Contextualizing User Perceptions about Biases for Human-Centered Explainable Artificial Intelligence

Chien Wen (Tina) Yuan
tinayuan@ntu.edu.tw
National Taiwan University
Taipei, Taiwan

Nanyi Bi
nanyibi@ntu.edu.tw
National Taiwan University
Taipei, Taiwan

Ya-Fang Lin
yml5563@psu.edu
Pennsylvania State University
State College, U.S.A.

Yuen-Hsien Tseng
samtseng@ntnu.edu.tw
National Taiwan Normal University
Taipei, Taiwan

ABSTRACT

Biases in Artificial Intelligence (AI) systems or their results are one important issue that demands AI explainability. Despite the prevalence of AI applications, the general public are not necessarily equipped with the ability to understand how the black-box algorithms work and how to deal with biases. To inform designs for explainable AI (XAI), we conducted in-depth interviews with major stakeholders, both end-users (n = 24) and engineers (n = 15), to investigate how they made sense of AI applications and the associated biases according to situations of high and low stakes. We discussed users' perceptions and attributions about AI biases and their desired levels and types of explainability. We found that personal relevance and boundaries as well as the level of stake are two major dimensions for developing user trust especially during biased situations and informing XAI designs.

CCS CONCEPTS

• **Human-centered computing** → **Empirical studies in collaborative and social computing**; **Empirical studies in HCI**.

KEYWORDS

Artificial Intelligence, Human-Computer Interaction (HCI), Explainable AI (XAI), Human-Centered Computing, Explainability, Transparency, AI bias

ACM Reference Format:
Chien Wen (Tina) Yuan, Nanyi Bi, Ya-Fang Lin, and Yuen-Hsien Tseng. 2023. Contextualizing User Perceptions about Biases for Human-Centered Explainable Artificial Intelligence. In *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems (CHI '23), April 23–28, 2023, Hamburg, Germany.* ACM, New York, NY, USA, 15 pages. https://doi.org/10.1145/3544548.3580945

CHI '23, April 23–28, 2023, Hamburg, Germany

ACM ISBN 978-1-4503-9421-5/23/04…$15.00
https://doi.org/10.1145/3544548.3580945

1 INTRODUCTION

Artificial intelligence (AI) techniques have seen great advancements over these years, rendering many smart applications no longer media hypes or science fictions but already commonplaces in people's daily lives. From consumer technologies like texting autocorrection, spam email filtering, or chatbot services to applications in professional domains like healthcare, financial services, transportation, or employment decisions, users have received or appropriated AI assistance to solve problems or make decisions (for a scoping review, see [1]). While it is important to continue revolutionizing AI model training and enhancing learning performance using computational techniques for overall accuracy, this study argues that we also need to better understand how users perceive AI-mediated applications and their results for a more explainable and transparent human-AI interaction, especially when AI presents biased results. Despite the prevalence of AI applications, the general public is not necessarily equipped with the ability and literacy to understand how the black-box algorithms work that help them accomplish different tasks [32]. There may be no clear mapping between technology features and how AI mediates the systems (e.g., how Facebook algorithmic content recommendation works has no correspondent features on Facebook for users to explore). Under such circumstance, if AI produces biased results, it can reduce users' trust and inhibit adoption.

Bias in AI can take place in the initial stage of data collection and get passed along in the training processes, such as model evaluation, model processing, and model deployment [29]. When AI models are deployed in different socio-cultural contexts, the computational biases may also be transformed to social ones like biases towards a certain gender, occupation, age, race, or culture. For instance, Google Photos was once found to have labeled two African Americans as gorillas in their application[1], which showed that biases in data collection or model training can yield significant consequences like discrimination, especially in areas that have long-standing racial issues. Revealing how the results are derived may help users understand and evaluate the biases. And this makes imposing a top-down transparency guideline for AI modeling less viable because it may not reflect the unique sociocultural contexts and authentic use cases and it may not cover what users would like

[1] https://www.forbes.com/sites/mzhang/2015/07/01/google-photos-tags-two-african-americans-as-gorillas-through-facial-recognition-software/?sh=789d9178713d

to know about the model and how would they like to know it under different situations.

