


# EID: Facilitating Explainable AI Design Discussions in Team-Based Settings

Jiehuang Zhang<sup>1,2</sup>  and Han Yu<sup>1</sup>

## ABSTRACT

Artificial intelligence (AI) systems have many applications with tremendous current and future value to human society. As AI systems penetrate the aspects of everyday life, a pressing need arises to explain their decision-making processes to build trust and familiarity among end users. In high-stakes fields such as healthcare and self-driving cars, AI systems are required to have a minimum standard for accuracy and to provide well-designed explanations for their output, especially when they impact human life. Although many techniques have been developed to make algorithms explainable in human terms, no design methodologies that will allow software teams to systematically draw out and address explainability-related issues during AI design and conception have been established. In response to this gap, we proposed the explainability in design (EID) methodological framework for addressing explainability problems in AI systems. We explored the literature on AI explainability to narrow down the field into six major explainability principles that will aid designers in brainstorming around the metrics and guide the critical thinking process. EID is a step-by-step guide to AI design that has been refined over a series of user studies and interviews with experts in AI explainability. It is devised for software design teams to uncover and resolve potential issues in their AI products and to simply refine and explore the explainability of their products and systems. The EID methodology is a novel framework that aids in the design and conception stages of the AI pipeline and can be integrated into the form of a step-by-step card game. Empirical studies involving AI system designers have shown that EID can decrease the barrier of entry and the time and experience required to effectively make well-informed decisions for integrating explainability into their AI solutions.

## KEYWORDS

explainable artificial intelligence (AI); design method; design tool; values sensitive design; ethics; design methodology

In the age of digitization and the fourth industrial revolution<sup>[1]</sup>, several enabling technologies include artificial intelligence (AI). AI systems are the key to new breakthroughs in important fields and numerous fields, such as video generation<sup>[2]</sup>, natural language processing<sup>[3]</sup>, algorithmic crowdsourcing<sup>[4,5]</sup>, government service provision<sup>[6]</sup>, and self-driving vehicles<sup>[7]</sup>. AI systems, especially machine learning and neural network based or deep learning systems, have allowed us to perform many tasks with greatly increased scale and finesse and technological breakthroughs that were believed to be unattainable without AI. While these advancements facilitated by AI technologies have resulted in great changes in the way we live and work<sup>[8]</sup>, there are complex and multi-faceted morality issues that warrant attention. Addressing these issues is important because the decision support systems otherwise solely handled by humans are increasingly being transferred to the responsibility of AI. Owing to this shift of autonomy and responsibility to algorithms, the chances of mistakes or improper decision-making must be addressed in a timely manner before side effects emerge. As part of a concerted approach, societies heavily reliant on AI systems must consider the repercussions of such a shift and regulate the advancement of AI towards a future direction with proper oversight where the benefits outweigh the potential harms<sup>[9]</sup>.

At the current stage of the AI community, most of the practitioners in the AI circle assess the performance of AI systems based on their accuracy scores and impact on computing

resources. Despite the usefulness of these metrics, they may not provide a complete representation of the inner workings of the decision-making process. Although state-of-the-art AI systems can assist or replace many processes in the workplace and peoples' personal lives, they generally lack explainability, and even the system designers are unable to fully explain how they work<sup>[10]</sup>. Despite being trained on factual and logical datasets, these algorithms are not invulnerable to mistakes of misjudgements and various other issues that can be difficult to detect<sup>[11]</sup>. In addition to being unable to understand how algorithms reach their decision output, there might be problems that even go undetected for a long period of time. For example, we refer to spectral heat-maps. Lapuschkin et al.<sup>[12]</sup> discovered that standard performance evaluation metrics are possibly unaware of certain types of issues in the decision-making. Ultimately, the consensus in the community is that the black box AI used in modern times encounters problems related to the transparency and explainability of its inner workings, especially in specific fields where the users and other stakeholders must understand how the output is being reached. Some examples of these fields are medical diagnosis and self-driving cars. These fields urgently require a satisfactory explanation to be generated or given when prompted.

In the context of AI, explainability is a complicated endeavour because it is multi-faceted, and its definition can be fluid depending on the context and type of requirements. Explainable artificial intelligence (XAI)<sup>[13]</sup> aims to create an arsenal of machine

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, jiehuang001@e.ntu.edu.sg](mailto:jiehuang.zhang, jiehuang001@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/>).

learning tools that enable users to understand, trust, and constructively regulate the advancing generation of AI systems<sup>[14]</sup>. These goals are within reach when designers intentionally create AI algorithms to have features that enhance explainability and are comprehensible from the perspective of human users. Despite the increasing difficulty of achieving explainable AI, progress has been accelerating in this field<sup>[15]</sup> (e.g., Layer-wise Relevance Propagation<sup>[16]</sup>). These steady improvements in explainability have laid the foundation for the vision of XAI as the community gains clarity about the function of complicated AI models, such as deep learning neural networks. In addition to creating tools and methods to enhance explainability in algorithms after they have generated an output (in other words, post hoc manner), we need to tackle the problems early in the design and conception (DC) stage. Methodological frameworks that will aid AI software development teams and research groups to integrate explainability measures in their AI products and/or services are currently lacking.

