definitions of stakeholder, members are asked to input in a stakeholder role relevant to their application scenario. It can be either a direct or indirect stakeholder. An example is shown in Figure 4.12.

After inputting a stakeholder role, the system randomly picks a fairness card out of ten available to show to the members. An example is shown in Figure 4.13. Even though it is randomised, the system will not duplicate the cards shown to each member to ensure the cards are spread out. These ten cards represent the 10 most widely used fairness definitions identified in the current machine learning literature [7, 67]. They are the same fairness cards used in Fairness In Design. Members are asked to read and comprehend both sides of the fairness card before the next step.

Once the members are confident to move on, they have to put themselves in the shoes of the stakeholder role they have written earlier and answer two questions. They have to think about how the fairness principle can provide benefits to the stakeholder as well as how it can cause negative consequences. An example of the reflection is shown in Figure 4.14.

As there are ten fairness principles in the FAIR application, it is recommended for the members to repeat their workflow multiple times if they have a team of less than 10 members so they can cover all ten principles. If the minimum number of

FIGURE 4.14: An example of a reflection done on a fairness card.

FIGURE 4.15: An example of a fairness review.

fairness cards given to each member is not met, the system will ask the members to select another stakeholder role. Each new round will start with stakeholder role selection and end with a different fairness card to complete the envisioning design steps.

After members have completed the minimum number of fairness cards, they have to wait for the remaining team members to submit their answers. Once all members have submitted their responses, the project will move to the peer review stage. Project card will be under the Needs Review tab. All members will receive both an in-app and email notification that the project is ready for review.

During the peer review stage, each member is shown another member's fairness input. This consists of the stakeholder role chosen, the fairness card that was given, and the responses to the two questions. An example can be seen in Figure

FIGURE 4.16: Results tab that is available in both facilitator and member account.

Figure 4.15. Based on the information, members have to rate the importance of the fairness principle to the given application scenario. The scale is based from 1 to 5, 1 being not important at all, 3 being moderately important, and 5 being very important. Members are asked for the reasons behind the ratings.

As facilitator can choose the minimum number of review each fairness principle has, the system will decide on how many reviews each member has to do to hit that number. The default setting in the system will compute according to the number of members in the project.

4.3.3 Both Facilitator and Member Account

When all the members have submitted their peer reviews, the system will notify both the facilitators and members that the project is completed so that they can view the results. The Results tab shown in Figure 4.16 is available in both facilitator and member dashboard. It allows both the facilitators and members to review and export the results of their project discussion. The results page shown in Figure 4.17 will help the project team rank the fairness principles from the most important to the least important according to the ratings given by the members. Each fairness principle can be expanded to see the reasoning behind the ratings given. The more inputs the system receives, the higher quality the ranking of fairness principles becomes. This information enables project teams to prioritise on

The screenshot shows the 'VIEW RESULTS' page of the FAIR application. At the top, there is a breadcrumb trail 'HOME > PROJECT AI > VIEW RESULTS' and a 'Sort By' dropdown menu. Below this is a table with two columns: 'Fairness Principle' and 'Rating'. The table lists five principles with their respective ratings and expandable details (indicated by a downward arrow icon).

HOME > PROJECT AI > VIEW RESULTS		Sort By ▾
Fairness Principle	Rating	
Demographic Disparity	5	▾
Treatment Equality	4.5	▾
Equalized Odds	4	▾
Test Fairness	3.5	▾
Equal Opportunity	3.5	▾

FIGURE 4.17: Results page of a project showing the computed fairness ranking.

the important principles and can be used to trace the rationale behind the adoption of the fairness principles in AI solutions.

4.4 Conclusion

FAIR application is an online platform which helps project teams to surface fairness concerns, identify the most relevant fairness principles, and promote perspective thinking in their application scenario. Not only that, it also serves as a medium to create and manage multiple projects in one area. The easy-to-follow steps can lower the barrier in incorporating fairness into the design discussion, and make the process efficient and engaging. With the increase in work-from-home and remote work arrangements due to Covid-19, having an online tool with asynchronous interaction will help to reduce the need to arrange meetings and provide the flexibility to work on fairness at their convenience.

Chapter 5

User Study

This chapter presents the details about the user study conducted to evaluate the usefulness of the methodology process Fairness in Design (FID), namely to surface fairness concerns, navigate complex ethical choices around fairness, and overcome blind spots and team biases. Section 5.1 will describe the demographics of participants. Section 5.2 will illustrate the design of the user study as well as the questionnaires used. Section 5.3 will elaborate on the procedure of the user study and Section 5.4 will discuss the analysis of results and my findings. Lastly, Section 5.5 concludes this work.

5.1 Participants

Ethics approval was obtained from the Institutional Review Board of our university before we began recruiting participants. All participants in this study were recruited primarily through snowball sampling, by emailing contacts working in Nanyang Technological University research centres. The participants were asked to share our recruiting message with colleagues working on AI systems. Only participants who are currently working in a project that can involve fairness and the project must be in an AI field are eligible for the study. They were compensated 10 Singaporean Dollar each for their time.

FIGURE 5.1: Survey demographics of the application domains (left) and technology areas (right).

In total, 24 participants were recruited for the user study. There were 18 males and 6 females between the ages of 24 to 45. The participants span a range of technology areas, including computer vision, conversational chatbots, predictive analytics, recommender systems, and natural language processing. There is a spread of application domains as well in areas such as healthcare, media and entertainment, retail, financial services, education, marketing, and social distancing. The technology areas and application domains are shown in Figure 5.1. Participants are allowed to select multiple options. The participants' role and the IDs are shown in Table 5.1.

