


Fairness in Design: A Framework for Facilitating Ethical Artificial Intelligence Designs

Jiehuang Zhang^{1,2} , Ying Shu¹, and Han Yu¹

ABSTRACT

As Artificial Intelligence (AI) and Digital Transformation (DT) technologies become increasingly ubiquitous in modern society, the flaws in their designs are starting to attract attention. AI models have been shown to be susceptible to biases in the training data, especially against underrepresented groups. Although an increasing call for AI solution designers to take fairness into account, the field lacks a design methodology to help AI design teams of members from different backgrounds brainstorm and surface potential fairness issues during the design stage. To address this problem, we propose the Fairness in Design (FID) framework to help AI software designers surface and explore complex fairness-related issues, that otherwise can be overlooked. We explore literature in the field of fairness in AI to narrow down the field into ten major fairness principles, which assist designers in brainstorming around metrics and guide thinking processes about fairness. FID facilitates discussions among design team members, through a game-like approach that is based on a set of prompt cards, to identify and discuss potential concerns from the perspective of various stakeholders. Extensive user studies show that FID is effective at assisting participants in making better decisions about fairness, especially complex issues that involve algorithmic decisions. It has also been found to decrease the barrier of entry for software teams, in terms of the pre-requisite knowledge about fairness, to address fairness issues so that they can make more appropriate related design decisions. The FID methodological framework contributes a novel toolkit to aid in the design and conception process of AI systems, decrease barriers to entry, and assist critical thinking around complex issues surrounding algorithmic systems. The framework is integrated into a step-by-step card game for AI system designers to employ during the design and conception stage of the life-cycle process. FID is a unique decision support framework for software teams interested to create fairness-aware AI solutions.

KEYWORDS

design methodology; complex system; digital transformation; ethical Artificial Intelligence (AI); fairness; software design

As the pace of the advancement of Artificial Intelligence (AI) technologies increases, it has become intertwined with our everyday lives^[1-4]. Alongside the Fourth Industrial Revolution, automation enabled by digital transformation such as AI systems can initiate a new paradigm in the way we work and live^[5]. Algorithmic tools are gradually replacing human decision-making. This is especially important implications in life-critical fields like medicine and heavy industries. Thus, we must consider desirable traits that AI shall embody and chart a course towards a paradigm of ethical AI^[6].

Currently, the key performance metrics for AI systems are about their effectiveness and efficiency. Most AI development teams are focused on these metrics. However, in recent times, we have been increasingly exposed to evidence that AI systems can be vulnerable to bias and fairness issues. In 2019, a large-scale risk-prediction algorithm in health care was found to be less likely to identify African Americans for intensive care management due to faulty metrics^[7]. A high-profile investigation in 2016 surrounding the recidivism software Correctional Offender Management Profiling for Alternative Sanctions (COMPAS) found that the algorithm falsely predicted a high recidivism rate for African American offenders^[8]. We now understand that biases in training data can negatively affect AI solutions, especially when underrepresented minority groups are involved^[9,10]. As a result of

these negative instances of fairness issues in AI systems, software development teams must identify and address fairness-related issues in their designs. As a community, we should consider integrating ethical values early in the design and conception stage^[11,12].

Digital Transformation (DT) has contributed to the increased attention of integrating ethical values to the design and conception of AI products and services. Reference [13] reviewed numerous works to build a framework that foregrounds DT as a process where technologies create disruptions. This process helps organisations tweak their value creation paths, improve structural changes, and remove barriers. This situation closely relates to the ethical AI life cycle where software teams aim to reduce the barrier to entry and discover and resolve potential undesirable issues or outcomes before they happen. Lee et al.^[14] presented a novel machine-learning system for topic modelling to review and analyse advanced DT technologies. Existing literature was automatically classified into several topics for further investigation^[15]. Lee et al.^[16] created an AI model using data science and reinforcement techniques to forecast pricing and raw material procurement in the petrochemical industry, thereby enabling business process automation. Methodologies to enable DT such as these frameworks are critical to continuous improvement and enhancing business core competitiveness.

¹ School of Computer Science and Engineering, Nanyang Technological University, Singapore 639798, Singapore

² Alibaba-NTU Singapore Joint Research Institute, Singapore 637335, Singapore

Address correspondence to jiehuang.zhang@e.ntu.edu.sg

© The author(s) 2023. The articles published in this open access journal are distributed under the terms of the Creative Commons Attribution 4.0 International License (<http://creativecommons.org/licenses/by/4.0/>).

The major challenges hindering an AI solution design team during the brainstorming stage concerning fairness issues are twofold.

(1) Diverse fairness notions: Fairness is complex, thus it can have diverse meanings and definitions in different contexts. More than a dozen notions for fairness exist in the current literature^[17]. Furthermore, statistical incompatibilities are observed when exploring these fairness metrics^[18, 19]. Hence, software development teams have trouble grasping these different notions and overcoming blind spots in the design process^[20].

(2) Diverse stakeholders and application scenarios: As AI solutions can be designed for any application scenario involving any stakeholders type, they may prioritise different notions of fairness. This situation results in additional challenges for an AI solution team to identify and consider the importance of potential fairness issues in their design.

Currently, there is a lack of a methodological framework for software development teams to identify fairness-related issues and incorporate mitigating designs into their AI solutions. This limitation hinders discussions about such issues by members of an AI solution team who are often from diverse backgrounds and may have different levels of familiarity with notions of fairness.