Our study proposes that AI-mediated applications may be perceived differently depending on their stakes and influences on users' lives. It is important to differentiate these dimensions to contextualize explainability and transparency for AI. This is especially important when AI shows biased results that yield collateral sociocultural consequences. As more and more systems programmed with AI are widely adopted in different domains and contexts, such as facial recognition algorithms and hiring/firing decision-making algorithms, users' lives are influenced whether knowingly or unknowingly [31]. It can be alarming if users do not know how the suggestions or decisions are derived yet still follow them, leading users to potentially follow the biased AI suggestions and make erroneous decisions. To fill this gap, our study intends to investigate how users perceive AI biases depending on the level of stake through a social and interactive lens.

Essentially, our research is in line with the call for more transparent AI, or the domain of explainable AI (XAI). Explainable AI promotes sets of tools, techniques, and algorithms to enhance users to develop an understanding, trust, or effective management of AI systems [15], aiming at the goals of achieving fairness, accountability, and transparency of AI models and applications [39]. The current XAI research space is largely supported by algorithmic accountability and interpretable modeling [2, 9, 25]. More and more researchers started to argue that making opaque computing algorithms transparent to users can enhance trust and has started to gain importance in various domains, such as analogical reasoning in medical diagnosis [20, 43], judicial reasoning for criminal, administrative, or civil cases [4, 8], or hiring decision-making in human resource [42].

Drawing on the human-centered XAI, we argue that transparency and explainability are not just traits of AI models but should be system affordances that invite users' interaction with the models/systems. To involve users in the picture, it is important to understand how users consider opacity in biased AI results and demand for transparency. From a social and interactive angle, our study attempts to complement previous computational XAI studies by involving users' perceptions in making sense of sociotechnical AI biases in scenarios of different stake levels. By sociotechnical AI biases, we are referring to the biases inherent or created by the data and/or algorithms but have an actual impact on human users and society.

With a bottom-up and user-centered approach, we conducted an interview study with general end-users (n = 24) and engineers (n = 15). The end-users were people who had experience in using various information and communication technologies (ICTs) but were not involved in AI development and deployment industries, whereas the engineers were those who worked in those industries and had first-hand and meaningful experience of interacting with AI. We found that the levels of stake and personal boundaries contextualize users' demands for AI explainability. The paper intends to make the following contributions:

- Understanding how users make sense of AI biases and how they expect the level of explainability to be;
- Exploring an additional and under-discussed dimension in XAI: AI biases in high-stake situations like medical diagnosis and low-stake situations like Google translation results;
- Aligning the types of biases and corresponding explanations with users' concerns;
- Offering insights into XAI design.

2 RELATED WORK

2.1 Explainable AI: An HCI Perspective

AI models, be they based on decision-trees, rule-based algorithms, or machine/deep learning (ML/DL), are difficult for average users to interpret. Those that are run on ML/DL are even more puzzling because the data structure is nested and nonlinear. The results or decisions yielded by AI models are not directly interpretable and understandable by users due to the opaqueness of the algorithms and a confounded notion of correlation for causation by users regarding the AI results [27]. While some researchers point out that algorithms themselves are neutral and free of biases and we only have to evaluate the results based on model performance and accuracy [36], others argue that ethics and fairness are entitled to users for they rely on the models to make decisions, especially in critical situations like granting parole or mortgage loan decisions [28].

To answer the call for more transparent and responsible use of AI, XAI researchers propose that AI algorithms should come with expressive and interpretable explanations to improve users' understanding and confidence in decision-making so that they can use the AI-generated outcomes to make justifiable decisions [3]. An auditable and provable process ensures transparency, trust, and fairness in AI models [15, 39]. In addition, XAI models may enhance users' control with greater model visibility over potential unknown flaws so that users can know where to debug or how to improve the models [3, 37]. Ultimately, it is suggested that new discoveries and insights may be derived from the explainable model results [3].

XAI researchers have tried to address opaqueness and enhance explainability in AI models with different methods [2]. Computational and mathematical approaches propose revealing interpretable learning processes with clearer derivation sequences of model decision [10, 34, 37]. Additionally, XAI research highlights how data or models are handled at which stage to ensure fairness from sources to prevent downstream biases [40]. However, despite the endeavors, practitioners such as engineers still found it difficult to operationalize the guidelines along the lifecycle of AI model development and deployment and sometimes the multi-dimensional and complex concepts (e.g., fairness, accountability, transparency, and privacy) are operationalized in uni-dimensional and simple ways in the models, such as using a checklist with a binary form of yes/no to indicate whether gender ratio is equally represented in the data for "gender equality" [30]. More importantly, general users may still feel at dark about algorithmic explanations based on technical approaches.