This work proposes the explainability in design (EID) methodology, a step-by-step framework that guides software design teams and research groups to systematically consider, explore, draw out, and resolve any explainability-related issues and problems in their AI systems. EID is designed to simultaneously elicit critical thinking during all stages of the AI life cycle and reduce the barrier of entry to allow lay people to participate in the conversation of ethical AI<sup>[17]</sup>. Explainability is one of the major pillars of ethical AI and, in many ways, allows for improvements in other pillars, such as privacy and fairness. Owing to the complicated nature of AI explainability, lay people often find the concepts and technicalities hard to understand, let alone contribute to this endeavour. To make matters worse, the desire to enhance explainability may lead to trade-offs in accuracy or efficiency because human or computing resources might be removed from the main objectives in the AI life cycle. EID aims to address these issues by introducing teams to the systematic process of brainstorming and discussing the likelihood of explainability-related problems for their AI products. Thus, software teams and research groups can identify or create explainability requirements and objectives specific to their AI system and context while stimulating perspective thinking from other groups of stakeholders, whether direct or indirect.

Empirical user studies involving 35 AI designers reveal that EID significantly enhances the ability of software design teams to identify, explore, and resolve ethical issues and problems related to explainability. Additionally, EID helps reduce the barrier of entry for team members to effectively participate in design and conception, allowing a great pool of participants to improve the AI software products and services. EID can assist software teams in diverse application domains, from e-commerce to facial recognition systems and beyond. With its agnostic application, EID benefits a large number of AI engineers and researchers in their work.

In Sections 1 and 2, we discuss the related works in AI explainability, compile a list of the core explainability principles, and then narrow the list down for use in the methodological framework. In Section 3, we focus on explaining the EID methodology in detail and how each step is executed. In Sections 4 and 5, we highlight user studies involving 35 AI technology professionals to evaluate the effectiveness of EID. Empirical study results suggest that EID is useful for assisting the participants in making good decisions and lowering the barrier to entry for software teams to address complex ethical issues. In Section 6, we address the limitations of the methodology and user study. In

Section 7, we conclude the paper by identifying promising future directions for our work.

## 1 Related Work

The field of XAI is quite large and complex, making the difficulty of creation of taxonomy high. For the purpose of this paper, we broadly classify the techniques into two main categories, post hoc and integrated approaches<sup>[13]</sup>, according to the stage of the life cycle where the techniques are applied. Integrated XAI refers to the building of explainability features during the design and construction of the algorithm, while post hoc XAI means that the explainability of an algorithm is only investigated after the output has been produced. Both categories have associated advantages and flaws that we will explore further in the paper and subsequent user studies. For now, the obvious advantage of post hoc explainability is that there is a low chance that this approach will interfere with the performance of the AI system. The research community in XAI is active and many new improved and novel methods of enhancing explainability in AI systems have emerged in recent times, such as Shapley values<sup>[18]</sup> and local interpretable model-agnostic explanations (LIME)<sup>[19]</sup>.

To the best of our knowledge, value sensitive design (VSD) is currently the prominent toolkit in the field of ethical AI methodological frameworks<sup>[20]</sup>. VSD is closely related to the field of human-computer interaction (HCI) and information systems design, and aims to resolve design issues by centring the analysis around ethical values such as privacy and fairness. These values are used in the workflows, allowing system designers to gain deeper insights and integrate with other methodological tools. The brainstorming sessions also consider the roles, values, and goals of both direct and indirect stakeholders, by stimulating perspective-taking in the process. According to Fig. 1, the main difference between direct and indirect stakeholders is that direct stakeholders use the AI product or service directly, while indirect stakeholders are impacted by the use but do not directly use the AI. Since the effect of AI use is less apparent on indirect stakeholders and as a result it has a higher probability of being overlooked, it might be better to allocate more time and attention to the analysis of indirect stakeholders. VSD has been the basis of two methodological card games, Envisioning Cards<sup>[21]</sup> and Judgement Call<sup>[22]</sup>. Envisioning Cards, as the name suggests, help to stimulate critical thinking and emphasize players' focus on timelines, stakeholder interests and values, as well as pervasiveness. They prompt participants to consider the long-term and likely systemic problems in system design. In contrast, Judgement Call is a card game that AI developer groups can use to surface ethical problems in an AI product. It consists of cards that primarily focus on scoring reviews and using wild cards that facilitate critical thinking.

Liao et al.<sup>[23]</sup> brought to light that while AI systems need explainability features, there is a consensus to address on the ground real-world user requirements before we understand algorithms. Liao et al.<sup>[23]</sup> invented a question bank in which user needs for explainability are portrayed as prototypical questions users might ask about the AI and use it as a study probe. Then, they consulted usability experience (UX) and design experts on the current gaps between XAI algorithmic work and practices. Reference [24] showed that there is a lack of a principled framework that can provide the basis for the development of an XAI framework. Then Kim et al.<sup>[24]</sup> came up with four foundational components that can assist to create a simple methodological framework to facilitate the design of XAI systems.