5.2 Design of User Study and Questionnaires

The purpose of user study is to evaluate the usefulness of the methodology process. Usefulness of FID is made up of three components: to navigate and make complex choices regarding fairness, to detect issues surrounding fairness, and to overcome

Participants' Role	Participant IDs
Data Labeler	15
Data Scientist	13,19
Designer	23,25
Project Manager	7,18,22
Researcher	1, 2, 3, 4, 5, 6, 11, 10, 12, 14, 20, 21
Social Scientist	24
Software Engineer	8,9,16

TABLE 5.1: Participant demographics: self-reported team role and participant ID.

team biases and blind-spots concerning fairness. Therefore, the three alternative hypotheses for this user study are stated below:

Hypothesis 1: After I use this methodology, I am better able to choose the fairness criteria that is most applicable to my specific scenario.

Hypothesis 2: After I use this methodology, I am better able to surface fairness concerns in my application scenario.

Hypothesis 3: After I use this methodology, I am better able to stimulate perspective taking from different stakeholders.

The user study was separated into two parts to test the hypotheses. It is a within-subject design such that all participants will experience the same treatment which is to go through the methodology process.

The first part of the study consisted of an online pre-study questionnaire. The aim of the questionnaire is to understand the participants' current viewpoint towards fairness in AI and their design process involving fairness, to measure their self-assessment of their ability in achieving the three hypotheses in their project, and to gather demographic details. The second part of the study consisted of a workshop session where participants gathered in teams to go through the FID methodology and a post-study questionnaire. The workshop session took place at least three weeks from the date the participants completed the pre-study questionnaire to prevent any memory recall effects as the self-assessment component on the questionnaire was repeated so we can compare the results. Participants were grouped together with their teammates in the same project, but if their teammates

were not available for the session, they would be assigned with someone else. Each workshop session has at least two participants.

The pre-study questionnaire comprised five sections. The first section contained demographic questions to understand the participants' background better like their project domain, technology area and role in their team. The next three sections asked for the participants' view of fairness in AI/ML, the existing practice of fairness in their team, and if they discovered potential fairness issues before in their projects. The last section was a self-assessment of their current ability in achieving the three aims of the tool. Since there are currently no standardized questionnaires that measure these constructs, the questions in the self-assessment were adapted from a survey used to study what AI practitioners need to develop fairer ML systems [18]. It consisted of 20 Likert scale questions with subscales for the three constructs. The participants indicated their level of agreement with statements on a 5-point scale ranging from "strongly disagree" to "strongly agree". There was an additional option labelled as "N/A" for participants who did not think about fairness in their project. As the adapted questions from the survey were all positive statements, we added in negative statements for all questions to check on the reliability of the responses. An example positive question is "I am able to navigate complex ethical choices around AI/ML fairness" whereas the negative counterpart question is "I do not know how to make decisions regarding fairness in AI/ML".

The post-study questionnaire comprised three sections. The first section asked the participants again regarding their view of fairness in AI/ML. The second section contained the same self-assessment questions as the pre-study questionnaire so we can compare and analyse the results. The last section contained questions to evaluate their experience in using FID such as if they will use it again, net promoter score, as well as to get feedback.

5.3 User Study Procedure

Participants were told the study will be in two parts. For the first part, they will be sent a link to the online pre-study questionnaire and the participant consent form through email. This process will take less than 30 minutes to complete. After they have finished the questionnaire, they were given a few dates and time slots

to choose from that are at least three weeks from the date they submitted the questionnaire. We coordinated the common date and timing between the team members and sent them a confirmation email with the details for the second part of the study.

There were a total of ten sessions held for the workshop. Team sizes for the workshop range from two to four members. Two and a half hours were allocated for each session. In each session, the participants were first introduced to the purpose of the workshop as well as the FID cards. Instructions on how to use the tool were given and explained to them. A facilitator was always present to provide guidance and assistance. The participants were then asked to start their discussion according to the steps of the tool. All of them completed the process till the last step of review and rank. After they wrapped up their discussion, they were given the link to an online post-study questionnaire. When they completed the questionnaire, they were then debriefed and provided with compensation.

5.4 Analysis of Results

From the user study, we were interested to find out if participants' competency in the three constructs improved significantly after using the FID tool. The competency was measured by the self-assessment component in the questionnaire. Order of items in the questionnaire was randomised. All 24 participants completed both pre and post study questionnaires at the end of the user study. Data from 3 participants were excluded from analysis as the participants indicated in the pre-study questionnaire that they did not think about fairness in their project. As a result, we were unable to determine their competency level before using the tool. Results from the Likert scale were relabelled into numbers to conduct statistical analysis.

5.4.1 Hypotheses Testing

Hypothesis 1 stated that going through the FID methodology tool would allow participants to better choose the fairness criteria that is most applicable to their specific application scenario. A paired-samples t-test was conducted to compare participants' competency in navigating and making complex fairness choices before

and after using the tool. There was a significant difference in the scores for before using the tool ($M = 21.86$, $SD = 37.23$) and after using the tool ($M = 26.19$, $SD = 30.26$), $t(20) = 3.52$, $p = .002$. These results suggest that the tool did increase the ability of participants in navigating and choosing the fairness criteria for their application scenario.

Hypothesis 2 stated that using the FID methodology tool would allow participants to better surface fairness concerns in their application scenario. A paired-samples t-test was conducted to compare participants' competency in detecting fairness issues before and after using the tool. There was a significant difference in the scores for before using the tool ($M = 22.43$, $SD = 45.86$) and after using the tool ($M = 26.14$, $SD = 21.13$), $t(20) = 2.91$, $p = .009$. These results suggest that the tool did increase the ability of participants in detecting fairness issues in their application scenario.