On the other hand, the principles revolving around XAI, such as fairness, accountability, transparency, and privacy, are all ethical concerns over AI use. "Ethics" is a fundamentally sociocultural concept, even when it relates to technical systems. In other words, XAI scholarship is essentially sociocultural and sociotechnical. In

light of users' needs, scholars start to call for integrating social dimensions and end-users' perspectives in the XAI research program [21, 24, 28]. Providing "layman" explanations about parameters and classifiers in an AI model is suggested [9, 25] to free from errors associated with "technosolutionism" [38].

Researchers have identified various stakeholder groups of XAI, including 1) AI developers who design and develop AI systems, 2) AI regulators who certify AI systems, 3) AI managers who oversee the deployment of AI in organizations, 4) AI users who use AI systems, and 5) people influenced by AI-based decisions [26]. Although more and more studies gradually started to investigate users' role in XAI, the "users" in the research are still relatively "technology-savvy," such as user experience researchers or designers [21], product/program managers, data scientists, researchers, or consultants [5, 23]. These users are either AI developers, AI managers, or AI regulators. Their understanding about how AI may function can be richer than the general AI users or people affected by AI-based decisions, who are relatively less studied in the literature. Lay people could be influenced by AI-based systems more than other experienced stakeholders because they may have less knowledge about AI systems and less awareness of the potential risks [47]. For example, a medical AI system with an overall 95% accuracy rate may achieve 99% accuracy on Asians but only 70% on Africans due to a lack of African data. An African AI developer may question the model's applicability in Africa, whereas an African AI novice may lack an awareness of this bias. While these previous studies have already shown that it can be difficult for engineers who deal with different processes of model training to trace back to the previous stage and debug the hidden layers of the learning algorithms [27], average end-users can find it hard to wrap their head around how AI works, especially when AI models present biased results.

To provide explanations suitable and useful for end-users, it is important to learn how opaqueness and transparency is construed based on users' expertise, preferences, and other contextual variables [3]. Users' understanding and trust of the system partly depends on the way they interact with the technology. As the ultimate goal of a system being user adoption, the solution to XAI should not be just "more AI" [28] but needs to incorporate an HCI angle to understand and empower users and their interactions with the systems. An integration of HCI and XAI can help develop transparent AI systems [2, 3, 21, 24, 28]. Many researchers have started to see the importance of this research trend but also pointed out that it has not yet well implemented and studied [28, 37].

Researchers propose that we have to understand users' mental models, or how users interact with and make sense of AI systems [12]. Gero et al. [13] found that users tended to overestimate AI's performance and demanded explanations when they got abnormal results. Users may hold prior biases and social expectations towards AI [28]. De Graaf and Malle [7] point out that people tend to assign human-like traits to intelligent agents and may expect explanations about AI to have the same traits. Wang et al. [44] leveraged HCI theories about human reasoning and attribution tendency to inform XAI and found that when users receive unexpected results from AI, they want AI to help them verify alternative hypotheses regarding what actually goes wrong. Additionally, users want to have data-driven reasoning to eliminate confirmation bias and they demand a coherent set of AI features, an access to source or supplementary data for trust, an understanding of how Bayesian modeling works, and alternative explanations suggested by the system. Studying multiple stakeholders, Brennen [5] pointed out several motivations for users to demand explainability, including debugging models, detecting bias, and building trust.

Since end-users are the major stakeholders who eventually adopt AI systems or experience the results of such systems, including end-users reflects a human-centered and contextualized direction in XAI. With the advances of AI technologies and scholarship in XAI, we propose to extend the previous literature that investigated everyday use of AI [28, 37] but further contextualize situations when AI makes biased suggestions and examine users' perceptions and reactions to such circumstances as biases in AI have emerged to be a salient issue that may cause sociotechnical consequences.

2.2 Uncovering Biases for Explainability and Transparency

One issue that results from algorithm opaqueness and raises the need for explainability is the biases that come with AI modeling and results. Biases can happen in different stages of AI development, from data collection to system deployment [29, 48]. For example, when the raw data sets used to train the models are biased in terms of data structure (e.g., more male data and less female data), such technical biases can be transformed to sociocultural ones (e.g., gender bias) and the level of consequences of such bias may vary depending on the stakes of the application scenarios (e.g., judicial systems vs. Google search results).

