A Design Methodology for Incorporating Privacy Preservation into AI Systems

Jiehuang Zhang[1,2] and **Han Yu**[2] ,
[1]School of Computer Science and Engineering, Nanyang Technological University
[2]Alibaba-NTU Singapore Joint Research Institute
{jiehuang001, han.yu}@ntu.edu.sg

Abstract

Artificial Intelligence (AI) has brought about paradigm shifts in how human societies work and live, from automating business processes to self-driving vehicles. With such tremendous impact comes ethical concerns. The privacy issue has been thrust to the forefront after multiple incidents involving well-known industry players. Although privacy preservation has been identified as a critical component of ethical AI, there is currently an absence of methodological tools to enable AI software teams to systematically surface and address privacy issues in the design and conception phase. We propose the Privacy-based Design (PbD) methodology to integrate privacy values early in the life cycle of AI products and services to address this gap. It allows AI software design teams to identify and analyse complex privacy issues in a systematic manner by guiding the envisioning of various scenarios. With PbD, we aim to reduce the barrier to entry, time and experience needed for AI practitioners to critically make well-thought-out decisions on incorporating privacy-preserving designs into AI solutions. User studies involving 29 participants found that the PbD methodology is useful and easy to use.

1 Introduction

Artificial Intelligence(AI) is a core part of the fourth industrial revolution [Schwab, 2017] and the digital age. It has enabled many advances in a plethora of fields such as healthcare [Tjoa and Guan, 2020], algorithmic crowdsourcing [Yu *et al.*, 2017] and autonomous driving [Zhang *et al.*, 2021]. The success of these AI technologies was made possible by the availability of big data generated in recent years, as well as novel machine learning techniques to achieve remarkable levels of performance. As new techniques facilitate the increasing impact of AI on everyday life [Makridakis, 2017], these technological breakthroughs allow the automation of tasks that brings many benefits.

Due to the far reaching impact of AI, there is a need to consider how it can adversely affect our societies due to potential ethical issues. Because of the improved capabilities of AI, many businesses and organizations are transferring more responsibility and autonomy to algorithmic systems. As a result, the possibility of mistakes or unintended side effects are more likely to happen [Amodei *et al.*, 2016]. For example, a significant event that brought privacy into the spotlight was Facebook's data breach in 2018 [Financial times, 2020] when the personal data of fifty million American voters from Facebook was gathered and then allegedly used by the political consultancy Cambridge Analytica. The incident raised questions on how giant technology companies can do more to protect their users' interests. To ensure that AI development benefits humanity as a whole, we must collectively monitor its advancement and guide its trajectory towards human-centred and ethical AI solution design [Croeser and Eckersley, 2019; Yu *et al.*, 2018].

Privacy preservation research in AI aims to address the question "What are the privacy challenges in Machine Learning (ML) and how can we solve them?" [Liu *et al.*, 2021]. Federated learning [Yang *et al.*, 2019b] is the primary approach adopted by this research field. Multiple survey papers have provided overviews of the state of this field from diverse perspectives [Liu *et al.*, 2021; Tan *et al.*, 2022; Lyu *et al.*, 2022; Zhang and Yu, 2022b; Shi *et al.*, 2023]. As more groundbreaking techniques to achieve privacy are discovered, we must also consider the need of integrating these principles early in the early stages of the AI software life cycle. However, the following challenges hinder design teams to incorporate considerations for privacy preservation into their AI solutions during the conceptualization phase:

1. **Diverse Privacy Notions and Privacy Preservation Techniques**: Privacy is a complex and multifaceted concept. It can have different definitions and requirements in different application scenarios. Furthermore, when privacy is prioritized, its requirements may impact other metrics such as performance and accuracy. Hence, AI solution design teams who are not well trained on this topic might be overwhelmed when trying to grasp the different notions and techniques during the AI product and service life cycle.

2. **Diverse Groups with Different Interests and Agendas in Various Domains**: To different stakeholders, different AI application scenarios may require different notions of privacy to be prioritized. This complexity poses

significant challenges to AI solution teams to allocate their limited resources to fulfil such requirements.

To address these challenges, we propose the Privacy-based Design (PbD) methodological framework. It is an extension of our previously proposed design methodologies for incorporating fairness [Shu *et al.*, 2021] and explainability [Zhang and Yu, 2022a] considerations into AI solutions. The objective of PbD aims to facilitate software teams to systematically analyse privacy issues during the conceptualization and brainstorming phase, by lowering the barrier to entry and eliciting systematic and deep thinking during team discussions. As an important part of constructing ethical AI products and services, the goal of the methodology is to create scaffolding for conversations among scientists and engineers, thereby enabling them to reach an optimal solution for privacy preservation in their AI solution designs. This is achieved by facilitating the process of brainstorming and investigating privacy requirements and topics surrounding the application domain and stimulating deep thinking from the shoes of various stakeholder communities. Through preliminary user studies involving 29 participants, we demonstrate that the proposed methodology is useful and easy to use.