Hypothesis 3 stated that using the FID methodology tool would allow participants to better stimulate perspective taking from different stakeholders. A paired-samples t-test was conducted to compare participants' competency in understanding different stakeholders before and after using the tool. There was a significant difference in the scores for before using the tool ($M = 11.62$, $SD = 12.75$) and after using the tool ($M = 14.19$, $SD = 6.66$), $t(20) = 3.34$, $p = .003$. These results suggest that the tool did increase the ability of participants in understanding the perspective of multiple stakeholders.

5.4.2 Attitude towards Fairness in AI/ML

We found that the attitude towards fairness in AI/ML has changed after they have used the tool. Before using the tool, on a 5-point Likert scale from "Not at all" to "A great deal", only 6 participants (25%) reported that they prioritise fairness in AI/ML "A great deal", "A lot" or "A moderate amount", and 18 participants (75%) indicated "Not at all" or "A little". Whereas after the workshop, majority of the participants (91.7%) expressed that they want to prioritise fairness in AI/ML in the future "A great deal", "A lot" or "A moderate amount", and with only 2 participants (8.3%) indicated "A little".

Likewise before using the tool, only 7 participants (29.2%) expressed “Yes” when asked if fairness in AI/ML is something they would like their team to prioritize more. Whereas after the workshop, 14 participants (58.3%) expressed “Yes” when asked the same question.

These results suggest that most of the participants realised the significance of fairness in AI design after going through the workshop such that they would like themselves and their teams to prioritize fairness more in their projects. We believe that this is because most participants do not have much exposure to fairness related concepts in their projects. Out of 24 participants, only 3 (12.5%) indicated that fairness and bias is currently part of their team’s AI/ML model-building checklist. Therefore, this workshop increased their awareness of the fairness principles in the AI/ML literature, as well as allowed them to recognise how fairness can impact their project.

5.4.3 Feedback of Fairness in Design Workshop

In our post-study questionnaire, we asked participants about their experience of the workshop. When asked how likely they were to use Fairness in Design in the future, 18 participants (75%) indicated “definitely yes” and “probably yes” to using it again. Seven participants stated the reason to be that the tool was useful and helpful in guiding and identifying fairness issues and concerns. One participant stated that “at least I can show the tool to the stakeholders who are concerned about fairness”. Another participant stated that “using this tool can save our time to work out all the possible problems and to avoid possible omissions about fairness issues”.

We also asked participants how likely they were to recommend Fairness in Design to a friend or colleague. Based on a scale of 1 to 10, 15 participants (62.5%) gave us a score of 7 and above. Similarly, the most common reason stated was because the tool helped to encourage them to think about fairness in AI and dissect vague issues.

When asked what they like about the tool, the majority of the participants liked it for introducing principles and concepts of fairness so that they can understand fairness from different aspects and angles. One participant stated that “the tool

can help to take into account the interests of different stakeholders”. Participants also shared various ways to improve on the methodological flow of the tool. Some feedback mentioned the need to customise the tool to more AI applications as the current version is more fitting for machine learning applications. Other feedback shared that more examples can be given as some of the fairness principles were very similar to each other and hard to distinguish.

5.5 Conclusion

In this chapter, a user study was conducted to evaluate the usefulness of Fairness in Design. Design and procedure of the user study and questionnaires were detailed. Results from the user study have shown us that Fairness in Design was indeed useful in navigating and choosing fairness criteria, surfacing fairness concerns, and stimulating perspective taking from different stakeholders. Not only that, Fairness in Design was able to let participants realise the importance of considering fairness values in their projects and provide them exposure to fairness in AI literature. Majority of the participants also expressed that they would like to use the methodological tool in the future and would recommend it to their friends or colleagues.

Chapter 6

Conclusion and Future Works

6.1 Conclusion

In Chapter 1 and Chapter 2, the background, motivation and objectives for this research are presented. I explained why incorporating human values in AI system design process is important and the reasons behind the focus on fairness value only. The need for more methodology tools to help AI practitioners in creating fairness discussion is also explained. Therefore, the objectives are to create a methodology to enable practitioners to navigate ethical choices in fairness, surface fairness concerns, highlight different stakeholder perspectives and on top of these, it must be easy to learn and engaging to use without the need for any prior knowledge from the users.

The Fairness in Design methodology is available in two mediums for different use cases. The steps of the methodology guide practitioners to think about the contexts of their application domain, let them be immersed with stakeholders' perspectives, encourage them to think about fairness issues and trade-offs from both technical and socio-cultural viewpoints, and enable them to vote for the most relevant and suitable fairness principles. It can be used without a data-set or ready algorithm.

Chapter 3 illustrates the cards version which is mainly for in-person team discussions. Content on the cards are easy to understand and the set of cards is also fully customisable to the users' needs. Whereas Chapter 4 illustrates the web application version which facilitates the discussion online. The application allows the

users to complete their tasks at their own place and time, which takes the pressure off them and enable them to think better. A user study was conducted on 24 participants to evaluate the effectiveness of the methodology. Details of the study and results were revealed in Chapter 5. Based on the results, the methodology was found to achieve its objectives to enable practitioners to navigate ethical choices in fairness, surface fairness concerns, highlight different stakeholder perspectives. Not only that, the tool was also found to increase AI practitioners' awareness of fairness principles and improved their attitude towards fairness in AI. Majority of the participants also stated that they would like to use the tool again and would also recommend it to their colleagues or friends.

To the best of our knowledge, Fairness in Design Cards and FAIR application are the first technical tools to facilitate AI solution development teams to incorporate fairness into their designs in an active, guided manner.

6.2 Future Work

In the future, I will expand both FDC and FAIR application to include other ethical AI aspects, such as data privacy and explainability, to provide more holistic design support.