In this paper, we propose a methodological framework called Fairness in Design (FID) to bridge this gap. It aims to help software design teams systemically consider fairness issues in AI systems, by lowering the barrier to entry as well as eliciting critical thinking. To address the first challenge, we develop a set of reference cards for 10 of the important notions of fairness^[17] to provide a quick guide for the design team. The format of the cards comprises a simple definition of the fairness notion, a simple equation to explain its statistical implications, and a simple application scenario to illustrate its implication on how the AI solution should make decisions. To address the second challenge, FID provides a game-like workflow of activities around the reference cards to guide the team lead on how to facilitate the discussions and guide team members on how to contribute to the discussions in an open-ended manner.

The main contributions presented in this paper are threefold. Firstly, we explore the state of the field of fairness in AI, highlighting the need for methodological frameworks to facilitate design processes. Secondly, we propose a methodology called FID to fill the gaps in the state of the field, in the process lowering the barrier of entry and eliciting critical thinking from AI design teams. Thirdly, we integrate the FID methodology into a multiplayer card game that guides design teams through systemic step-by-step workflows for a streamlined design process.

In Section 1, a discussion of related works in the field of fairness in AI is provided. We list the core fairness principles that already exist in literature and narrow them down for use in the methodological tool. In Section 3, we highlight a user study involving 24 AI technology professionals to evaluate FID effectiveness. The results show that FID is effective at assisting participants in making better decisions about fairness, especially in complex issues that involve algorithmic decisions. It has also been found to decrease the barrier of entry for software teams to address fairness issues so that they can make better, more appropriate design decisions relating to fairness concerns in their application. We also discuss the limitations of our user research. In Section 4, we conclude the study and identify several promising future directions for our work.

1 Related Work

The AI research community has spent significant effort to

formulate notions of fairness mathematically to support algorithmic fairness research^[21–23]. Dwork et al.^[24] and Kusner et al.^[25] divided the notions of fairness into two main categories: individual fairness and group fairness. Individual fairness revolves around one person, and requires that people with similar attributes receive similar outcomes from AI decision-making^[24]. Here we briefly list the statistical definitions under individual fairness, which are used in our user study. Individual fairness notions include: (1) Fairness Through Awareness^[24], (2) Fairness Through Unawareness^[26], (3) Counterfactual Fairness^[25], and (4) Fairness in Relational Domain. Group fairness, on the other hand, requires that different groups to be treated equally^[24]. There are six statistical definitions under group fairness: (1) Conditional Statistical Parity^[27], (2) Demographic Parity^[23], (3) Equal Opportunity^[28], (4) Equalised Odds^[28], (5) Test Fairness^[18], and (6) Treatment Equality^[29].

Madaio et al.^[30] proposed a co-design checklist to leverage industry practitioners' experience for designing fairness-aware AI. They highlighted that despite organisations publishing high-level principles to guide the ethical development of AI products, there had been a disconnect between intention and execution. The complexity and abstract nature of AI ethics increase the difficulty of operationalizing it for practitioners. Thus, the proposed ethics checklists in these organisations enable the co-designing of ethical AI with active participation from the practitioners working on AI products. However, these checklists do not provide actionable frameworks of guidance to help an AI solution design team to organize their brainstorming activities to uncover fairness-related issues specific to their application scenarios. Another toolkit—Fairlearn—allows developers to improve the fairness of their AI systems^[31]. It aims to mitigate fairness-related harms by framing the fairness issue as a socio-technical problems. However, Fairlearn is limited to addressing only unfairness in classification and regression models. In addition, it is not designed for the team brainstorming stage, but more as a mitigation tool when fairness issues emerge during later stages of development.

Existing ethical AI design methodologies are mainly derived from the Value Sensitive Design (VSD) methodology^[32]. It is a theoretical framework that focuses on integrating human values into the design of technologies in a principled and methodical way throughout the design processes. VSD can help users envision different situations from the perspectives of direct and indirect stakeholders. This analysis helps the users create a list of stakeholders to emulate, and explore how they are impacted by specific technology designs^[32]. Indirect stakeholders, people who do not use the target technological artefacts directly but are affected by their use, tend to be overlooked by system designers. This is despite them being as important as direct stakeholders, sometimes even more.

The two prominent VSD-inspired technology design methodological tools are (1) the Envisioning Cards^[33], and (2) the Judgment Call game^[34]. Envisioning Cards use specially designed cards to help stimulate critical thinking from the technology designers. They cover four main criteria: stakeholders, time, values, and pervasiveness. These cards allow designers to consider the long-term and potential systemic issues in technology design. As these cards are not specifically designed for AI ethics-related issues, the directions of discussion prompted by the envisioning cards do not provide adequate guidance for design teams to uncover AI ethics-related issues in their proposed AI solutions. To address this shortcoming, the Judgement Call game provides specialised set of cards to cover major AI ethics dimensions (e.g., fairness, privacy, explainability, security, and robustness). In

addition, it provides stakeholder cards and rating cards to further guide the brainstorming activities by individual team members around specific ethical values. Nevertheless, the methodology assumes that the participants are well-versed in the nuanced notions of the various ethical AI dimensions involved.