Fig. 1 Metrics of explainability in AI categorised into 6 types.

To date, no software engineering design methodology has been developed to guide an AI solution development team in brainstorming and determining what XAI principles should be incorporated into a given AI system design. Hence, the proposed EID methodological framework is designed to fill this gap.

## 2 Preliminary

In this paper, we have classified the principles of XAI into the three main types as shown in Fig. 2<sup>[3]</sup>: (1) Transparency, (2) interpretability, and (3) explainability. This classification allows lay people to focus on the most relevant or important principles of XAI and brainstorm how to implement in their application scenarios.

(1) **Transparency:** An algorithm is transparent when it can be viewed in a stand-alone manner. An algorithm can feature different degrees of understandability.

(a) **Model transparency:** degree of human understandability of how the components (e.g., filters used and layers of a neural network (NN)) of the trained model contribute to the output.

(b) **Design transparency:** degree of human understandability of design decisions made to create the machine learning model.

(c) **Algorithmic transparency:** degree of human understandability of the training process that resulted in the trained machine learning model.

The model and design transparency can be distinguished by the specific section of the transparency analysis: Model transparency focuses on the way individual components are sequenced, and design transparency refers to the design choices made by the

engineer to enhance the transparency of the algorithm. For example, a simple algorithm such as decision trees are inherently transparent. For enhanced model transparency, the components of the algorithm must be easily explainable to a layperson. To an untrained layperson, the two types of transparency are largely indistinguishable. The goal of this classification system is to inform them of the minute but significant differences.

(2) **Interpretability:** the ability to explain or to provide the meaning in understandable terms to a human.

(a) **Integrated interpretability:** design of an ML model that involves specific design choices for better understandability.

(b) **Post hoc interpretability:** ability to analyse information pertaining to how the output of a trained ML model is obtained from the input.

(3) **Explainability:** Associated with the notion of explanation as an interface between humans and a decision maker, the focus is on the human and how the human can understand the mechanics of an algorithm.

Differentiating between interpretability and explainability is also crucial. The main difference is the human factor: explainability indicates how a human being understands the output explanation, and interpretability is the degree of understandability of the algorithmic decision-making process that is not based on human factors but relative to other algorithms. In this methodology, the humans concerned are the AI system designers and direct and indirect stakeholders. The subtle differences in these principles allow for interesting trade-offs and interplays when the specifics are given due to attention and analysis.

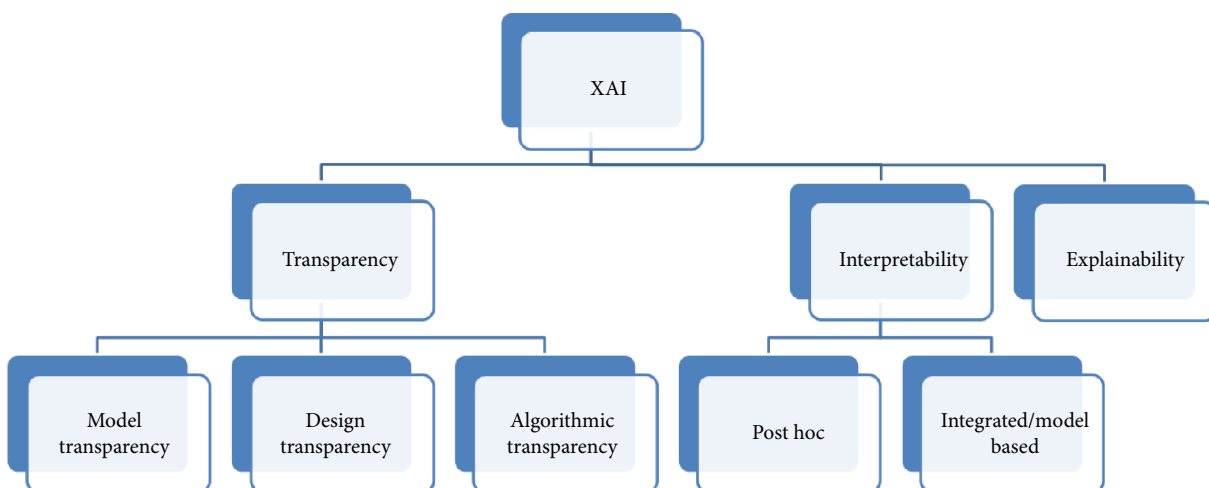


Fig. 2 Comparison between direct and indirect stakeholders.

### 3 Explainability in Design Methodology

EID methodology is integrated into the gameplay processes of a card game. The methodology is designed to be relevant to a wide range of application scenarios, in other words, model agnostic, allowing the team to adjust fine details to tailor the methodology to their AI product or service. With the barrier to entry to use EID being lowered, the lay participants and experienced people can join the process and collectively brainstorm towards realistic improvements. The workflow for a team to use EID to facilitate team discussions around XAI issues is shown in Fig. 3. We discuss the details of the step-by-step guide to EID workflow.