Not only that, I will also extend the FAIR application to incorporate more analytical processes (e.g., cost-benefit analysis) to evaluate the trade-off between adopting additional fairness principles and achieving the design goals.

Appendix A

Application Type Cards

FIGURE A.1: Life-Critical Systems.

FIGURE A.2: Industrial and Commercial Uses.

FIGURE A.3: Office, Home, Entertainment Applications.

FIGURE A.4: Exploratory, Creative, Collaborative Applications.

FIGURE A.5: Sociotechnical Systems.

Appendix B

Fairness Cards

B.1 Group Fairness Cards

FIGURE B.1: Demographic Parity/ Statistical Parity (Front).

FIGURE B.2: Demographic Parity/ Statistical Parity (Back).

FIGURE B.3: Conditional Statistical Parity (Front).

FIGURE B.4: Conditional Statistical Parity (Back).

FIGURE B.5: Equalized Odds (Front).

FIGURE B.6: Equalized Odds (Back).

FIGURE B.7: Equal Opportunity (Front).

FIGURE B.8: Equal Opportunity (Back).

FIGURE B.9: Treatment Equality (Front).

FIGURE B.10: Treatment Equality (Back).

FIGURE B.11: Test Fairness (Front).

FIGURE B.12: Test Fairness (Back).

B.2 Individual Fairness Cards

FIGURE B.13: Fairness Through Awareness (Front).

FIGURE B.14: Fairness Through Awareness (Back).

FIGURE B.15: Fairness Through Unawareness (Front).

FIGURE B.16: Fairness Through Unawareness (Back).

FIGURE B.17: Counterfactual Fairness (Front).

FIGURE B.18: Counterfactual Fairness (Back).

FIGURE B.19: Fairness in Relational Domain (Front).

FIGURE B.20: Fairness in Relational Domain (Back).

Appendix C

Questionnaires

C.1 Pre-Study Questionnaire

13/08/2021

Fairness Questionnaire

Fairness Questionnaire

The purpose of this questionnaire is to understand current opinions, practices, and issues in creating fair artificial intelligence (AI) and machine learning (ML) systems.

By "fairness" we refer, very broadly, to cases where AI / ML systems behave differently for certain individuals or groups (e.g., age, race, or gender groups) in systemic, undesirable ways.

*Required

1. Index *

2. Gender *

Mark only one oval.

☐ Male

☐ Female

3. Age *

4. Which of the following best describes your role on your team? (Please check all that apply) *

Tick all that apply.

- ☐ Data Labeler
- ☐ Data Scientist
- ☐ Designer
- ☐ Domain/Content Expert
- ☐ Executive/ General Manager
- ☐ Product Manager
- ☐ Project/Program Manager
- ☐ Researcher
- ☐ Social Scientist
- ☐ Software Engineer
- ☐ Technical Lead/Manager

Other: ☐ _____

5. Which of the following best describe the application domains of your team's machine learning products/services? (Please check all that apply) *

Tick all that apply.

- ☐ Defense/Military
- ☐ Media and Entertainment
- ☐ Education
- ☐ Financial Services
- ☐ General-purpose Machine Learning Tools (eg. APIs)
- ☐ Financial Services
- ☐ Healthcare
- ☐ Government/Public Sector
- ☐ Hiring/Recruiting
- ☐ Retail
- ☐ Marketing

Other: ☐ _____

6. Which of the following technology areas best describe your team's machine learning products/services (Please check all that apply) *

Tick all that apply.

- ☐ Computer Vision/ Image Analysis
- ☐ Conversational AI/ Chatbots
- ☐ Decision support
- ☐ Natural Language Processing
- ☐ Robotics/Cyber-physical Systems
- ☐ Search/ Information Retrieval
- ☐ Speech and Voice
- ☐ User Modeling/ Adaptive Hypermedia
- ☐ Predictive Analytics
- ☐ Recommender Systems

Other: ☐ _____

7. Which of the following are currently part of your team's AI/ML model-building checklist? (Please check all that apply) *

Tick all that apply.

- ☐ Explainability and Transparency
- ☐ Compliance
- ☐ User Control over Data and Models
- ☐ Privacy
- ☐ Fairness and Bias
- ☐ None of the above
- ☐ I don't know

View towards fairness in AI/ML

8. To what extent do you currently prioritize fairness in AI/ML? *

Mark only one oval.

- ☐ Not at all
☐ A little
☐ A moderate amount
☐ A lot
☐ A great deal

9. How long ago did your team begin prioritizing fairness in AI / ML? *

Mark only one oval.

- ☐ We are just beginning
☐ Within the last 6 months
☐ 6 months to a year ago
☐ 1 to 2 years ago
☐ 2 to 5 years ago
☐ More than 5 years ago

10. Is fairness in AI / ML something you would like your team to prioritize more? *

Mark only one oval.

- ☐ No
☐ Maybe
☐ Yes

11. In general, do you think other people on your team would like to prioritize fairness more? *

Mark only one oval.

- ☐ Definitely not
☐ Probably not
☐ Might or might not
☐ Probably yes
☐ Definitely yes

Practice of fairness in your team

12. Does your team consider fairness when collecting data or designing training/test datasets? (If your team has no control over data collection and training/test set design, please select 'N/A') *

Mark only one oval.

- ☐ No
☐ Yes
☐ N/A

13. How often does your team typically consider fairness at each of the following steps of the data pipeline? (If your team has no control over a given step, please select 'N/A') *

Mark only one oval per row.