The proposed FID methodology addresses the limitations of envisioning cards in the Judgement Call game. Compared with these methodologies, FID provides a more nuanced and actionable guide on how to organise team discussions on specific notions of fairness in order to prioritise their project development resources.

2 Proposed FID Methodology

FID comprises two forms, the online tool and the physical card game. We first develop the physical card game for in-depth in person user studies but later completed the online tool in order to scale and reach more people in the COVID-19 pandemic. The cards are meant to be used by diverse users that require a methodology to uncover potential ethical problems in their AI products. The methodology works for many types of application domains, from e-commerce to computer vision based systems. Some background experience can enhance the effectiveness of FID, users are not required to have any prior knowledge to use the methodology, as it is designed to be inclusive to laypeople. The workflow for using FID to facilitate AI design team discussion is shown in Fig. 1.

We have consolidated the fairness definitions and notions from the literature into 10 principles, further categorised into group fairness and individual fairness. These principles are represented as cards*, to be used during the reflection of stakeholders' perspectives in the FID workflow.

(1) Team members must decide on the application domain that sets the environment for the rest of the session. This domain can be a fictional or a real-world product, depending on the needs of the team. As much detail as possible of the AI system or product should be included for FID to achieve a meaningful output because different considerations and trade-offs in different domains exist, which, in turn, can affect the decision-making process.

(2) Team members should pick an application card type that best describes their application domain. We adopt six application domain categories from Shneiderman's classification for usability motivation in Human-Computer Interaction literature: (a) life-critical systems; (b) industrial and commercial uses; (c) office,

*The complete set of cards can be found at: FID website: <https://sites.google.com/view/zhang-jiehuang/fairness-in-design>.

home, and entertainment; (d) exploratory and creative; (e) collaborative applications; and (f) socio-technical applications.

(3) The application domain has a set of stakeholders that are key to the analysis. We differentiate two stakeholder types, namely, direct and indirect stakeholder^[8], in this step. Direct stakeholders are people who use the AI system in the application domain directly, whereas indirect stakeholders are people who do not use the system directly but are impacted by its use. Each team member identifies one stakeholder group and brainstorms issues that they may face from the perspective of that stakeholder.

(4) Each team member draws a fairness principle card and applies it to the application domain. If the fairness principle card is inapplicable to the application domain, then he can choose to discard it and draw another. These cards illustrate fairness metrics according to the ten fairness principles previously discussed and how they are applied to AI systems.

(5) Each team member then applies the chosen fairness metric to the application domain and stimulates the potential problems or solutions that the stakeholder faces. We ask the question—What can go right or wrong for that stakeholder to elicit critical thinking from the team member? He then writes his thought process on the card.

(6) The team compiles all member responses and randomises them before reading them out loud. In this way, the responses are anonymous, and team members can be encouraged to be truthful. The team can discuss and evaluate the responses to decide if they are valid and worth addressing. Team members then rate fairness principles on a Likert scale to evaluate their importance to the application domain.

(7) Once the process is completed, the team can repeat the process by going back to Step (3) and conducting a new stakeholder analysis, or simply conclude the session.

Figure 2 shows an example user interface of FID. In this case, the user is at Step (3) of writing an envisioning review from his adopted stakeholder perspective. The given fairness notion, in this case, is demographic parity for which there is a card-based quick reference guide for the user to refer to. He is asked to think about what can go right and what can go wrong for the current design of his AI solution under the given context. In addition, to support ideation, the design input and decisions recorded by FID can be a useful source of information for tracing the origin of design issues. A demonstration video can be found at https://youtu.be/nnowNLss_wQ.

3 Empirical Evaluation

We conduct user studies to empirically evaluate the proposed FID framework.

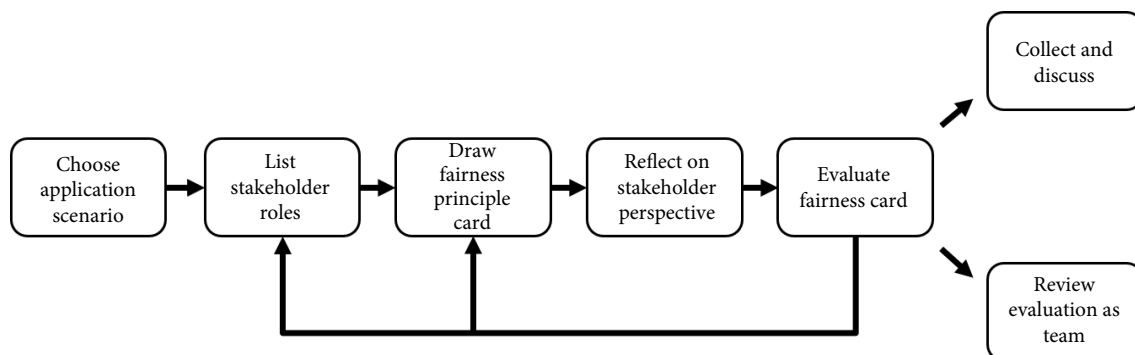


Fig. 1 FID workflow.

Fig. 2 An example user interface of FID (writing an envisioned stakeholder review for the given fairness notion—demographic parity).