(1) In Step 1, each team member selects a scenario that establishes the context for the entire user study. The scenario can be real or fictitious but preferably in a field where explainability is of high priority, and the dataset of the scenario partially or fully contains a dataset generated by people. If possible, most of the team should be acquainted with the scenario such that many fine details are considered in the flow of the user study. In selected scenarios, several considerations and trade-offs can sometimes impact the decision-making process.

(2) In Step 2, participants must select the type of application card that accurately reflects their scenario. We have chosen some considerations from the work from Shneiderman and Hochheiser's classification for usability motivation in the human-computer interaction literature<sup>[25]</sup>. They are (a) life-critical systems, (b) industrial and commercial uses, (c) office, home, and entertainment, (d) exploratory, creative, and collaborative applications, and (e) socio-technical applications.

(3) In the next Step 3, the team must find and list the types of stakeholders that are central to the process. Armed with the list of direct and indirect stakeholders, each individual must stimulate the perspective of a stakeholder and perform an analysis of the principles pertaining to that stakeholder. To facilitate this process, the EID methodology includes a list of exploratory questions in this step. Some examples include:

- (a) Other than the stakeholders, who can the output explanations be targeted at?
- (b) Does the timing of the explanation generated matter to the user?
- (c) How can the frequency of use impact the trust levels in the user?
- (d) Does the emotional state of the user impact other factors?
- (e) How can the algorithm perform this operation in a more transparent way?
- (f) Determine the scope and depth of the explanations needed in this scenario.

As a result of this step, each individual participant should grasp the perspective of their stakeholder deeply, which in turn facilitates the XAI principles to be explored. These requirements

are even more pronounced when the analysis of indirect stakeholders occurs and the impact on them may not be initially apparent.

(4) In the next Step 4, the team members must select the highest-rated explainability principles relevant to the scenario. They have to justify the selection and why the other principles were not chosen. When a specific principle is chosen repeatedly, we can identify it as the default principle that works well for the scenario.

(5) In the final Step 5, the leader of the group collects the responses of the team and randomly shuffles them before reading them to the team. With the element of anonymity, the members are spurred to be more forthright in their decision-making processes. The leader will assess the responses to decide if any further action or step is to be taken.

At the end of the workflow, the leader of the group can consider returning to Step 3 to initiate a new stakeholder analysis if necessary. The members can also shift their focus to another part of the study to ensure good coverage. If the results of the workflow are satisfactory, then the process can be ended.

The methodology aims to provide the following deliverables:

- (1) Select the explainability metrics most suitable for the scenario.
- (2) Determine priorities of specialised requirements for the format, scope, and type of explanations.
- (3) Measure the differences in the knowledge and experience levels of team members.
- (4) Facilitate the discovery of explainability issues and where in the AI system it is located.

### 4 Empirical Evaluation

We conducted user studies to empirically evaluate the proposed explainability in the design framework and our hypotheses.

We recruited 35 participants for the user study. All of the participants are experienced researchers or engineers who are currently working or have previously worked on software systems design involving AI technologies. We also considered additional criteria, for example, the ability to understand basic explainability concepts surrounding the ML literature and consenting to be recorded. We recruited participants of a diverse age range to investigate how the methodology can impact users with different levels of seniority. According to Fig. 4, most participants fell in the 20–30 years age group, which is representative of the typical target users of our framework.

In the pre-study questionnaire, we asked participants to report how they prioritise ethical considerations in their AI solution development experience. As shown in Fig. 5, most participants chose explainability, fairness, and privacy as the top 3 ethical considerations, followed by accountability and compliance. This

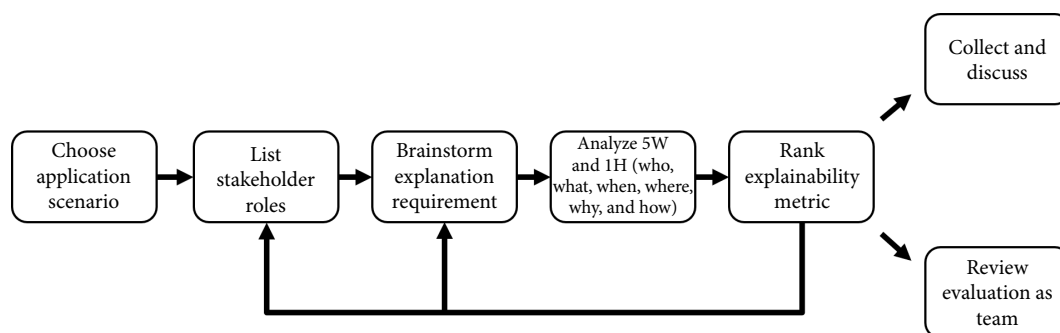


Fig. 3 Explainability in design workflow.