2 Related Work

Privacy preservation is a vital part of making AI safe and beneficial for all. There is increasing public awareness about large companies compromising on data security and user privacy. There has been much backlash in response to these scandals, and many countries are improving their laws to address data privacy and security [Yang *et al.*, 2019a]. For example, the European Union (EU) instituted the General Data Protection Regulation (GDPR) in order to enhance the protection of public users' personal privacy and security [EuropeanUnion, 2016].

Newly emerging techniques in AI and machine learning (ML) continue to increase the intricacy of privacy preservation. The challenges and problems associated with making AI privacy respecting have been a central focus of the research community, given the time-sensitive nature of the problem before more government and regulatory laws are introduced to protect data. Such works can be further grouped into subcategories, depending on whether the techniques are meant for dataset or model protection, as well as whether ML techniques are used for offence or defence.

Most responsible AI methodologies and frameworks are influenced by the Value Sensitive Design (VSD) approach [Friedman *et al.*, 2017], which was developed in human-computer interaction (HCI) information systems design (ISD). VSD gives importance to the ethical values of both direct and indirect stakeholders and uses various methods to engage with diverse values based on the application. This allows designers to gain insights and integrate with other methodologies. The main workflow of VSD involves stimulating the perspectives of stakeholders and analyzing how their values are affected. Direct stakeholders directly use the AI product and are impacted, while indirect stakeholders are not users but are still affected.

VSD has delivered two exploratory card games, Judgement Call [Ballard *et al.*, 2019] and Envisioning Cards, to facilitate ethical AI design. Envisioning Cards encourage critical thinking about stakeholders, time, values, and motivation to consider systemic long-term problems. Judgement Call is a turn-based card game that AI development teams can use to identify moral problems in an AI product, using cards that focus on virtuous value, stakeholders, and review ratings to encourage experimental thinking. Based on the Judgement Call game design, the Fairness in Design (FID) [Shu *et al.*, 2021] and Explainability in Design (EID) [Zhang and Yu, 2022a] approaches have been proposed to provide more focused guidance for AI design teams on envisioning challenges and opportunities with regard to fairness and explainability, respectively. The proposed PbD approach extends FID and EID to provide support for incorporating privacy-preservation into AI solution designs.

3 Preliminaries

For this section, we have classified the techniques of privacy into the respective four categories as shown in Figure 1: 1) Attack and Threat Models, 2) Private Machine Learning Schemes, 3) Privacy Attacks, and 4) Machine Learning-enhanced Privacy Protection.

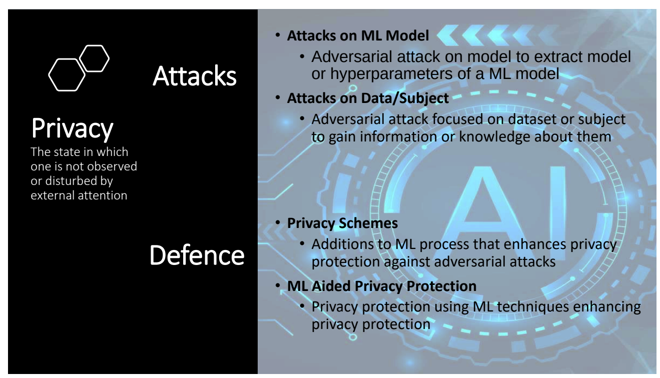

Figure 1: An overview of the main principles of privacy in AI

1. **Attacks on ML Model**: Adversarial attack on ML models to extract entire model or hyperparameters of an ML model. Examples include:
   (a) Model Extraction Attacks: Attacks aims to copy or "extract" an AI model on a high level, resulting in a function with parameters and coefficients that resembles the original model [Tramèr *et al.*, 2016]
   (b) Feature Estimation Attacks: Feature estimation attacks seek to estimate specific features or statistical properties of the training dataset. This type of attack is initiated through model inversion, power side-channel attacks or shadow model [Fredrikson *et al.*, 2015]
   (c) Membership Inference Attacks: Such an attack involves determining whether a particular data point belongs to the training dataset [Shokri *et al.*, 2017]
   (d) Model Memorization Attacks: Such an attack seeks to recover exact feature values on individual sam-

ples and involves stealing model parameters and coefficient values [Song *et al.*, 2017]