3.1 Study design

A total of 24 participants (18 males and 6 females) are recruited for our user study. All participants are experienced researchers or engineers who are currently working or have previously worked on software systems involving AI. Additional recruitment criteria include the ability to understand the basic fairness concept and consent to be recorded. We recruit participants of a diverse range of ages to investigate how the methodology can impact usage by users with different seniority levels. Most participants fall into the 30s age group (Fig. 3), which is representative of the typical target users of FID.

Before the start of the actual user study, we instruct participants to complete a pre-study questionnaire through Google forms to understand how they prioritise ethical considerations in their AI solution development experience (if at all). As illustrated in Fig. 4, most participants indicate explainability and transparency as the most important criteria in this question. This result is understandable, as there has been increased attention in this area of machine learning in the research community. Although fairness

and bias rank third from last, participants do not mean that fairness is any less important. Rather, it is primarily because of the lack of support for the complex, multifaceted fairness concept to be considered during the design stage.

As shown in Fig. 5, most participants are working on AI solutions in the healthcare application domain. We include a redundancy test in the questionnaire by presenting the same statement twice, once in a positive way and once in a negative way. For example, we present the following positive statement, “I can compare across different possible strategies for addressing fairness issues in my application area”, and subsequently the negative statement, “I do not know how to evaluate different fairness solutions”. Using the redundancy check, we can detect invalid responses. Moreover, we ensure that the post-study questionnaire is completed within three days of the user study.

The three main hypotheses for this study are as follows:

H1: The FID methodology helps participants determine the fairness criteria that are the most relevant to their AI applications.

H2: The FID methodology helps participants surface fairness concerns in their AI applications.

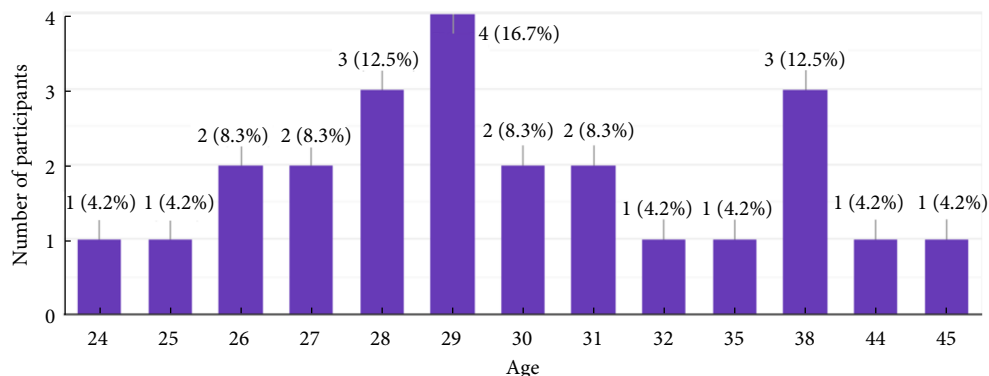


Fig. 3 Demographics of the participants.

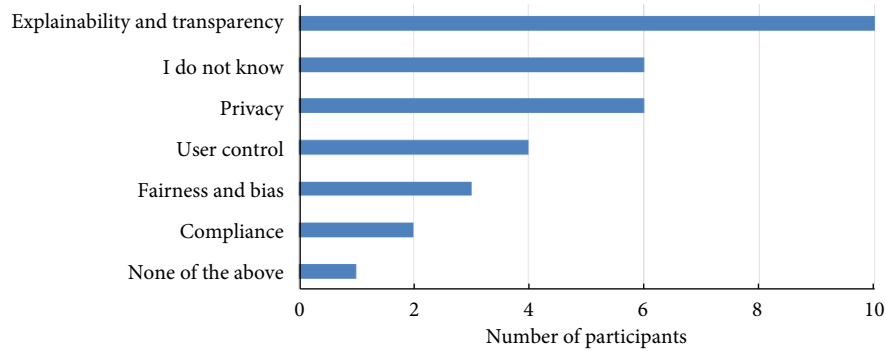


Fig. 4 Participants' ethical AI prioritisation.

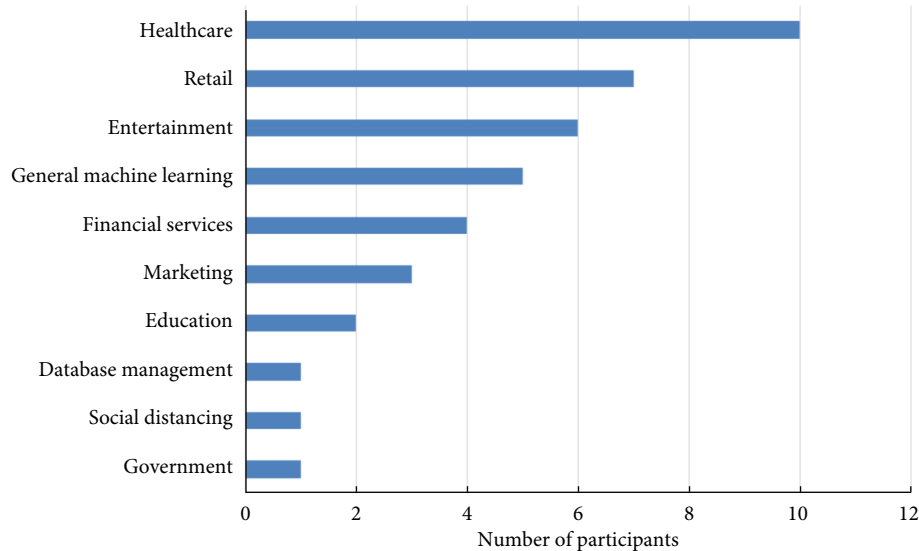


Fig. 5 Participants' experience developing AI applications.