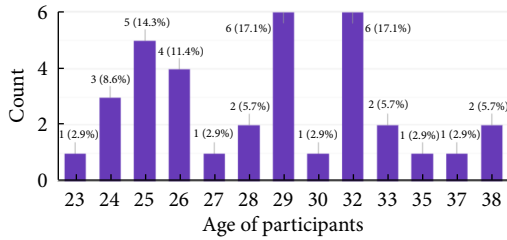


Fig. 4 Demographics of the participants.

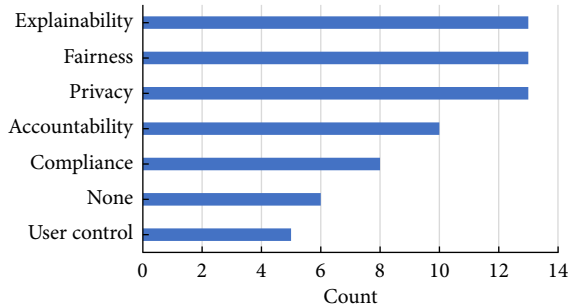


Fig. 5 Participants' ethical AI prioritisation.

finding corresponds with the explainability aspect of ethical AI that we have chosen as the focus of our user study.

We also asked participants to identify the application domain where they have worked on their AI products and services. As shown in Fig. 6, most of the participants are working in the healthcare sector, general-purpose machine learning (ML) applications, and government-related projects. To improve consistency in our questionnaire, we included a redundancy test by asking the same question twice, once in a positive way and once in a contrasting negative way. For example, we first asked the following positive question, "I can navigate complex ethical choices around AI/ML explainability", and subsequently, the negative question, "I do not know how to make decisions regarding explainability in AI/ML". By using this redundancy check, we can detect and discard invalid responses. Furthermore, we informed the participants to complete the post-study questionnaire as soon as possible after the study. All the participants completed the questionnaire on the first day.

The 3 main hypotheses for this user study are as follows.

- (1) The EID methodology helps participants determine the explainability criteria that are the most relevant to their AI applications.
- (2) The EID methodology helps participants surface explainability concerns in their AI applications.
- (3) The EID methodology helps participants envision the perspectives from different stakeholders.

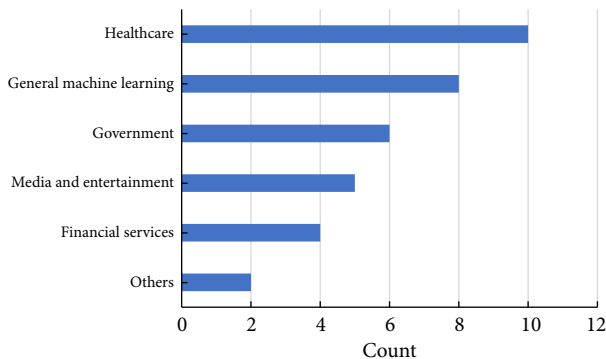


Fig. 6 Participants' application domains.

We created the pre- and post-study questionnaires for the participants to self-assess their understanding of the explainability concepts and how to apply them to their AI products and services. Each hypothesis was designed to assess the individual ability of the participants to choose an applicable explainability solution, brainstorm and draw out explainability concerns, and stimulate the thinking process and perspectives of stakeholders. The participants were asked to rate their understanding of explainability problems on a Likert scale of 1 to 5, 1 being "strongly disagree" (SD) 2 being "disagree" (D), 3 being "neutral" (N), 4 being "agree" (A), and 5 being "strongly agree" (SA). Using the results of the questionnaire, we conducted statistical data analysis to evaluate the three hypotheses.

## 5 Result and Analysis

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

### 5.1 Hypothesis 1

**Hypothesis 1:** The EID methodology helps participants determine the explainability criteria that are the most relevant to their AI applications.

Figure 7 illustrates the results from the participant's responses to questions related to H1. The responses were largely negative and followed a distribution roughly centred on "disagree", signalling a general lack of confidence in the self-assessment by the participants. This result was expected because most of the participants have not actively worked on explainability issues in the AI domain and therefore are unlikely to be competent or experienced in this area. This finding also signified that the distribution of the participants' capabilities to make design decisions related to the explainability aspect of AI was typical of a population of AI solution designers. After the participants used EID in the empirical study sessions, a significant increase in the number of participants who responded with "agree" and a significant decrease in the number of "disagree" and "strongly disagree" responses were observed. These results indicated that the participants found the methodology to be useful in helping them think about the explainability principle in their application domains.

As shown in Fig. 8, the participants' average response scores in the post-study were significantly higher than those in the pre-study. 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 and confidence interval with Cronbach alpha at 0.7256.

### 5.2 Hypothesis 2

**Hypothesis 2:** The EID methodology helps participants surface

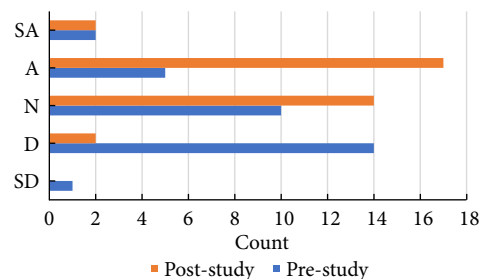


Fig. 7 Participants' self-reported capability of making design decisions related to explainability before and after using EID.