	Never	Rarely	Sometimes	Very Often	Always	N/A
Before collecting data	☐	☐	☐	☐	☐	☐
During data collection	☐	☐	☐	☐	☐	☐
When designing training datasets (eg/ curation, dataset augmentation)	☐	☐	☐	☐	☐	☐
When designing test datasets	☐	☐	☐	☐	☐	☐
When developing AI/ML models	☐	☐	☐	☐	☐	☐
When evaluating AI/ML models	☐	☐	☐	☐	☐	☐

Fairness issues in your project

14. Has your team ever found fairness issues (or potential fairness issues) in your products/ services (at any stage of development or deployment)? *

If your team did not try to find fairness issues, please select N/A

Mark only one oval.

- ☐ No
☐ Yes
☐ N/A

15. At which of the following stages does your team typically find fairness issues or potential fairness issues in your products/services? *

If your team did not try to find fairness issues in a given step, please select N/A

Mark only one oval per row.

	Never	Rarely	Sometimes	Very Often	Always	N/A
During initial conception or specification of a product/service	☐	☐	☐	☐	☐	☐
During data collection	☐	☐	☐	☐	☐	☐
During the design of training sets (e.g., curation, dataset augmentation)	☐	☐	☐	☐	☐	☐
During the design of test sets	☐	☐	☐	☐	☐	☐
During development of ML models	☐	☐	☐	☐	☐	☐
During early-stage testing/prototyping of product/service	☐	☐	☐	☐	☐	☐
During late-stage testing of product/service before deploying	☐	☐	☐	☐	☐	☐
Once product/service has been deployed in the real world	☐	☐	☐	☐	☐	☐

16. If your team uses metrics/key performance indicators (KPIs) to monitor for potential fairness issues, please briefly describe what metrics or tests you use. *

If your team did not use any metric or KPI, please write N/A

17. Do you suspect there are fairness issues in your products/services that your team has not yet found? *

Mark only one oval.

- ☐ Definitely not
- ☐ Probably not
- ☐ Might or might not
- ☐ Probably yes
- ☐ Definitely yes

Self-assessment

If you did not think about fairness in your project, please select N/A

18. Rate how strongly you agree with each statement. *

Mark only one oval per row.

	Strongly Disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree	N/A
I am able to navigate complex ethical choices around AI/ML fairness	☐	☐	☐	☐	☐	☐
I do not know how to make decisions regarding fairness in AI/ML	☐	☐	☐	☐	☐	☐
I can specify and enforce fairness criteria to train my AI/ML models	☐	☐	☐	☐	☐	☐
I am not sure which fairness criteria I should use for my AI/ML models	☐	☐	☐	☐	☐	☐
I am able to compare across different possible strategies for addressing a fairness issue in my application area	☐	☐	☐	☐	☐	☐
I do not know how to evaluate different fairness solutions	☐	☐	☐	☐	☐	☐
I can prioritize which of the fairness issues I have found most needed to be addressed	☐	☐	☐	☐	☐	☐
I am unable to determine which fairness issues are	☐	☐	☐	☐	☐	☐

more important to
focus on first

I understand what
kinds of fairness
issues can arise in my
application area

☐☐☐☐☐☐

I am not sure of the
implication of
fairness in my
application area

☐☐☐☐☐☐

I am able to detect
potential fairness
issues during the
process of developing
AI/ML models

☐☐☐☐☐☐

I am unable to identify
fairness issues that
might occur when
creating AI/ML
models

☐☐☐☐☐☐

I have a process to
help anticipate
fairness issues I
might otherwise miss

☐☐☐☐☐☐

I do not know how to
surface fairness
issues that I might
have neglected or
overlooked

☐☐☐☐☐☐

I am able to
understand what side
effects a particular
"fix" for a fairness
issue might have on
my users'/customers'
experience

☐☐☐☐☐☐

I am unsure of the
repercussions on my
users/customers
when I use a certain

☐☐☐☐☐☐

13/08/2021

Fairness Questionnaire

fairness criteria to
solve a fairness issue

I can figure out which
subgroups of
stakeholders are
important to think
about, in terms of
fairness

☐☐☐☐☐☐

I am unsure of which
stakeholders to
prioritise when it
comes to fairness

☐☐☐☐☐☐

I can understand what
stakeholders of my
products and services
will consider fair vs
unfair

☐☐☐☐☐☐

I do not know how my
stakeholders might
interpret fairness

☐☐☐☐☐☐

This content is neither created nor endorsed by Google.

Google Forms

C.2 Post-Study Questionnaire

13/08/2021

Post-Study Questionnaire

Post-Study Questionnaire

*Required

1. Index *

2. Is fairness in AI / ML something you would like your team to prioritise more? *

Mark only one oval.

- ☐ No
☐ Maybe
☐ Yes

3. To what extend will you want to prioritise fairness in AI/ML in the future? *

Mark only one oval.

- ☐ Not at all
☐ A little
☐ A moderate amount
☐ A lot
☐ A great deal

Self-assessment

4. Rate how strongly you agree with each statement. *

Mark only one oval per row.