Take Amazon's recruiting system for example[2]; according to Reuter's report, the system integrated previous applications and hiring biases so that the model results favored male applicants over female ones over the years. Another case is a risk management system that reinforces patterns of racism and classism in criminal justice [4]. The biases in the above-mentioned scenarios originated in data sets, got amplified through modeling, and resulted in serious consequences in applications that jeopardize users' rights and lives [5]. While it is crucial to address technical biases in model development to prevent from further reinforcing them into social ones, it is also important to start from the bottom-up to know how users make sense of bias in AI results to inform explainability design.

In order to address the issue of explainability, many public and private organizations propose guidelines or tenets to ensure fairness in AI models, such as European Union[3], Microsoft[4], or Google[5]. The common denominators of these guidelines include fairness, accountability, transparency, and user privacy for AI [17]. The guidelines do not instruct how model development and deployment processes should accommodate and adjust in a more dynamic way, which increases the difficulty in applying them in various contexts [23]. More importantly, with little justification, it is unclear how some of the guidelines are derived and feasibly useful to actual users [44].

[2]https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G
[3]https://ec.europa.eu/futurium/en/ai-alliance-consultation
[4]https://www.microsoft.com/en-us/research/project/guidelines-for-human-ai-interaction/
[5]https://ai.google/responsibilities/responsible-ai-practices/

Since AI-enabled applications can be found in everyday life and many of them may present different types of bias, this study takes interest in what users' perspectives are when they learn about and encounter AI biases and how their notion of explainability is developed around AI biases. To extend the line of work in XAI and human-AI interaction, a more nuanced focus on studying how users make sense of AI and AI biases can help uncover users' trust mechanisms and inform transparent and explainable AI interface designs.

We propose to investigate what constitutes explainability according to two important stakeholder groups: end-users and engineers (broadly referring to those who deal with data or modeling in the AI pipeline). We hope this approach can address the challenges identified in the previous works, such as a difficulty to operationalize complex concepts like fairness, transparency, and explainability in AI modeling for engineers or high thresholds that prevent end-users from understanding AI. We posit that a more contextualized, user-centered approach that reflects stakeholders' needs facing AI biases may help inform designs of XAI user interface.

In addition to the various considerations and dimensions of the existing XAI methods, such as algorithmic adequacy, users' usability concerns, or prevention from user vulnerability [14], we also propose another dimension to contextualize bias and associated explainability in AI models: the level of stake for the application scenarios. Adadi and Berrada [3] argue that in some low-stake contexts, such as an AI system for targeted advertising placements, users may not need high and detailed model interpretability. In contrast, in situations like public transportation, medical conditions, or criminal justice where the stakes are high, higher and more detailed interpretability is demanded. In other words, the required level of explainability should be adjusted based on the stakes and consequences of the situations. This again reflects the insufficiency and inflexibility of cookie-cutter benchmarks and leads to another reason why a bottom-up approach is needed to understand how users judge and interpret different use contexts for AI explainability. We propose our research questions:

RQ1: How do stakeholders, both end-users and engineers, make sense of AI and some potential associated sociotechnical biases? How do end-users seek and engineers provide explanations of AI models especially when biases are present?

RQ2: How do stakeholders, both engineers and end-users, perceive biases and associated level of explainability in contexts of high or low stakes?

3 METHOD

3.1 Procedure and Participants

The first three authors conducted one-on-one, in-depth interviews with both general end-users (n = 24) and AI product developers (n = 15).

For general end-users, we recruited our interviewees from both social media platforms like Facebook and a local Bulletin Board System (BBS), where research opportunities are listed. The recruiting criteria included: 1) having experience using technologies that they thought used AI techniques; 2) not being part of the AI modeling or implementing processes; and 3) aged 18 and above. We tried to gather a relatively diverse sample in terms of gender (12 females, 12 males), education level (11 graduate, 13 undergraduate), age (from 24 - 55; M = 34.75; SD = 6.95), income level (M = USD$24,506; median = USD$17,591, SD = USD$15,263), and occupations, including freelancers, marketing, sales, health professionals, designers, editors, consultants, scientists, customer service representatives, educators, secretaries, administrators, finance, law, etc. We also asked participants' perceived understanding of AI on a Likert scale of 1-5 (1 being knowing very little and 5 being knowing very well; M = 2.81; SD = .76).