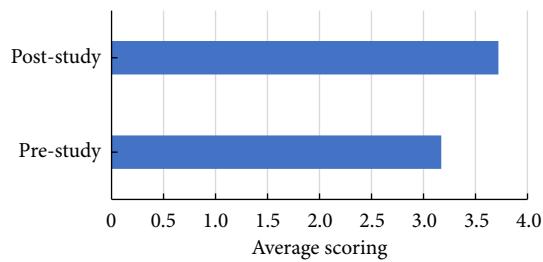


Fig. 8 Participants’ average scoring for the pre- and post-studies for Hypothesis 1.

explainability concerns in their AI applications. This hypothesis pertains to the participants’ self-assessment of their competency in discovering issues or concerns about the explainability aspect.

Figure 9 illustrates the results from the participants’ responses about drawing out explainability concerns by focusing on Hypothesis 2. This question indicates the self-assessed competency of participants to identify in advance what kind of explainability issues can occur in their application domains. Similar to previous results, the responses were roughly centred on “agree”. In contrast, after using the EID methodology, a significant increase was observed in the number of responses for “strongly agree” and “agree”, and a corresponding decrease was found in the number of responses for “strongly disagree” and “disagree”. The results indicated that EID is effective in identifying potential issues or concerns about explainability in advance.

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 Fig. 10. 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 0.7456.

### 5.3 Hypothesis 3

**Hypothesis 3:** The EID methodology helps participants envision the perspectives of different stakeholders. We conceptualised this hypothesis to assess the competence of participants to stimulate thinking in the perspective of other relevant groups of

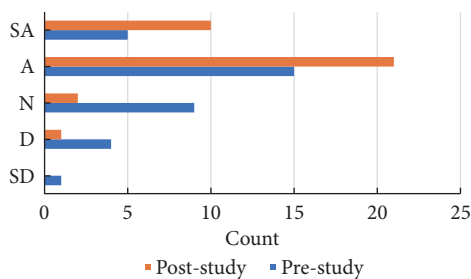


Fig. 9 Participants’ self-reported capability of surfacing explainability concerns before and after using EID.

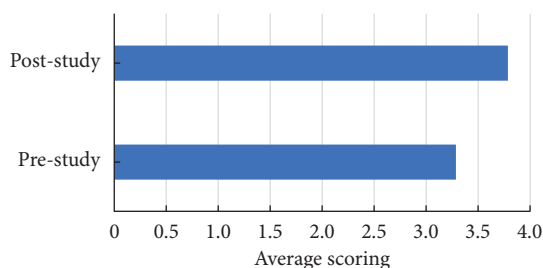


Fig. 10 Participants’ average scoring for the pre- and post-studies for Hypothesis 2.

stakeholders.

Figure 11 highlights the results from the participants’ responses on thinking from the perspective of stakeholders. The question focuses on Hypothesis 3 and challenges the participants to visualise the perspective of 2 types of stakeholders, namely, the direct and indirect stakeholders. For example, direct stakeholders can be the end user of the AI system or the system designer and engineers, and indirect stakeholders include the family members of the end users. The proportion of “strongly disagree” and “disagree” greatly decreased, and many participants changed their answers to “agree” and “strongly agree”. Hence, the methodology greatly improved the self-assessed ability of participants to stimulate stakeholder thinking, a valuable skill set in AI development teams.

According to Fig. 12, the average of questionnaire responses increased slightly in the post-study compared with that in the pre-study. However, when we conducted a student’s t-test analysis of the questionnaire results from Hypothesis 3, the null hypothesis cannot be rejected.

## 6 Discussion and Limitation

We found EID to be an effective framework for promoting conversations and eliciting critical thinking about explainability in AI. A series of user studies revealed that EID introduced the participants to explainability and provided them with deep insights into the complexities of explainability in AI. In the post-study, 19 participants expressed their confidence in making complex decisions around AI explainability. This number was higher than the seven participants before the study. However, the EID methodology should be used early in the design and conception stage of the AI life cycle because the metrics must be integrated early into the product or service. These design decisions will also impact the rest of the AI pipeline. Although the EID framework is a good starting point for software teams with no experience in explainability, the field of AI explainability can become quite complex and fragmented, especially when going deep into the technical details. We simplified the explainability

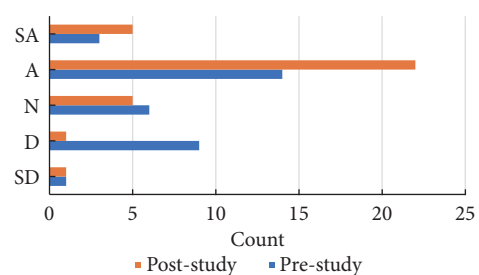


Fig. 11 Participants’ self-reported capability of thinking from stakeholders’ perspective before and after using FID.