H3: The FID methodology helps participants envision the perspectives from different stakeholders.

We design our pre- and post-study questionnaires for participants to conduct self-assessment of their fairness-related techniques. Each hypothesis is intended to assess the individual ability of participants to choose an applicable fairness solution, brainstorm and surface fairness concerns, and approach problems from a stakeholder perspective. Participants are required to rate their understanding of fairness problems on a Likert scale of 1 to 5. On the basis of questionnaire results, we perform a statistical analysis to evaluate the effectiveness of the three proposed hypotheses.

3.2 Result and analysis

In this section, we analyse the results from the empirical studies by presenting the findings related to each hypothesis.

3.2.1 Hypothesis 1

Hypothesis 1: The FID methodology helps participants determine the fairness criteria that are the most relevant to their AI applications.

Figure 6 illustrates the results from participant responses to questions related to H1. In the pre-study, participant responses follow a normal distribution centred on "Neutral". Thus, the distribution of their capabilities to make design decisions related to the fairness aspect of AI is typical of a population of AI solution designers. After using FID in the empirical study sessions, there is a significant increase in the number of participants who

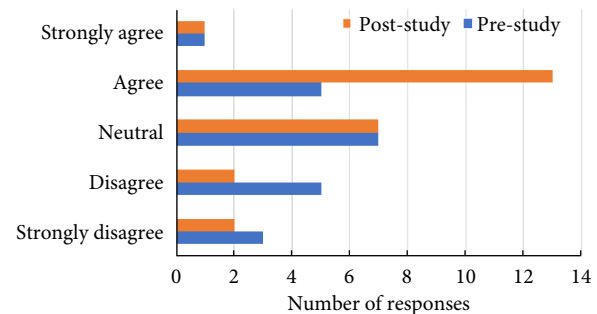


Fig. 6 Participants' self-reported capability of making design decisions related to fairness before and after using FID.

responded with "Agree", while the number of "Disagree" and "Strongly disagree" responses decreased. The number of "Strongly agree" and "Neutral" responses remains unchanged. Participants found FID to be useful for helping them think about design decisions related to incorporating fairness into AI solutions. As shown in Fig. 7, the participants' average response scores in the post-study are significantly higher than in the pre-study.

After conducting a students' t-test analysis of questionnaire results from H1, we conclude that the null hypothesis can only be rejected at the 90% confidence level.

3.2.2 Hypothesis 2

Hypothesis 2: The FID methodology helps participants surface fairness concerns in their AI applications. This hypothesis pertains

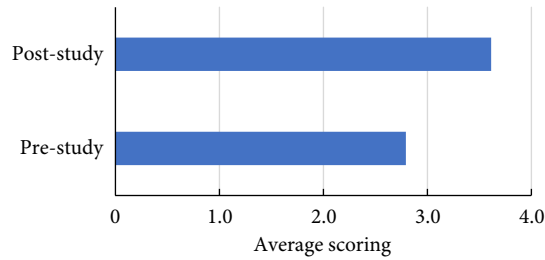


Fig. 7 Participants' average scoring for the pre- and post-studies for Hypothesis 1.

to the participants' self-assessment of their competency in discovering fairness issues ahead of time.

As shown in Fig. 8, we illustrate the results from the participants' responses about surfacing fairness concerns, focusing on H2. This question pinpoints the self-assessed ability of participants to identify ahead of time what type of fairness issues can arise in their specific application domain. As before participant responses are roughly normally distributed and centred on neutral. After using FID, a significant increase in the number of responses for "Agree" and "Strongly agree", as well as a corresponding decrease in the number of responses for "Disagree" and "Strongly disagree". The results indicate that FID is effective in assisting participants to surface potential fairness issues in their application domains.

For H2, we find that the questionnaire response averages increased by more than 0.5 in the post-studies compared to the pre-studies (Fig. 9). After conducting a students' t-test, we are able to reject the null hypothesis at 95% confidence level.

3.2.3 Hypothesis 3

Hypothesis 3: The FID methodology helps participants envision the perspectives from different stakeholders. We conceptualised this hypothesis to assess the ability of participants to stimulate thinking in the perspective of other relevant groups of people.

Figure 10 highlights the results from the participant responses on thinking from the perspective of stakeholders. This question focuses on H3 and challenges the participants to visualise and think from the perspective of two types of stakeholders, direct stakeholders such as the end-user and legal and marketing staff,

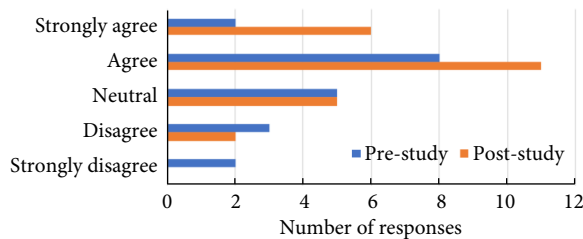


Fig. 8 Participants' self-reported capability of surfacing fairness concerns before and after using FID.