For engineers, we posted our recruiting messages in AI-related Facebook groups and snowball-sampled via personal connections. We recruited participants that covered various industries spanning medical care, finance, manufacturing, or consumer electronic products. Our participants took care of different stages and tasks of AI product development ranging from need-finding to deployment. They were in different pay grades ranging from interns to CEO of startups or large corporate. We conducted one-on-one, in-person interviews with our engineer interviewees except for 3 online interviews with participants who work in the U.S.(E06, E08, E13) (see Table 1.) For those participants, the interviews were conducted over video conferencing tools of the interviewee's choice.

For both end-users and engineers, the interviews ranged between 40 to 90 minutes in length. All interviewers were done in Mandarin. We first explained the purpose of our research, explicitly told the participants that they could skip the questions if they did not want to answer, and had participants sign a consent form before starting the interview. After the interview, each end-user participant was compensated NTD $300 (roughly USD $10) for their participation. Considering the average salary of the engineers, we compensated our engineer participants NTD $500 (roughly USD $18) for their time. Our study was approved by the X University IRB (Number to be added).

3.2 Interview Protocol

There are two parts of interview questions for both end-user and engineer participants. The first part pertains to their general experiences and attitudes towards using and/or developing AI-mediated services or systems they identified. For end-user participants, we asked questions about their understanding of AI and interactions with different AI products. The protocol also contains questions regarding why our participants think the applications/technologies implement AI techniques and how these techniques may work, such as using their behavioral logs for recommendations or predictions on the sites. We also probed if our participants had ever come across any odd AI-mediated outcomes and how they accounted for them. We did not limit the types of AI applications in the protocol because our end-user participants might not necessarily interact with a specific kind of AI. In addition, we probed their general understanding of AI applications and the perceived potential biases embedded in the model development and deployment. For engineers, in addition to the questions that we asked our end-user participants, we focused on their individual and collaborative experiences of developing AI products at different stages and their perceptions about AI biases.

The second part of the interview used a scenario-based protocol that covered six biased AI applications we drew from the news,

	Gender	Age	Years of AI experience	AI dev. tasks	Job title	Domain	Education
E01	Male	35-44	12	1234	CEO	AI platform	PhD
E02	Male	25-34	1	1234	Software Engineer	Security, advertisement	Master
E03	Male	20-24	3	123	ML engineer, Academic researcher	Food	Undergraduate
E04	Male	35-44	2	123	Software Engineer	Finance	Undergraduate
E05	Male	25-34	1	1234	Software Engineer	Manufacturing	Master
E06	Male	35-44	2	123	Data Scientist	Finance	Master
E07	Male	35-44	16	1234	Technical Associate	Security, IoT, consumer electronics	Incomplete PhD
E08	Male	20-24	0	1234	CV intern	Manufacturing	Master
E09	Male	25-34	3	1234	Sr. Engineer	Medical care, consumer electronics	Master
E10	Male	35-44	4	1234	Data Scientist	Manufacturing, finance	Master
E11	Male	35-44	5	1234	AI team Lead	Robot	Master
E12	Male	35-44	12	1234	Engineering Manager	Medical care, manufacturing, consumer electronics, research	PhD
E13	Female	25-34	3	23	AI/ML Research Scientist	Medical care	PhD
E14	Male	25-34	4	1234	R&D Lead	Medical care	Master
E15	Male	25-34	3	123	Product manager	Medical care, advertisement	Master

Table 1: The demographic and background information about engineer participants.
***AI development tasks include (1) need finding and problem definition, (2) data collecting and processing, (3) AI model building and evaluation, (4) AI model deployment**

including high-stake ones like credit limit decision (credit card algorithm sparks gender bias against female applicants[6]), criminal face detection (computer vision algorithm contains racial discrimination[7]), and employee hiring (recruiting-engine algorithm shows bias against women[8]) as well as low-stake ones like photo searching (search engine results of CEO always show male figures[9]), text autocomplete (smart texting services give gender-based autocomplete suggestions[10]), and machine translation across languages (smart translation services miss gender and cultural sensitivity when translating contents across languages[11]). The major distinction between high- and low-stake scenarios is whether the AI results were used to make inferential decisions that had critical influences such as people's finance, job rights, or civil rights. All low-stake scenarios displayed biased results but no inferential decisions were made based on the results. These biased scenarios were chosen for the following reasons: 1) they were popular applications/materials drawn from the press so that our participants were more likely to hear about and resonate with them; 2) they concern different types of sociotechnical biases like gender or race biases.