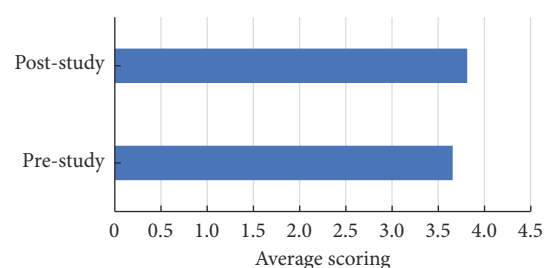


Fig. 12 Participants’ average scoring for the pre- and post-studies for Hypothesis 3.

principles by narrowing them down into six main categories to facilitate our user studies without going too deep into the complexities of AI explainability. With additional time and resources, we can provide a balanced view of many aspects of explainability in the ethical AI field. For this reason, most AI system designers are unfamiliar with explainability and/or see it as an unnecessary trade-off for performance or efficiency. Most AI developers are not willing to forgo the reduction in their algorithm's performance unless a specific requirement is imposed to include explainability. This response has been duly noted; however, we believe that, eventually, teams can deliver a system that minimises the compromise on performance and explainability. Over the course of multiple user studies, we modified the EID framework or user study procedures for improvement. For example, we realised that in some application domains, some explainability principles were irrelevant to the participants. Thus, we decided to give the participants the option to discard irrelevant explainability principles. Our use study consisted of only 35 participants. Recruiting a large number of people to conduct a larger-scale online study is necessary to evaluate the EID framework. Additionally, self-reported preferences often do not align well with participants' actual behaviours<sup>[36]</sup>. Whether the findings from existing and past work contribute to significant improvements in our methodology remains an open question. Given that explainability in AI is a large field, dividing the investigation into different subfields seems reasonable.

## 7 Conclusion and Future Work

This work provided an overview of the state of the field in ethical AI design, along with the gaps in the literature and proposed solutions. Using existing toolkits, such as the theory of VSD and various recent studies, we proposed and tested a methodological framework, EID. This framework assists software design teams in facilitating complicated ethical choices around explainability. Owing to its efficiency in terms of time and effort and a low entry knowledge barrier, EID effectively allows team members to improve the decision-making process for explainability in their AI products and services.

In proposed future work, we aim to conduct larger-scale online-only user studies to evaluate the effectiveness of our EID and assess if the project goals are attainable. The format of our user studies is currently conducted in teams of more than one member that are working or have worked on AI products previously. As the new normal of working from home has been in effect for the past years, the online form of EID will be taking priority in our proposed future research goals. Then the team members can collaborate and use the EID methodology online accordingly. We aim to include project management functions for all team members to manage work allocation and timeline management. In addition to the above, we plan to identify specific application domains such as autonomous vehicles<sup>[27]</sup> and medical healthcare diagnosis<sup>[28]</sup> to apply our methodology and investigate deeper on how the different complexities interact.

## Acknowledgment

This work was supported in part by 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); Nanyang Technological University, Nanyang Assistant Professorship (NAP); 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 & 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: 16 August 2022; Revised: 28 September 2022; Accepted: 30 September 2022

## References

- [1] K. Schwab, *The Fourth Industrial Revolution*. New York, NY, USA: Currency, 2017.
- [2] C. Liu, Y. Dong, H. Yu, Z. Shen, Z. Gao, P. Wang, C. Zhang, P. Ren, X. Xie, L. Cui, et al., Generating persuasive visual storylines for promotional videos, in *Proc. 28<sup>th</sup> ACM International Conference on Information and Knowledge Management (CIKM'19)*, Beijing, China, 2019, pp. 901–910.
- [3] X. Guo, B. Li, H. Yu, and C. Miao, Latent-optimized adversarial neural transfer for sarcasm detection, in *Proc. 2021 Annual Conference of the North American Chapter of the Association for Computational Linguistics (NAACL-HLT'21)*, Virtual event, 2021, pp. 5394–5407.
- [4] H. Yu, Z. Shen, and C. Leung, Bringing reputation-awareness into crowdsourcing, in *Proc. 9<sup>th</sup> International Conference on Information, Communications and Signal Processing (ICICS'13)*, Tainan, China, 2013, pp. 1–5.
- [5] H. Yu, C. Miao, Y. Chen, S. Fauvel, X. Li, and V. R. Lesser, Algorithmic management for improving collective productivity in crowdsourcing, *Scientific Reports*, vol. 7, no. 1, p. 12541, 2017.
- [6] Y. Zheng, H. Yu, L. Cui, C. Miao, C. Leung, and Q. Yang, SmartHS: An AI platform for improving government service provision, in *Proc. 30<sup>th</sup> AAAI Conference on Innovative Applications of Artificial Intelligence (IAAI-18)*, New Orleans, LA, USA, 2018, pp. 7704–7711.
- [7] J. Zhang, Y. Shu, and H. Yu, Human-machine interaction for autonomous vehicles: A review, in *Proc. 23<sup>rd</sup> International Conference on Human-Computer Interaction*, Virtual event, 2021, pp. 190–201.
- [8] S. Makridakis, The forthcoming artificial intelligence (AI) revolution: Its impact on society and firms, *Futures*, vol. 90, pp. 46–60, 2017.
- [9] S. Croeser and P. Eckersley, Theories of parenting and their application to artificial intelligence, in *Proc. 2019 AAAI/ACM Conference on AI, Ethics, and Society*, Honolulu, HI, USA, 2019, pp. 423–428.
- [10] M. Heinert, Artificial neural networks—how to open the black boxes? in *Proc. First Workshop Application of Artificial Intelligence in Engineering Geodesy (AIEG 2008)*, Vienna, Austria, 2008, pp. 42–62.
- [11] Y. Shu, J. Zhang, and H. Yu, Fairness in design: A tool for guidance in ethical artificial intelligence design, in *Proc. 23<sup>rd</sup> International Conference on Human-Computer Interaction*, Virtual event, 2021, pp. 500–510.
- [12] S. Lapuschkin, S. Wäldchen, A. Binder, G. Montavon, W. Samek, and K. -R. Müller, Unmasking clever Hans predictors and assessing what machines really learn, *Nature Communications*, vol. 10, no. 1, p. 1096, 2019.
- [13] F. K. Došilović, M. Brčić, and N. Hlupić, Explainable artificial