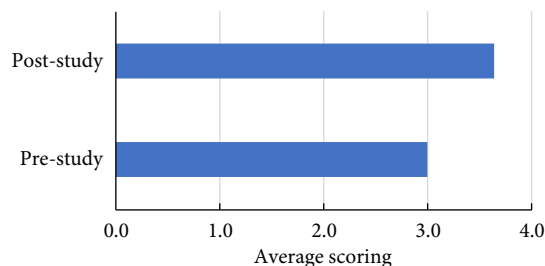


Fig. 9 Participants' average scoring for the pre- and post-studies for Hypothesis 2.

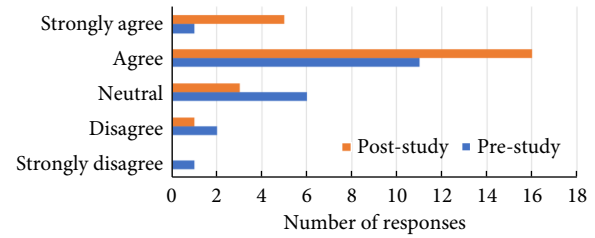


Fig. 10 Participants' self-reported capability of thinking from stakeholders' perspective before and after using FID.

and indirect stakeholders such as families of end users. The proportion of "Disagree" and "Strongly disagree" greatly decreased and the participants changed their answers to "Agree" and "Strongly agree". Specifically, the frequency of "Strongly agree" increased from 1 in pre-study to 5 in post-study. The methodology can greatly improve the self-assessed ability of participants to stimulate stakeholder perspective, which is a valuable skill set in AI development teams.

For H3, we find that the questionnaire response averages increased by more than 0.5 in the post-studies compared to the pre-studies (Fig. 11). Conducting a students' t-test analysis of questionnaire results from Hypothesis 3, the null hypothesis can only be rejected at 90% confidence level.

3.3 Discussion

FID is found effective in promoting conversations about surfacing fairness concerns of AI projects. However, it is only suitable for projects that are in the design stage, and the reason is that fairness metrics must be built into the system early before other tasks. Testing for fairness can be performed in the later stages of the AI life cycle; system designers can use the methodology to extrapolate outcomes or impacts of using certain fairness notions and principles. Depending on the specific application domain, the team may have experienced the typical performance of the algorithmic system and how to introduce fairness considerations in the operations. The FID methodology can be enhanced to create a significant impact on the entire AI pipeline process, and this direction can be promising for future work. We highlight one of the important questions in our questionnaire regarding making decisions about fairness in their projects. We find that a significant number of participants are more confident after the study, and 14 "Agree" or "Strongly agree" in the post-study compared with 6 in the pre-study.

Moreover, given that fairness is a complex and multifaceted topic in AI development, most system designers are unfamiliar with it or see it as an unnecessary trade-off for performance or efficiency. That is, most AI developers are unwilling to forgo the reduction in their algorithm performance unless a specific requirement is requested. This response is expected in most engineers and practitioners in the community, but we believe the direction of the field is that, eventually, teams can deliver a system

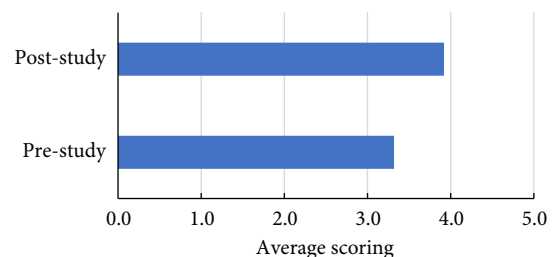


Fig. 11 Participants' average scoring for the pre- and post-studies for Hypothesis 3.

that minimises the compromise on both values. The aim of the FID methodology is to initiate a framework for AI design teams to embark on the vision of a balanced and fair AI system without compromising on algorithm performance or efficiency.

Another consideration in this work is that fairness can be perceived or defined dynamically depending on contexts and application domains. For example, in a large group of employees in a multinational corporation, what is fair for the senior management may not be received the same way as lower-ranked employees. With the rise of the gig economy, the dynamic pairing of workers and jobs means that market conditions can change drastically within a short period. In these volatile application domains, the FID methodology can be adapted to vary several underlying assumptions to enhance the effectiveness of the entire process. For example, over the course of multiple user studies, we discover that target systems specifically collect data on protected and unprotected individuals/groups and then subsequently make decisions on the basis of these data. We also tweak the process numerous times in response to the feedback obtained. For example, we realise the need to provide the option of discarding irrelevant fairness principle cards to participants because some principles are more important to the application scenario than others.

3.4 Limitation

Our user study only includes 24 recruited participants. Although this number is sufficient for observing some recurring conclusions about FID, some flaws in the toolkit only become apparent in large-scale tests. Testing and improving the methodology and then scaling it up to conduct testing on a large sample size of participants take significant amounts of time and effort. We are planning for a larger-scale study with only online participants to conduct a more in-depth evaluation of the FID methodology. Self-reported preferences often do not align well with actual user behaviours^[35]. Whether findings from existing and past works contribute to significant improvements in our methodology remains unclear. As the fairness notions elaborated to users in FID are adapted from AI literature, the FID tool is useful for algorithms that collect data to train machine learning models. We also consider minority groups or communities that are at high risk of being intentionally or unintentionally overlooked.