	Strongly Disagree	Disagree	Neither agree nor disagree	Agree	Strongly agree	N/A
I am able to navigate complex ethical choices around AI/ML fairness	☐	☐	☐	☐	☐	☐
I do not know how to make decisions regarding fairness in AI/ML	☐	☐	☐	☐	☐	☐
I can specify and enforce fairness criteria to train my AI/ML models	☐	☐	☐	☐	☐	☐
I am not sure which fairness criteria I should use for my AI/ML models	☐	☐	☐	☐	☐	☐
I am able to compare across different possible strategies for addressing a fairness issue in my application area	☐	☐	☐	☐	☐	☐
I do not know how to evaluate different fairness solutions	☐	☐	☐	☐	☐	☐
I can prioritize which of the fairness issues I have found most needed to be addressed	☐	☐	☐	☐	☐	☐
I am unable to determine which fairness issues are more important to focus on first	☐	☐	☐	☐	☐	☐
I understand what kinds of fairness issues can arise in my application area	☐	☐	☐	☐	☐	☐
I am not sure of the implication of fairness in my application area	☐	☐	☐	☐	☐	☐
I am able to detect potential fairness issues during the process of developing AI/ML models	☐	☐	☐	☐	☐	☐

13/08/2021

Post-Study Questionnaire

I am unable to identify fairness issues that might occur when creating AI/ML models	☐	☐	☐	☐	☐	☐
I have a process to help anticipate fairness issues I might otherwise miss	☐	☐	☐	☐	☐	☐
I do not know how to surface fairness issues that I might have neglected or overlooked	☐	☐	☐	☐	☐	☐
I am able to understand what side effects a particular "fix" for a fairness issue might have on my users'/customers' experience	☐	☐	☐	☐	☐	☐
I am unsure of the repercussions on my users/customers when I use a certain fairness criteria to solve a fairness issue	☐	☐	☐	☐	☐	☐
I can figure out which subgroups of stakeholders are important to think about, in terms of fairness	☐	☐	☐	☐	☐	☐
I am unsure of which stakeholders to prioritise when it comes to fairness	☐	☐	☐	☐	☐	☐
I can understand what stakeholders of my products and services will consider fair vs unfair	☐	☐	☐	☐	☐	☐
I do not know how my stakeholders might interpret fairness	☐	☐	☐	☐	☐	☐

Your experience

5. On a scale of 1 to 10, how likely are you to recommend this tool to a friend or colleague? *

Mark only one oval.

	0	1	2	3	4	5	6	7	8	9	10	
Not at all likely	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	☐	Extremely likely

6. What are the primary reasons for the score you give us? *

7. Will you use this tool in your future projects? *

Mark only one oval.

- ☐ Definitely not
☐ Probably not
☐ Might or might not
☐ Probably yes
☐ Definitely yes

8. What are the primary reasons for your answer? *

13/08/2021

Post-Study Questionnaire

9. What do you like about the tool? *

10. How can this tool be improved? *

This content is neither created nor endorsed by Google.

Google Forms

List of Author's Publications¹

Conference Proceedings

- **Ying Shu**, Jiehuang Zhang and Han Yu, “Fairness in Design: A Tool for Guidance in Artificial Intelligence Design,” in *Proceedings of the 13th International Conference on Social Computing and Social Media (SCSM’21)*, 2021.

¹The superscript * indicates joint first authors

Bibliography

- [1] Klaus Schwab. *The fourth industrial revolution*. Currency, 2017. [1](#)
- [2] Spyros Makridakis. The forthcoming artificial intelligence (ai) revolution: Its impact on society and firms. *Futures*, 90:46–60, 2017. [1](#)
- [3] Cary Coglianese. Algorithmic regulation. *The Algorithmic Society: Technology, Power, and Knowledge*, 2020. [1](#)
- [4] Kate Crawford, Meredith Whittaker, Madeleine Clare Elish, Solon Barocas, Aaron Plasek, and Kadija Ferryman. The ai now report. *The Social and Economic Implications of Artificial Intelligence Technologies in the Near-Term*, 2016. [2](#)
- [5] Wei Xu. Toward human-centered ai: a perspective from human-computer interaction. *Interactions*, 26(4):42–46, 2019. [2](#)
- [6] Batya Friedman and Helen Nissenbaum. Bias in computer systems. *ACM Transactions on Information Systems (TOIS)*, 14(3):330–347, 1996. [2](#)
- [7] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. A survey on bias and fairness in machine learning. *arXiv preprint arXiv:1908.09635*, 2019. [2](#), [3](#), [10](#), [11](#), [17](#), [35](#)
- [8] Kate Crawford. Artificial intelligence’s white guy problem. *The New York Times*, 25(06), 2016. [2](#)
- [9] Adrienne Yapo and Joseph Weiss. Ethical implications of bias in machine learning. In *Proceedings of the 51st Hawaii International Conference on System Sciences*, 2018. [2](#)
- [10] Julia Angwin, Jeff Larson, Surya Mattu, and Lauren Kirchner. Machine bias: there’s software used across the country to predict future criminals. and it’s biased against blacks. *ProPublica*, 2016. [2](#)
- [11] Ziad Obermeyer, Brian Powers, Christine Vogeli, and Sendhil Mullainathan. Dissecting racial bias in an algorithm used to manage the health of populations. *Science*, 366(6464):447–453, 2019. [2](#)
- [12] John Havens. *Heartificial intelligence: embracing our humanity to maximize machines*. Jeremy P. Tarcher/Penguin, 2016. [3](#)