2. **Privacy Schemes**: A privacy-preserving scheme is a collection of techniques or algorithms that assist ML models to improve their defence against adversarial privacy attacks.
   (a) Encryption: Homomorphic Encryption applies a computation to encrypt data, allowing sensitive data to be used as a training dataset. However adds an order of magnitude to computation complexity [Bost *et al.*, 2014]
   (b) Obfuscation: Obfuscation mechanisms aim to reduce the precision of privacy attacks using the introduction of noise to the coefficients of the model [McPherson *et al.*, 2016]
   (c) Aggregation: Aggregation techniques involve multiple parties joining an ML scheme while at the same time aiming to hide their own datasets, to be applied during training or after training. Federated Learning (FL) is grouped in this section
3. **Attack and Threat Models**: Types of attack an adversary can employ to access a model's parameters.
   (a) Identification Attack: Specifies a user's details or identification on a shared dataset [Li *et al.*, 2016], when anonymization is reversed, the attack is called re-identification
   (b) Inference Attack: This type of attack's objective is to explore data to obtain information on a target [Nasr *et al.*, 2019]
   (c) Linkage Attack: The counter party's goal is to steal the subject's information by comparing or cross-referencing different datasets from the source
4. **ML Privacy Protection**: Preemptive privacy measures targeted at mitigating privacy risks
   (a) Risk Assessment Protection: Evaluate and predict the risk for users during the process of accessing and sharing information. Algorithms are employed to predict data streams to find risks and subsequently deploy countermeasures
   (b) Personal Privacy Management: Policy evaluation, user preference prediction, and management, of the behaviour of the user
   (c) Private Data Release: Disseminate datasets with a privacy assurance

4 The Privacy by Design Methodology

The Privacy by Design (PbD) Methodology takes the form of a tangible card game, which can be utilized by individuals with varying levels of expertise, ranging from novices to professionals. The purpose of this methodology is to encourage brainstorming and the identification of potential privacy issues within AI while remaining adaptable to different applications. The design team is given the freedom to dictate the dimensions of their intended domain application, making the methodology application agnostic.

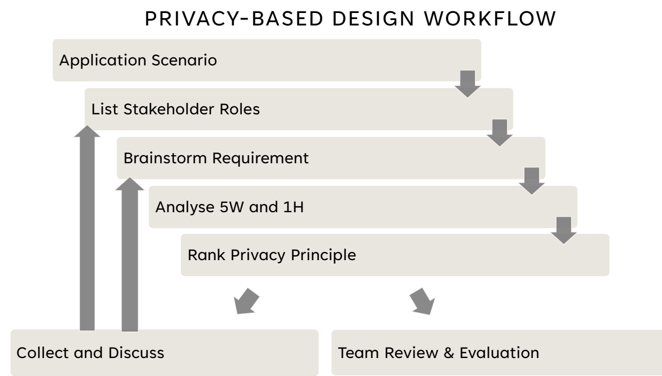

Figure 2: The Privacy-Based Design Workflow

The step-by-step framework for a software design team to employ PbD to encourage conversations in the software team surrounding privacy issues is illustrated in Figure 2. The systemic guide to using the methodology is discussed below:

1. Initially, the AI design team is required to select an application domain, which will provide the base environment for the purpose of the study. The environment, depending on the teams, can either be a genuine or imaginary setting, with a strong preference for a domain where privacy is a significant concern. It is ideal for team members to possess knowledge about the domain, enabling them to incorporate as many details as possible throughout the user study. This is particularly important as some domains may require specific considerations and compromises that can influence the usage of this methodological framework.
2. During Step Two, users are required to select a card from the classification system that corresponds to their context domain. The classification system used for this purpose is based on Shneiderman's work on usability motivation in the field of Human-AI Interaction (HCI) [Shneiderman and Hochheiser, 2001].
3. In Step Three of the PbD methodology, the group needs to conduct an exploratory analysis and recognize the stakeholders who play a vital role in the end-to-end AI pipeline. Direct stakeholders are those who frequently use the product or service, while indirect stakeholders are not the end-users but are still influenced by the deployment [Friedman *et al.*, 2017]. The team members are required to take on the perspective of a stakeholder and carry out an in-depth examination of the privacy details concerning that stakeholder. The PbD methodology presents several guiding questions to streamline the thought process. These critical thinking guides revolve around the who, what, when, where, why, and how of privacy-related topics. For instance:

After finishing all the steps in the methodology, the team can have an insightful and deeper appreciation of privacy concerns and their application domain. They can choose to further investigate a specific topic by brainstorming and discussing it in detail as a team. This process of delving deeper into a particular topic can help uncover any complex privacy issues that may have been missed otherwise.

The deliverables of the framework include the listed outputs:

1. Understanding and Selecting the privacy principles that are relevant for the environment.
2. Rank priorities of specialized requirements for the analysis of privacy measures.
3. Quantifying improvements in the privacy knowledge levels of methodology users.
4. Create a thinking guide to determine where to focus their attention and focus on during sprints

The methodology encourages a collaborative approach to AI design, with individuals stimulating the perspectives of direct and indirect stakeholders, and then conducting an in-depth exploration of privacy attributes. This approach helps ensure that all potential privacy concerns are addressed and that stakeholders' perspectives are taken into account in the design process. Using these outputs, the team can then use the insights and information gained to make informed decisions about improving their processes.

5 Empirical Evaluation

In this section, we discuss the process and analyse the results of our experiments with recruited participants to evaluate the proposed PbD Methodology and our hypotheses.

5.1 Study Design

We recruited 29 participants through the snowball sampling method. Our criteria for the qualification of the potential participants are such that they possessed experience being part of a team that worked on AI technologies. All of the participants were researchers, scientists or engineers that were able to understand privacy concepts in AI/ML, as well as consenting to be recorded and their insights published. We recruited participants from a diverse age range to investigate how the PbD Framework can affect end users of different levels of seniority. However, the majority of participants belong to the 20-30-year-old age group, as is the profile of the usual proposed users of the PbD framework.