- intelligence: A survey, in *Proc. 2018 41<sup>st</sup> International Convention on Information and Communication Technology, Electronics and Microelectronics (MIPRO)*, Opatija, Croatia, 2018, pp. 210–215.
- [14] D. Gunning and D. W. Aha, Darpa's explainable artificial intelligence (XAI) program, *AI Magazine*, vol. 40, no. 2, pp. 44–58, 2019.
- [15] F. Xu, H. Uszkoreit, Y. Du, W. Fan, D. Zhao, and J. Zhu, Explainable AI: A brief survey on history, research areas, approaches and challenges, in *Proc. 8<sup>th</sup> CCF International Conference on Natural Language Processing and Chinese Computing*, Dunhuang, China, 2019, pp. 563–574.
- [16] G. Montavon, A. Binder, S. Lapuschkin, W. Samek, and K. -R. Müller, Layer-wise relevance propagation: An overview, in *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning*, W. Samek, G. Montavon, A. Vedaldi, L. K. Hansen, and K. -R. Müller, eds. Cham, Switzerland: Springer, 2019, pp. 193–209.
- [17] H. Yu, Z. Shen, C. Miao, C. Leung, V. R. Lesser, and Q. Yang, Building ethics into artificial intelligence, in *Proc. 27<sup>th</sup> International Joint Conference on Artificial Intelligence (IJCAI'18)*, Stockholm, Sweden, 2018, pp. 5527–5533.
- [18] M. Ancona, C. Oztireli, and M. Gross, Explaining deep neural networks with a polynomial time algorithm for shapley value approximation, in *Proc. 36<sup>th</sup> International Conference on Machine Learning*, Long Beach, CA, USA, 2019, pp. 272–281.
- [19] M. T. Ribeiro, S. Singh, and C. Guestrin, “Why should I trust you?”: Explaining the predictions of any classifier, in *Proc. 22<sup>nd</sup> ACM SIGKDD International Conference on Knowledge Discovery and Data Mining*, San Francisco, CA, USA, 2016, pp. 1135–1144.
- [20] B. Friedman, Value-sensitive design, *Interactions*, vol. 3, no. 6, pp. 16–23, 1996.
- [21] 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.
- [22] 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.
- [23] Q. V. Liao, D. Gruen, and S. Miller, Questioning the AI: Informing design practices for explainable AI user experiences, in *Proc. 2020 CHI Conference on Human Factors in Computing Systems*, Honolulu, HI, USA, 2020, pp. 1–15.
- [24] M. -Y. Kim, S. Atakishiyev, H. K. B. Babiker, N. Farruque, R. Goebel, O. R. Zaïane, M. -H. Motallebi, J. Rabelo, T. Syed, H. Yao, et al., A multi-component framework for the analysis and design of explainable artificial intelligence, *Machine Learning and Knowledge Extraction*, vol. 3, no. 4, pp. 900–921, 2021.
- [25] B. Shneiderman and H. Hochheiser, Universal usability as a stimulus to advanced interface design, *Behaviour & Information Technology*, vol. 20, no. 5, pp. 367–376, 2001.
- [26] 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.
- [27] S. Atakishiyev, M. Salameh, H. Yao, and R. Goebel, Explainable artificial intelligence for autonomous driving: A comprehensive overview and field guide for future research directions, arXiv preprint arXiv: 2112.11561, 2021.
- [28] E. Tjoa and C. Guan, A survey on explainable artificial intelligence (XAI): Toward medical XAI, *IEEE Transactions on Neural Networks and Learning Systems*, vol. 32, no. 11, pp. 4793–4813, 2020.