4 Conclusion and Future Work

In this study, we identify the research gaps in ethical AI literature and highlight the need for a methodological tool that allows for deep analyses of potential ethical issues. After exploring fairness and VSD literature, we identify significant fairness principles and develop a methodology to assist software designers in understanding fairness issues and then create strategies to address them and overcome biases. Our conversations with product teams reveal that they usually view fairness as an afterthought, and many barriers to considering fairness issues in their teams exist. We design this FID methodology to have a low barrier to entry and thus it is easy for laypeople to use effectively. Our target audience of AI product teams finds the methodology effective and useful to explore and make fair decisions in their application domains. With this methodology, we hope to inspire others in the ethical AI research community to construct more methodological frameworks that assist AI product teams in considering ethics in their AI systems. To the best of our knowledge, FID is the first technical tool to facilitate AI solution development teams to incorporate fairness into their designs. Empirical results from our

user study involving 24 AI solution developers show that FID can improve design teams' understanding of fairness concepts and is perceived to be useful for their projects.

In subsequent research, we will be looking to scale the usage of our FID methodology to a larger user base such that more improvements and tweaks can be made. Leveraging on circumstance synchronisation, stakeholders, and technologies in DT, we aim to optimise complex concepts and processes for the benefit of laypeople. We also plan to extend the methodology to other ethical values, such as privacy and explainability, and create a unified methodological framework that is the go-to tool for considering ethics in AI.

Acknowledgment

This work was supported in part by Nanyang Technological University, Nanyang Assistant Professorship (NAP); Alibaba Group through Alibaba Innovative Research (AIR) Program and Alibaba-NTU Singapore Joint Research Institute (JRI) (No. Alibaba-NTU-AIR2019B1), Nanyang Technological University, Singapore; the National Research Foundation, Singapore under its AI Singapore Programme (AISG Award) (No. AISG2-RP-2020-019); the RIE 2020 Advanced Manufacturing and Engineering (AME) Programmatic Fund (No. A20G8b0102), Singapore; the Joint SDU-NTU Centre for Artificial Intelligence Research (C-FAIR); and Future Communications Research and Development Programme (No. FCP-NTU-RG-2021-014). Any opinions, findings, and conclusions or recommendations expressed in this material are those of the author(s) and do not reflect the views of the National Research Foundation, Singapore.

Dates

Received: 4 May 2022; Revised: 27 September 2022; Accepted: 28 September 2022

References

- [1] K. Schwab, *The Fourth Industrial Revolution*. New York, NY, USA: Currency, 2017.
- [2] Y. Zheng, H. Yu, L. Cui, C. Miao, C. Leung, and Q. Yang, SmartHS: An AI platform for improving government service provision, in *Proc. 30th Innovative Applications of Artificial Intelligence Conference*, New Orleans, LA, USA, 2018, pp. 7704–7711.
- [3] X. Guo, B. Li, H. Yu, and C. Miao, Latent-optimized adversarial neural transfer for sarcasm detection, in *Proc. 2021 Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT'21)*, Online, 2021, pp. 5394–5407.
- [4] M. Lei, Z. Rao, H. Wang, Y. Chen, L. Zou, and H. Yu, Maceral groups analysis of coal based on semantic segmentation of photomicrographs via the improved U-net, *Fuel*, vol. 294, p. 120475, 2021.
- [5] S. Makridakis, The forthcoming artificial intelligence (AI) revolution: Its impact on society and firms, *Futures*, vol. 90, pp. 46–60, 2017.
- [6] H. Yu, Z. Shen, C. Miao, C. Leung, V. R. Lesser, and Q. Yang, Building ethics into artificial intelligence, in *Proc. 27th International Joint Conference on Artificial Intelligence (IJCAI'18)*, Stockholm, Sweden, 2018, pp. 5527–5533.
- [7] Z. Obermeyer, B. Powers, C. Vogeli, and S. Mullainathan, Dissecting racial bias in an algorithm used to manage the health of populations, *Science*, vol. 366, no. 6464, pp. 447–453, 2019.
- [8] J. Angwin, J. Larson, S. Mattu, and L. Kirchner, Machine bias: There's software used across the country to predict future criminals. And it's biased against blacks, *ProPublica*, <https://www.propublica.com>.