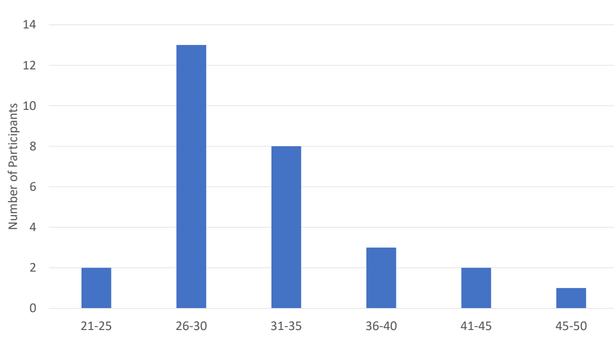

Figure 3: Demographics of the Participants.

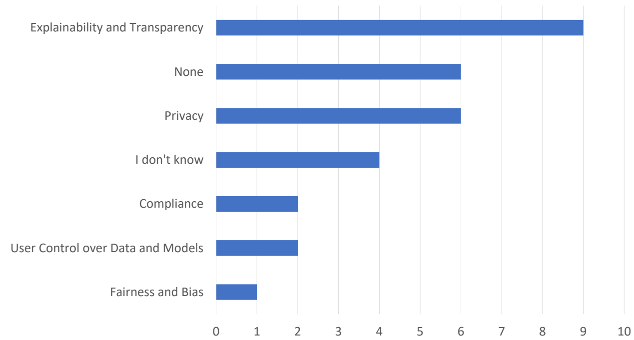

Figure 4: Participants' ethical AI prioritisation.

Before commencing the user study, we enquired about the order of priority of the type of responsible AI considerations in the AI product pipeline. According to Figure 5, nine users preferred explainability and transparency as their top considerations, while a significant portion of 6 participants indicates that none of the ethical values as part of the considerations for the software development pipeline. This observation was reflected in the many feedback from the users that performance and efficiency were greatly valued over ethical AI principles. Most of the initiatives in ethical AI tend to be a reaction to government compliance processes or regulatory pressure. While privacy may not be at the top of the priority list, there is a need for more toolkits to assist software design teams to enhance privacy in their workflows.

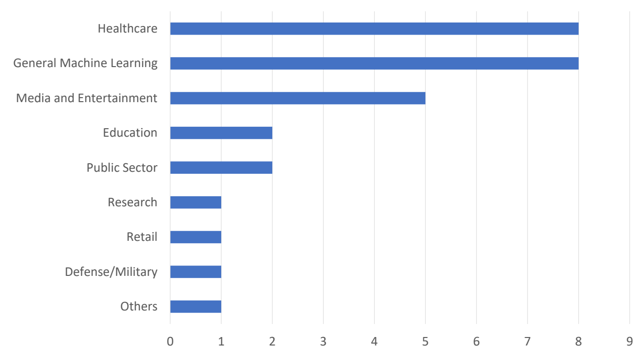

Figure 5: Participants' Application Domains.

We also asked participants about their application domain which will form the context of the user study. According to Figure 6, most of them are working in the healthcare, general machine learning and media and entertainment sectors. The application domain will impact the decisions made while navigating the process of the PbD methodology. As a measure to improve the consistency and validation of the questionnaire results, we included a redundancy test by asking the same question twice. Based on the redundancy check, we identify and discard invalid responses. Furthermore, the participants were instructed to complete the post-study questionnaire immediately after the user study, and most participants completed it on the same day as the user study.

We designed the questionnaire based on the 3 hypotheses:

1. PbD assists users to select the privacy concept that is appropriate for their applications.
2. PbD improves participants' ability to identify privacy concerns in their AI applications.
3. PbD helps users to stimulate the perspectives of different stakeholders.

Both the pre-study and post-study questionnaires consist of a main section where participants conduct a self-assessment of their understanding and ability to apply privacy concepts to their AI products and services. Each hypothesis is designed after exploring the literature on advances in the field of AI privacy and designed to rate the participant's individual ability to brainstorm and surface privacy issues, design relevant and optimal strategies, as well as to stimulate the perspectives of stakeholders. They had to give themselves a score of their understanding of AI privacy issues on a Likert Scale of 1 to 5, 1 being "strongly disagree" (SA) and 5 being "strongly agree" (SA). We conducted data analysis and hypothesis testing based on the results of the self-assessment questionnaire.

6 Results and Analysis

6.1 Hypothesis 1

Hypothesis 1: PbD assists users to select the privacy concept that is appropriate for their applications.