- [13] Batya Friedman, David G Hendry, and Alan Borning. A survey of value sensitive design methods. *Foundations and Trends in Human-Computer Interaction*, 11(2):63–125, 2017. [3](#), [9](#), [10](#), [16](#)
- [14] Mary Flanagan, Daniel C Howe, and Helen Nissenbaum. Values at play: Design tradeoffs in socially-oriented game design. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 751–760, 2005. [3](#)
- [15] Sahil Verma and Julia Rubin. Fairness definitions explained. In *2018 ieee/acm international workshop on software fairness (fairware)*, pages 1–7. IEEE, 2018. [3](#), [11](#), [17](#)
- [16] Alexandra Chouldechova. Fair prediction with disparate impact: A study of bias in recidivism prediction instruments. *Big data*, 5(2):153–163, 2017. [3](#), [12](#), [17](#)
- [17] Jon Kleinberg, Sendhil Mullainathan, and Manish Raghavan. Inherent trade-offs in the fair determination of risk scores. *arXiv preprint arXiv:1609.05807*, 2016. [3](#), [12](#)
- [18] Kenneth Holstein, Jennifer Wortman Vaughan, Hal Daumé III, Miro Dudik, and Hanna Wallach. Improving fairness in machine learning systems: What do industry practitioners need? In *Proceedings of the 2019 CHI conference on human factors in computing systems*, pages 1–16, 2019. [4](#), [8](#), [42](#)
- [19] Partnership on AI. Ai tenets. URL <https://www.partnershiponai.org/2016/09/industry-leaders-establish-partnership-on-ai-best-practices/>. [7](#)
- [20] Organisation for Economic Co-operation and Development. Oecd ai principles. URL <https://www.oecd.org/going-digital/ai/principles/>. [7](#)
- [21] European Union High level Expert Group. Ethics guidelines for trustworthy ai: Building trust in human-centric ai. URL <https://ec.europa.eu/futurium/en/ai-alliance-consultation/guidelines>. [7](#)
- [22] G20. G20 ministerial statement on trade and digital economy. URL <https://www.mofa.go.jp/files/000486596.pdf>. [7](#)
- [23] Microsoft. The future computed: Artificial intelligence and its role in society. URL https://news.microsoft.com/cloudforgood/_media/downloads/the-future-computed-english.pdf. [8](#)
- [24] Google. Ai at google: our principles. URL <https://www.blog.google/technology/ai/ai-principles/>. [8](#)
- [25] Anna Jobin, Marcello Ienca, and Effy Vayena. The global landscape of ai ethics guidelines. *Nature Machine Intelligence*, 1(9):389–399, 2019. [8](#)

- [26] Andrew Abbott. Professional ethics. *American journal of sociology*, 88(5): 855–885, 1983. [8](#)
- [27] Gabriel Abend. 5. the good of american society. In *The Moral Background*, pages 207–259. Princeton University Press, 2014. [8](#)
- [28] Luke Stark and Anna Lauren Hoffmann. Data is the new what? popular metaphors & professional ethics in emerging data culture. 2019. [8](#), [11](#)
- [29] Brent Mittelstadt. Principles alone cannot guarantee ethical ai. *Nature Machine Intelligence*, 1(11):501–507, 2019. [8](#)
- [30] Michael Veale, Max Van Kleek, and Reuben Binns. Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making. In *Proceedings of the 2018 chi conference on human factors in computing systems*, pages 1–14, 2018. [8](#)
- [31] Batya Friedman. Value-sensitive design. *interactions*, 3(6):16–23, 1996. [9](#)
- [32] Batya Friedman, Peter Kahn, and Alan Borning. Value sensitive design: Theory and methods. *University of Washington technical report*, (2-12), 2002. [9](#)
- [33] Keld Bodker, Finn Kensing, and Jesper Simonsen. *Participatory IT design: designing for business and workplace realities*. MIT press, 2009. [9](#)
- [34] Ronald K Mitchell, Bradley R Agle, and Donna J Wood. Toward a theory of stakeholder identification and salience: Defining the principle of who and what really counts. *Academy of management review*, 22(4):853–886, 1997. [9](#)
- [35] Batya Friedman, Peter H Kahn Jr, Jennifer Hagman, Rachel L Severson, and Brian Gill. The watcher and the watched: Social judgments about privacy in a public place. *Human-Computer Interaction*, 21(2):235–272, 2006. [9](#)
- [36] Lisa P Nathan, Batya Friedman, Predrag Klasnja, Shaun K Kane, and Jessica K Miller. Envisioning systemic effects on persons and society throughout interactive system design. In *Proceedings of the 7th ACM conference on Designing interactive systems*, pages 1–10, 2008. [9](#), [10](#)
- [37] Alexei Czeskis, Ivayla Dermendjieva, Hussein Yapit, Alan Borning, Batya Friedman, Brian Gill, and Tadayoshi Kohno. Parenting from the pocket: Value tensions and technical directions for secure and private parent-teen mobile safety. In *Proceedings of the sixth symposium on usable privacy and security*, pages 1–15, 2010. [9](#), [10](#)
- [38] Kari Edison Watkins, Brian Ferris, Yegor Malinovskiy, and Alan Borning. Beyond context-sensitive solutions: Using value-sensitive design to identify needed transit information tools. In *Urban Public Transportation Systems 2013*, pages 296–308. 2013. [9](#)