We first asked if our participants had experienced any odd and biased results they got from any AI services and then used the scenario-based prompts to guide them to talk about the sociotechnical biases that we focused on in the study. For each participant, we might choose different scenarios depending on which scenarios our participants had experiences in or felt they could elaborate on, but we always included both high- and low-stake applications in the interview. The reason why we provided a list of scenarios for both high- and low-stake was because when a participant could not relate to a specific scenario of either high- or low-stake, we used another scenario of the same stake level to carry on the interview. Note that the scenarios were used to elicit participants' opinions or experiences instead of counterbalanced experiment conditions. We asked the end-user participants how they perceived the biases and associated stakes in the scenarios, how they would respond to these situations, and how they thought about the AI systems that yielded biased results. For the engineer participants, we also asked the same set of questions and additionally probed what they thought went wrong in the AI development process and how they would implement these applications and deal with the biases had they been the developer.

3.3 Data Analysis

We audio-recorded the interviews and later transcribed them. The first three authors first individually and then collaboratively coded the transcripts using Atlas.ti Cloud Service, which allows us to co-edit the notes, codes, and themes.

We started with coding the data of the end-user participants. During the first round of open coding, the first three authors coded individually and met regularly to ensure we had coherent interpretations of the codes and discussed emerging codes and themes. We exchanged perspectives about the codes for a more comprehensive

[6]https://www.washingtonpost.com/business/2019/11/11/apple-card-algorithm-sparks-gender-bias-allegations-against-goldman-sachs/
[7]https://sitn.hms.harvard.edu/flash/2020/racial-discrimination-in-face-recognition-technology/
[8]https://www.reuters.com/article/us-amazon-com-jobs-automation-insight-idUSKCN1MK08G
[9]https://www.washingtonpost.com/business/2019/01/03/searching-images-ceos-or-managers-results-almost-always-show-men/
[10]https://www.reuters.com/article/us-alphabet-google-ai-gender-idUSKCN1NW0EF
[11]https://analyticsindiamag.com/google-translate-has-gender-bias-and-it-needs-fixing/

understanding of the data. In the following rounds of deeper coding, we iteratively refined the codes from the first round and grouped them into categories. To this end, we swapped coded transcripts and reviewed the quotes of each code identified by others to ensure that we agreed with one another's codes. Then we reviewed the data again at the code level to ensure every quote under each code indeed reflects what the code denotes. During this stage, we also refined the codes and categorized codes under salient themes.

Then, we coded the engineer participants' transcripts following the same procedure. We tried to identify common themes we found in both participant groups. For example, some of the code names were identified first in the engineer data and may be carried over to the end-user data if the concepts were transferrable. Then, the sets of end-user and engineer data were connected using common themes. After all themes from both participant groups were stabilized, the first two authors conducted axial coding to connect and structure the meaningful themes, which we report in the following section.

4 RESULTS

In this section, we report the salient themes that emerged from the axial coding that help answer our research questions. We use "U" to denote the end-users and "E" the engineers for the quotes.

4.1 Making Sense of Biases in AI

Our first research question pertains to the sense-making process. Specifically, we are interested in how end-users as well as engineers make sense of AI in terms of its limitations, especially under the circumstances of biases. The results can be categorized into two main themes: 1) Bias being part of reality, and 2) Attribution of AI bias sources.

4.1.1 Bias being part of reality. When asked about bias, many interviewees commented that biases are part of reality and reflect the social and cultural norms where the data are collected or through which the AI is applied to some extent. For instance, U02 commented on the biased results of Google image search, where the proportion of gender in specific occupations (e.g., bartender, CEO, and nurse) is different from the actual statistics. Despite the results being gender-stereotypical, our participants did not necessarily consider such incidents surprising or misleading.

> "I am not surprised by this result. The image search just truthfully reflects the gender stereotypes people have. I don't think there is anything it [Google] can do to change the underlying bias in our minds; it's just a reflection after all. If our mind changes, it will change consequently. " (U09, F)

In light of such biases, our participants pointed out that the AI-mediated results should be interpreted based on the contexts. For instance, one engineer mentioned that if the data were obtained from Middle Eastern countries, such as Saudi Arabia, the masculine culture may be captured in the data and reflected in the results.