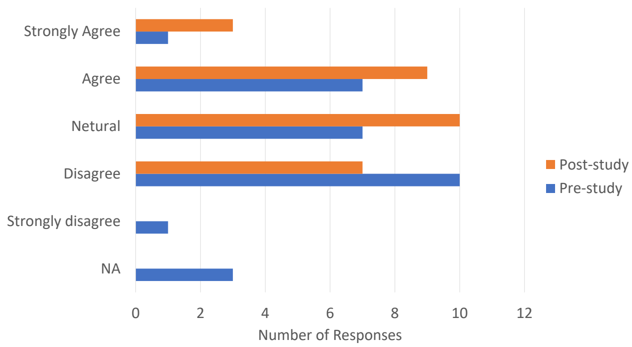

Figure 6: Participants' self-reported capability of making design decisions related to privacy before and after using PbD.

According to figure 6, it can be observed that the responses were mainly negative or neutral and can be said to follow a distribution roughly centred on "Disagree". This indicates that many of the participants are not so confident in their abilities to select the optimal privacy principle for their application domain. Since privacy is not a significant concern in many AI software teams, we expected that many participants are not well-versed in this area and require assistance in doing so. The findings also indicate that the distribution of the participants' self-assessed abilities to make relevant decisions pertaining to privacy is representative of a typical population of AI solution designers. After the participants proceeded through the PbD methodological tool, there was a significant increase in the number of participants who responded with 'Agree' and 'Strongly Agree', while the corresponding number of responses with 'Disagree' decreased significantly. This observation showed that the participants perceived the PbD methodology to be effective in enabling them to think critically about the privacy criteria that are relevant and optimal for their application scenarios.

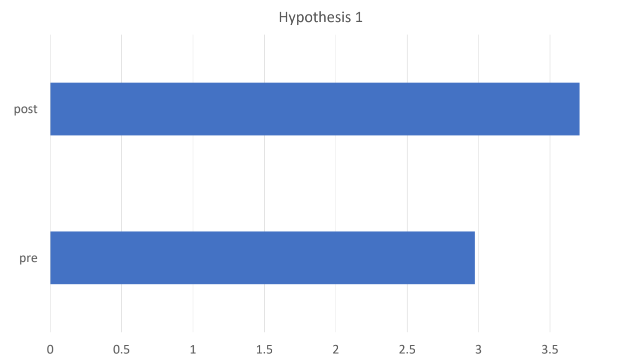

Figure 7: Participants' average scoring for the pre-and post-studies for hypothesis 1

According to figure 7, the average response score of the participants in the post-study was significantly higher than those in the pre-study, an increase of more than 0.6. After conducting statistical analysis and on the basis of the student's t-test of questionnaire results from H1, we concluded that the null hypothesis can be rejected at a 95 percent confidence interval with Cronbach alpha at 0.7462.

6.2 Hypothesis 2

Hypothesis 2: PbD improves participants' ability to identify privacy concerns in their AI applications.

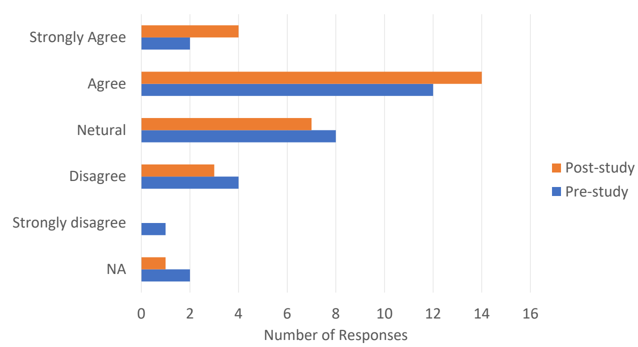

Figure 8: Participants' self-reported capability of surfacing privacy concerns before and after using PbD.

According to figure 8, the results provide an overview of the participants' responses to surfacing or identifying privacy concerns in the AI pipeline ahead of time focusing on hypothesis 2. This is a useful skill that enables early detection of problems that can become exacerbated in the later stages of development. Similar to the previous observations, the pre-study responses were roughly centred on 'Agree', with

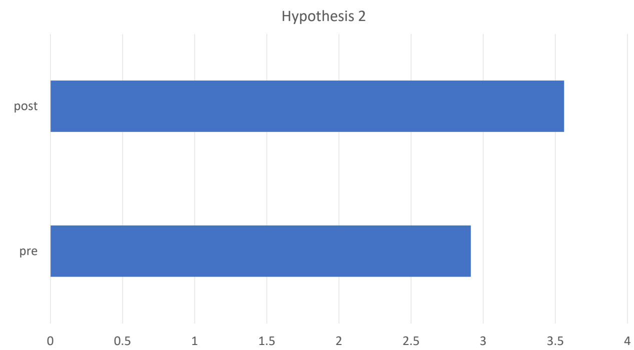

Figure 9: Participants' average scoring for the pre-and post-studies for hypothesis 2.