- [org/article/machine-bias-risk-assessments-in-criminal-sentencing](https://www.cs.dartmouth.edu/~ccpalmer/teaching/cs89/Resources/Papers/AIs%20White%20Guy%20Problem%20-%20NYT.pdf), 2016.
- [9] K. Crawford, Artificial intelligence's white guy problem, *The New York Times*, <https://www.cs.dartmouth.edu/~ccpalmer/teaching/cs89/Resources/Papers/AIs%20White%20Guy%20Problem%20-%20NYT.pdf>, 2016.
- [10] A. Yapo and J. Weiss, Ethical implications of bias in machine learning, in *Proc. 51st Hawaii International Conference on System Sciences*, Hawaii, HI, USA, 2018, pp. 5365–5372.
- [11] J. Havens, *Heartificial Intelligence: Embracing Our Humanity to Maximize Machines*. New York, NY, USA: Jeremy P. Tarcher/Penguin, 2016.
- [12] B. Friedman, D. G. Hendry, and A. Borning, A survey of value sensitive design methods, *Foundations and Trends in Human-Computer Interaction*, vol. 11, no. 2, pp. 63–125, 2017.
- [13] G. Vial, Understanding digital transformation: A review and a research agenda, *The Journal of Strategic Information Systems*, vol. 28, no. 2, pp. 118–144, 2019.
- [14] C. -H. Lee, C. -L. Liu, A. J. Trappey, J. P. T. Mo, and K. C. Desouza, Understanding digital transformation in advanced manufacturing and engineering: A bibliometric analysis, topic modeling and research trend discovery, *Advanced Engineering Informatics*, vol. 50, p. 101428, 2021.
- [15] C. -H. Lee, A. J. Trappey, C. -L. Liu, J. P. T. Mo, and K. C. Desouza, Design and management of digital transformations for value creation, *Advanced Engineering Informatics*, vol. 52, p. 101547, 2022.
- [16] C. -Y. Lee, B. -J. Chou, and C. -F. Huang, Data science and reinforcement learning for price forecasting and raw material procurement in petrochemical industry, *Advanced Engineering Informatics*, vol. 51, p. 101443, 2022.
- [17] N. Mehrabi, F. Morstatter, N. Saxena, K. Lerman, and A. Galstyan, A survey on bias and fairness in machine learning, arXiv preprint arXiv: 1908.09635, 2019.
- [18] A. Chouldechova, Fair prediction with disparate impact: A study of bias in recidivism prediction instruments, *Big Data*, vol. 5, no. 2, pp. 153–163, 2017.
- [19] J. Kleinberg, S. Mullainathan, and M. Raghavan, Inherent trade-offs in the fair determination of risk scores, arXiv preprint arXiv: 1609.05807, 2016.
- [20] K. Holstein, J. W. Vaughan, H. Daumé, M. Dudik, and H. Wallach, Improving fairness in machine learning systems: What do industry practitioners need? in *Proc. 2019 CHI Conference on Human Factors in Computing Systems*, Glasgow, UK, 2019, pp. 1–16.
- [21] R. Binns, Fairness in machine learning: Lessons from political philosophy, *Proceedings of Machine Learning Research*, vol. 81, pp. 149–159, 2018.
- [22] B. Hutchinson and M. Mitchell, 50 years of test (un)fairness: Lessons for machine learning, in *Proc. Conference on Fairness, Accountability, and Transparency*, Atlanta, GA, USA, 2019, pp. 49–58.
- [23] S. Verma and J. Rubin, Fairness definitions explained, in *Proc. 2018 IEEE/ACM International Workshop on Software Fairness (FairWare)*, Gothenburg, Sweden, 2018, pp. 1–7.
- [24] C. Dwork, M. Hardt, T. Pitassi, O. Reingold, and R. Zemel, Fairness through awareness, in *Proc. 3rd Innovations in Theoretical Computer Science Conference*, Cambridge, MA, USA, 2012, pp. 214–226.
- [25] M. J. Kusner, J. R. Loftus, C. Russell, and R. Silva, Counterfactual fairness, arXiv preprint arXiv: 1703.06856, 2017.
- [26] N. Grgic-Hlaca, M. B. Zafar, K. P. Gummadi, and A. Weller, The case for process fairness in learning: Feature selection for fair decision making, presented at Symposium on Machine Learning and the Law at the 29th Conference on Neural Information Processing Systems, Barcelona, Spain, 2016.
- [27] S. Corbett-Davies, E. Pierson, A. Feller, S. Goel, and A. Huq, Algorithmic decision making and the cost of fairness, in *Proc. 23rd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, Halifax, Canada, 2017, pp. 797–806.
- [28] M. Hardt, E. Price, and N. Srebro, Equality of opportunity in supervised learning, arXiv preprint arXiv: 1610.02413, 2016.
- [29] R. Berk, H. Heidari, S. Jabbari, M. Kearns, and A. Roth, Fairness in criminal justice risk assessments: The state of the art, *Sociological Methods & Research*, vol. 50, no. 1, pp. 3–44, 2021.
- [30] M. A. Madaio, L. Stark, J. W. Vaughan, and H. Wallach, Co-designing checklists to understand organizational challenges and opportunities around fairness in AI, in *Proc. 2020 CHI Conference on Human Factors in Computing Systems (CHI'20)*, Honolulu, HI, USA, 2020, pp. 1–14.
- [31] S. Bird, M. Dudik, R. Edgar, B. Horn, R. Lutz, V. Milan, M. Sameki, H. Wallach, and K. Walker, Fairlearn: A toolkit for assessing and improving fairness in AI, Tech. Rep. MSR-TR-2020-32, Microsoft, Redmond, WA, USA, 2020.
- [32] B. Friedman, Value-sensitive design, *Interactions*, vol. 3, no. 6, pp. 16–23, 1996.
- [33] B. Friedman and D. Hendry, The envisioning cards: A toolkit for catalyzing humanistic and technical imaginations, in *Proc. SIGCHI Conference on Human Factors in Computing Systems*, Austin, TX, USA, 2012, pp. 1145–1148.
- [34] S. Ballard, K. M. Chappell, and K. Kennedy, Judgment call the game: Using value sensitive design and design fiction to surface ethical concerns related to technology, in *Proc. 2019 on Designing Interactive Systems Conference*, San Diego, CA, USA, 2019, pp. 421–433.
- [35] E. Zell and Z. Krizan, Do people have insight into their abilities? A metasynthesis, *Perspectives on Psychological Science*, vol. 9, no. 2, pp. 111–125, 2014.