- [39] Batya Friedman, Ian Smith, Peter H Kahn, Sunny Consolvo, and Jaina Selawski. Development of a privacy addendum for open source licenses: Value sensitive design in industry. In *International Conference on Ubiquitous Computing*, pages 194–211. Springer, 2006. [9](#)
- [40] Jessica K Miller, Batya Friedman, Gavin Jancke, and Brian Gill. Value tensions in design: the value sensitive design, development, and appropriation of a corporation’s groupware system. In *Proceedings of the 2007 international ACM conference on Supporting group work*, pages 281–290, 2007. [9](#), [10](#)
- [41] Lisa P Nathan, Predrag V Klasnja, and Batya Friedman. Value scenarios: a technique for envisioning systemic effects of new technologies. In *CHI’07 extended abstracts on Human factors in computing systems*, pages 2585–2590, 2007. [10](#)
- [42] Jill Palzkill Woelfer, Amy Iverson, David G Hendry, Batya Friedman, and Brian T Gill. Improving the safety of homeless young people with mobile phones: Values, form and function. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1707–1716, 2011. [10](#)
- [43] Daisy Yoo, Alina Huldtgren, Jill Palzkill Woelfer, David G Hendry, and Batya Friedman. A value sensitive action-reflection model: evolving a co-design space with stakeholder and designer prompts. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 419–428, 2013. [10](#)
- [44] Tamara Denning, Alan Borning, Batya Friedman, Brian T Gill, Tadayoshi Kohno, and William H Maisel. Patients, pacemakers, and implantable defibrillators: Human values and security for wireless implantable medical devices. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 917–926, 2010. [10](#)
- [45] Batya Friedman and David Hendry. The envisioning cards: a toolkit for catalyzing humanistic and technical imaginations. In *Proceedings of the SIGCHI conference on human factors in computing systems*, pages 1145–1148, 2012. [10](#)
- [46] Stephanie Ballard, Karen M Chappell, and Kristen Kennedy. Judgment call the game: Using value sensitive design and design fiction to surface ethical concerns related to technology. In *Proceedings of the 2019 on Designing Interactive Systems Conference*, pages 421–433, 2019. [10](#), [16](#), [35](#)
- [47] Chelsea Barabas, Madars Virza, Karthik Dinakar, Joichi Ito, and Jonathan Zittrain. Interventions over predictions: Reframing the ethical debate for actuarial risk assessment. In *Conference on Fairness, Accountability and Transparency*, pages 62–76. PMLR, 2018. [10](#)
- [48] Ifeoma Ajunwa and Daniel Greene. Platforms at work: Automated hiring platforms and other new intermediaries in the organization of work. In *Work and labor in the digital age*. Emerald Publishing Limited, 2019. [10](#)

- [49] Nripsuta Ani Saxena. Perceptions of fairness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 537–538, 2019. [11](#)
- [50] Nripsuta Ani Saxena, Karen Huang, Evan DeFilippis, Goran Radanovic, David C Parkes, and Yang Liu. How do fairness definitions fare? examining public attitudes towards algorithmic definitions of fairness. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 99–106, 2019. [11](#)
- [51] Reuben Binns. Fairness in machine learning: Lessons from political philosophy. In *Conference on Fairness, Accountability and Transparency*, pages 149–159. PMLR, 2018. [11](#)
- [52] Ben Hutchinson and Margaret Mitchell. 50 years of test (un) fairness: Lessons for machine learning. In *Proceedings of the Conference on Fairness, Accountability, and Transparency*, pages 49–58, 2019. [11](#)
- [53] Derek Leben. Normative principles for evaluating fairness in machine learning. In *Proceedings of the AAAI/ACM Conference on AI, Ethics, and Society*, pages 86–92, 2020. [11](#), [12](#)
- [54] Cynthia Dwork, Moritz Hardt, Toniann Pitassi, Omer Reingold, and Richard Zemel. Fairness through awareness. In *Proceedings of the 3rd innovations in theoretical computer science conference*, pages 214–226, 2012. [11](#), [17](#)
- [55] Matt J Kusner, Joshua R Loftus, Chris Russell, and Ricardo Silva. Counterfactual fairness. *arXiv preprint arXiv:1703.06856*, 2017. [11](#), [17](#)
- [56] Pratik Gajane and Mykola Pechenizkiy. On formalizing fairness in prediction with machine learning. *arXiv preprint arXiv:1710.03184*, 2017. [11](#)
- [57] Google Research. What if tool. URL <https://pair-code.github.io/what-if-tool/>. [12](#)
- [58] Pedro Saleiro, Benedict Kuester, Loren Hinkson, Jesse London, Abby Stevens, Ari Anisfeld, Kit T Rodolfa, and Rayid Ghani. Aequitas: A bias and fairness audit toolkit. *arXiv preprint arXiv:1811.05577*, 2018. [12](#)
- [59] Michael A Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. Co-designing checklists to understand organizational challenges and opportunities around fairness in ai. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pages 1–14, 2020. [12](#)
- [60] DrivenData. Deon: An ethics checklist for data scientists. URL <http://deon.drivendata.org/>. [12](#)
- [61] Ben Shneiderman. Universal usability: A research agenda for human-computer interaction to empower every citizen. In Rae A. Earnshaw, Richard A. Guedj, Andries van Dam, and John A. Vince, editors, *Frontiers of Human-Centered Computing, Online Communities and Virtual Environments*, pages 179–189. Springer, 2001. [16](#), [32](#)

- [62] Nina Grgic-Hlaca, Muhammad Bilal Zafar, Krishna P Gummadi, and Adrian Weller. The case for process fairness in learning: Feature selection for fair decision making. In *NIPS Symposium on Machine Learning and the Law*, volume 1, page 2, 2016. [17](#)
- [63] Sam Corbett-Davies, Emma Pierson, Avi Feller, Sharad Goel, and Aziz Huq. Algorithmic decision making and the cost of fairness. In *Proceedings of the 23rd acm sigkdd international conference on knowledge discovery and data mining*, pages 797–806, 2017. [17](#)
- [64] Moritz Hardt, Eric Price, and Nathan Srebro. Equality of opportunity in supervised learning. *arXiv preprint arXiv:1610.02413*, 2016. [17](#)
- [65] Richard Berk, Hoda Heidari, Shahin Jabbari, Michael Kearns, and Aaron Roth. Fairness in criminal justice risk assessments: The state of the art. *Sociological Methods & Research*, page 0049124118782533, 2018. [17](#)
- [66] Golnoosh Farnadi, Behrouz Babaki, and Lise Getoor. Fairness in relational domains. In *Proceedings of the 2018 AAAI/ACM Conference on AI, Ethics, and Society*, pages 108–114, 2018. [17](#)
- [67] Ying Shu, Jiehuang Zhang, and Han Yu. Fairness in design: A tool for guidance in artificial intelligence design. In *Proceedings of the 13th International Conference on Social Computing and Social Media (SCSM’21)*, 2021. [35](#)