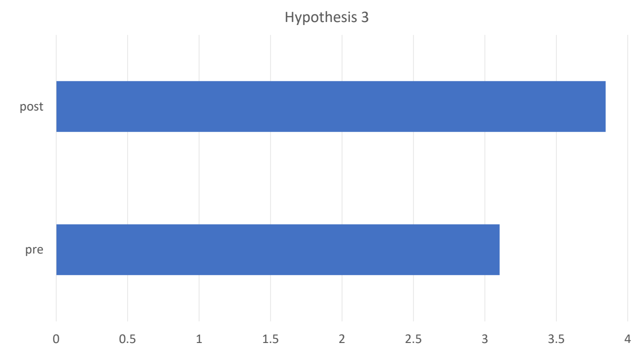

Figure 11: Participants' average scoring for the pre-and post-studies for hypothesis 3.

a slight increase in the number of responses of 'Agree' and 'Strongly Agree' post-study. This observation might be reflective of participants' relative confidence in being able to detect privacy issues during the development process.

For hypothesis 2, we found that the average questionnaire response increased by more than 0.5 in the post-study compared with that in the pre-study, according to 10. After conducting a student's t-test, we were only able to reject the null hypothesis at a 90 percent confidence level with the Cronbach alpha at 0.7156.

6.3 Hypothesis 3

Hypothesis 3: PbD helps users to stimulate the perspectives of different stakeholders.

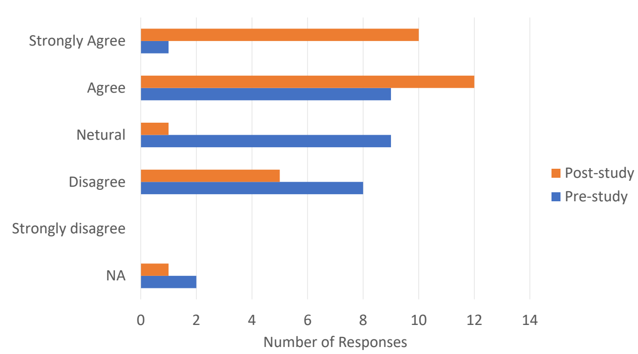

Figure 10: Participants' self-reported capability of stimulating stakeholders' perspective before and after using PbD.

Figure 10 illustrates the overview of participants' responses on stimulating the perspectives of various stakeholders, both direct and indirect. According to figure 10, there was a significant increase in the number of responses of 'Strongly Agree', from 1 to 10 after the use study. We found that the methodology actively facilitates critical thinking and enables participants to filter through irrelevant information to find issues that stakeholders are concerned with. Indirect stakeholders are usually overlooked in most development scenarios, and over the course of the user study, we constantly asked questions on addressing the needs of these communities that may not be directly impacted by the use of technology. Hence, the methodology significantly improved the participants' self-assessed ability to think critically from the perspective of stakeholders.

According to figure 11, the average of questionnaire responses increased by about 0.6 in the post-study compared with that in the pre-study. After conducting a student's t-test, we were able to reject the null hypothesis at a 95 percent confidence level with the Cronbach Alpha at 0.7683.

7 Discussions and Limitations

Over the course of multiple user studies, we found that the context of the application domain greatly impacts how decisions were made when measures for ethical AI are required. Most participants feedback that the PbD methodology is effective for promoting discussions and facilitating critical thinking regarding the various issues surrounding privacy in artificial intelligence. Especially when ethical issues surrounding privacy and other aspects can be difficult to uncover, deep insights driven by comprehensive exploratory thinking can avoid unnecessary issues in the future. However, ultimately measures designed to enhance ethical AI systems can also be perceived as trade-offs for performance. User 3 shared that the reality of privacy in the priority list of most large organisations:

"Realistically, privacy is low on business priorities as it is usually only appreciated in hindsight. However, I do recommend employing this methodology nonetheless as this study gave me the opportunity to pause and reflect on areas of weaknesses in my business with regard to privacy management. Being able to plan ahead is critical for companies to stay ahead of the game in the long run."

Although the issue of trade-off is a major hindrance in enhancing privacy in AI systems, the vision of the research community is that eventually, teams can deliver a system that minimises the compromise on performance and ethical values.

Despite the methodological toolkit being a useful starting point for software or AI design teams with no experience in addressing ethical AI issues, the complexities and fragmented state of the field can be a significant challenge to overcome, especially when mired in technical details. Participant 7 noted that the methodology can be used to generate industry best practices and guidelines to plan the trajectory of building ethical algorithmic systems:

"This methodology serves as a good starting point to incorporate privacy principles and considerations into ML projects. Further on, it can be helpful to provide a simplified view of the complications when dealing with ethics and inspire best practices in the process of enhancing private AI systems."

Furthermore, there are many stakeholders involved in building and deploying large-scale AI systems, and each group of stakeholders with diverging interests add to the difficulty of building privacy in AI. For the purpose of this study, we summarised the existing AI privacy principles into 4 main groups of techniques to encourage the exploratory thinking process. With additional time and resources, the team will be able to build a more nuanced and balanced view of many aspects of privacy and ethical AI. Participant 19 provided a glimpse into the future direction of methodological tools for building ethical AI:

"This tool can serve as a privacy framework to be used in the life-cycle of AI products, from design and conception to deployment. To enable the framework to be conducted effectively, the researchers can aim to standardise or estimate the requirements, time and resources needed for each step of the AI product pipeline. In this way, teams can work with this information to better achieve their key goals, objectives and deliverables."

Additionally, we found that for several participants, their internal model of privacy in AI was changed after being introduced to the methodology. Due to the lack of importance placed on privacy and the more general ethical AI principles, AI team members may not fully understand the implications and procedures of building these measures. Participant 11 noted that:

"After the discussion and deep dive into the methodology, the concept of privacy is completely different from what I expected. Upon learning what encompasses privacy in AI/ML, I believe it is now even more important to implement preventive measures against breaches of privacy."

After each user study, the team discussed potential ways to improve the process and each step of the methodology. In some application scenarios, the participants shared more in-depth principles and techniques that they employed. While in other groups, less relevant concepts were discarded and new topics were introduced to further facilitate discussion. Participant 27 commented on the requirements of these ethical AI tool-kits:

"The framework provides a clear structure to navigate the complexities and challenges in privacy-sensitive application environments. It is comprehensive, clear and easy to follow. However, it can be a significant challenge to build application-agnostic methodological frameworks, given the dynamic requirements of each field. One possible direction moving forward is to group similar application domains together and define the common objectives, timelines and specific deliverables to provide a systematic way to address this important field of ethical AI and privacy".

The user study consisted of 30 participants, many of which are experts in their field of AI/ML. However, to evaluate the PbD framework more effectively, a larger-scale online study is required. We propose to use a crowdsourcing tool such as Amazon Mechanical Turk to recruit participants from the public for this large-scale study. Furthermore, there are reports that self-assessed preferences and abilities usually do not align completely with participants' actual behaviours [Zell and Krizan, 2014]. Whether the findings from current and past work can value add to objectives in our methodological tool is still an open research question. When future works in AI privacy is implemented, dividing the investigation into multiple sub-categories seems to be the right course of action.

8 Conclusions and Future Work

The authors of this paper introduced a new methodology called PbD to help design teams tackle complex ethical dilemmas related to privacy in the development of AI products and pipelines. They identified gaps in current ethical AI design methodologies and developed PbD using the VSD theory and recent studies. PbD is designed to be user-friendly and time-efficient to facilitate its adoption by design teams. The authors plan to conduct user studies to assess the effectiveness of PbD in various application domains, such as banking, finance, autonomous vehicles[Atakishiyev *et al.*, 2021], and medical diagnosis [Tjoa and Guan, 2020]. They also plan to include project management functions to make PbD more accessible online. By promoting the use of PbD, the authors hope to improve the analysis of the ethical implications of AI products.

Acknowledgments

This research is supported, in part, by the National Research Foundation Singapore and DSO National Laboratories under the AI Singapore Programme (AISG Award No: AISG2-RP-2020-019); Alibaba Group through Alibaba Innovative Research (AIR) Program and Alibaba-NTU Singapore Joint Research Institute (JRI) (Alibaba-NTU-AIR2019B1), Nanyang Technological University, Singapore; and the Nanyang Assistant Professorship (NAP).

References

[Amodei *et al.*, 2016] Dario Amodei, Chris Olah, Jacob Steinhardt, Paul Christiano, John Schulman, and Dan Mané. Concrete problems in ai safety. *arXiv preprint arXiv:1606.06565*, 2016.

[Atakishiyev *et al.*, 2021] Shahin Atakishiyev, Mohammad Salameh, Hengshuai Yao, and Randy Goebel. Explainable artificial intelligence for autonomous driving: A comprehensive overview and field guide for future research directions. *arXiv preprint arXiv:2112.11561*, 2021.

[Ballard *et al.*, 2019] 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 (DIS'19)*, pages 421–433, 2019.

[Bost *et al.*, 2014] Raphael Bost, Raluca Ada Popa, Stephen Tu, and Shafi Goldwasser. Machine learning classification over encrypted data. *Cryptology ePrint Archive*, 2014.

[Croeser and Eckersley, 2019] Sky Croeser and Peter Eckersley. Theories of parenting and their application to artificial intelligence. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*, pages 423–428, 2019.

[EuropeanUnion, 2016] EuropeanUnion. Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation). 2016.

[Financial times, 2020] Financial times. Facebook Data Breach, 2020.

[Fredrikson *et al.*, 2015] Matt Fredrikson, Somesh Jha, and Thomas Ristenpart. Model inversion attacks that exploit confidence information and basic countermeasures. In *Proceedings of the 22nd ACM SIGSAC conference on computer and communications security*, pages 1322–1333, 2015.

[Friedman *et al.*, 2017] 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.

[Li *et al.*, 2016] Huaxin Li, Haojin Zhu, Suguo Du, Xiaohui Liang, and Xuemin Shen. Privacy leakage of location sharing in mobile social networks: Attacks and defense. *IEEE Transactions on Dependable and Secure Computing*, 15(4):646–660, 2016.

[Liu *et al.*, 2021] Bo Liu, Ming Ding, Sina Shaham, Wenny Rahayu, Farhad Farokhi, and Zihuai Lin. When machine learning meets privacy: A survey and outlook. *ACM Computing Surveys (CSUR)*, 54(2):1–36, 2021.

[Lyu *et al.*, 2022] Lingjuan Lyu, Han Yu, Xingjun Ma, Chen Chen, Lichao Sun, Jun Zhao, Qiang Yang, and Philip S. Yu. Privacy and robustness in federated learning: Attacks and defenses. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[Makridakis, 2017] Spyros Makridakis. The forthcoming artificial intelligence (ai) revolution: Its impact on society and firms. *Futures*, 90:46–60, 2017.

[McPherson *et al.*, 2016] Richard McPherson, Reza Shokri, and Vitaly Shmatikov. Defeating image obfuscation with deep learning. *arXiv preprint arXiv:1609.00408*, 2016.

[Nasr *et al.*, 2019] Milad Nasr, Reza Shokri, and Amir Houmansadr. Comprehensive privacy analysis of deep learning: Passive and active white-box inference attacks against centralized and federated learning. In *2019 IEEE symposium on security and privacy (SP)*, pages 739–753, 2019.

[Schwab, 2017] Klaus Schwab. *The fourth industrial revolution*. Currency, 2017.

[Shi *et al.*, 2023] Yuxin Shi, Han Yu, and Cyril Leung. Towards fairness-aware federated learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2023.

[Shneiderman and Hochheiser, 2001] Ben Shneiderman and Harry Hochheiser. Universal usability as a stimulus to advanced interface design. *Behaviour & Information Technology*, 20(5):367–376, 2001.

[Shokri *et al.*, 2017] Reza Shokri, Marco Stronati, Congzheng Song, and Vitaly Shmatikov. Membership inference attacks against machine learning models. In *2017 IEEE symposium on security and privacy (SP)*, pages 3–18. IEEE, 2017.

[Shu *et al.*, 2021] Ying Shu, Jiehuang Zhang, and Han Yu. Fairness in design: A tool for guidance in ethical artificial intelligence design. In *International Conference on Human-Computer Interaction*, pages 500–510. Springer, 2021.

[Song *et al.*, 2017] Congzheng Song, Thomas Ristenpart, and Vitaly Shmatikov. Machine learning models that remember too much. In *Proceedings of the 2017 ACM SIGSAC Conference on computer and communications security*, pages 587–601, 2017.

[Tan *et al.*, 2022] Alysa Ziying Tan, Han Yu, Lizhen Cui, and Qiang Yang. Towards personalized federated learning. *IEEE Transactions on Neural Networks and Learning Systems*, 2022.

[Tjoa and Guan, 2020] Erico Tjoa and Cuntai Guan. A survey on explainable artificial intelligence (xai): Toward medical xai. *IEEE Transactions on Neural Networks and Learning Systems*, 2020.

[Tramèr *et al.*, 2016] Florian Tramèr, Fan Zhang, Ari Juels, Michael K Reiter, and Thomas Ristenpart. Stealing machine learning models via prediction apis. In *USENIX security symposium*, volume 16, pages 601–618, 2016.

[Yang *et al.*, 2019a] Qiang Yang, Yang Liu, Tianjian Chen, and Yongxin Tong. Federated machine learning: Concept and applications. *ACM Trans. Intell. Syst. Technol.*, 10(2):12:1–12:19, January 2019.

[Yang *et al.*, 2019b] Qiang Yang, Yang Liu, Yong Cheng, Yan Kang, Tianjian Chen, and Han Yu. Federated learning. *Synthesis Lectures on Artificial Intelligence and Machine Learning*, 13(3):1–207, 2019.

[Yu *et al.*, 2017] Han Yu, Chunyan Miao, Yiqiang Chen, Simon Fauvel, Xiaoming Li, and Victor R Lesser. Algorithmic management for improving collective productivity in crowdsourcing. *Scientific reports*, 7(1):1–11, 2017.

[Yu *et al.*, 2018] Han Yu, Zhiqi Shen, Chunyan Miao, Cyril Leung, Victor R Lesser, and Qiang Yang. Building ethics into artificial intelligence. In *Proceedings of the 27th International Joint Conference on Artificial Intelligence (IJCAI'18)*, pages 5527–5533, 2018.

[Zell and Krizan, 2014] Ethan Zell and Zlatan Krizan. Do people have insight into their abilities? a metasynthesis. *Perspectives on Psychological Science*, 9(2):111–125, 2014.

[Zhang and Yu, 2022a] Jiehuang Zhang and Han Yu. A methodological framework for facilitating explainable ai design. In *International Conference on Human-Computer Interaction*, pages 437–446. Springer, 2022.

[Zhang and Yu, 2022b] Yanci Zhang and Han Yu. Towards verifiable federated learning. In *Proceedings of the 31st International Joint Conference on Artificial Intelligence (IJCAI'22)*, pages 5686–5693, 2022.

[Zhang *et al.*, 2021] Jiehuang Zhang, Ying Shu, and Han Yu. Human-machine interaction for autonomous vehicles: A review. In *International Conference on Human-Computer Interaction*, pages 190–201. Springer, 2021.