> "I think the data source is critical. For instance, if we collect textual data from newspapers, magazines, or websites in Arabic countries, it will naturally lean toward more patriarchal ideas and biases. My guess is that cultural differences will affect the Google translation results." (E05, M)

Users' sense-making about the biased results may also reflect their sociocultural heuristics. Therefore, it is possible that people may unknowingly have biased expectations to begin with, and are thus less aware of the bias in AI. For instance, an end-user commented on an example of a gender stereotype where females tended to do more impulsive consumption, which was why an AI system granted a woman a lower credit line than her male counterpart:

> "My explanation for why the woman had a high credit score but was granted a low credit line is that the data of the financial history of females show more impulsive and unnecessary expenses than males. This is a rather universal phenomenon; females have more impulsive purchases than males." (U12, M)

While not every user was aware of their own biases and AI's biases, the benefited social group may turn oblivious to the bias and maintain their usage.

> "I benefited from such a system [that is biased towards males] so I will keep using it for the vested interest...even though I know I shouldn't." (E15, M)

To the same issue of gender stereotypes, an engineer pointed out that it is futile to blame AI-mediated systems. According to him, all users and the sociocultural context need an overhaul.

> "My inclination is to start from scratch...I mean education. We need to teach people to stop saying biased words before building a fair system. Blaming the system is a red herring that distracts us from the real issue. The system really just takes in whatever we feed them...we users are the accomplices" (E15, M)

Interestingly, our participants pointed out that the biased results by AI actually help them reflect the hidden biased structures our society may have been failing to externalize and embody.

> "I think the biased image search results actually reflect biases on a deeper level that we need to fix. Maybe it's the news media that really fosters the bias as they focus more on reporting male CEOs. So it's not really AI's fault." (U19, M)

The fact that many people consider AI's biases being part of reality and a mirror of a society's social and cultural norms drives our participants to argue that biases should be changed from a social perspective, rather than a technical perspective. And according to our participants, the fact that bias is sociotechnical, it evolves with time as well.

> Everyone has to know such and such stereotypes exist to tweak their data collection process and model building. In addition, I don't think a bias-free AI model exists forever; it has to evolve with society. For instance, I can change all the "he"s in my textual data to "he/she" to include females, but what about gender neutrals? The system has to keep up with the society." (E13, F)

> "We are seeing all these biased image search results because, although gender equality has been improved

> around the past 10 years, what we are seeing today goes beyond what has been accumulated in the last decade. The idea of gender equality was still not a thing back then. In the next 20 years, we might see something very different because the database is constantly being updated. I can't just reduce the database to fit what's mainstream today, because it might limit what I get." (U03, F)

In short, our participants had noticed that AI may generate biased results and they reckoned that such biases reflect the social reality to a large extent. Note that if the participants' own perceptions were aligned with AI's biases, we observed a tendency that they were less likely to spot the biased issue. Generally, revealing bias was considered a good thing in that it triggers public awareness and change, but they did not think it is an issue that could be solved by AI or the engineers alone. Addressing a certain bias requires more fundamental social change and the AI system should evolve with time.

4.1.2 Attribution of AI bias sources. When spotting biases in AI-enabled results, our participants also attributed AI biases to different sources, which can be roughly categorized into the following types.

External sources: It's all "AI"'s faults. Compared with engineers who may understand better how AI biases emerged, our end-user participants attributed AI biases to an erroneous algorithm, but they could not delineate how the algorithms have gone wrong.

> "I think it has something to do with the facial features and how the algorithms take them in? Maybe a Caucasian's facial feature is more prominent on a computer screen? I don't know, but I feel it may be a technical problem." (U06, F)

Some end-users also associated the algorithm bias with the developers and the engineers. Engineers, although agreeing that algorithms could be a source of bias, provided more detailed accounts of how algorithms could have gone wrong. For instance, one engineer listed how framing effect bias could lead to a problematic algorithm.

> "I think the autocorrection is biased because the designers are biased. They have stereotypes for certain groups. I won't like it, but I will probably adjust myself to using it because it is hard for the big companies to change." (U15, F)
> "People who work in the pipeline possess different domain knowledge and may have different ways of thinking, therefore they are limited in framing the question for modeling. For instance, a programmer could be good at programming but not doing business. The model developed by this programmer could have gone wrong from the beginning." (E14, M)

In addition to the algorithms and the engineers behind the algorithms, the biased results were perceived to be associated with the company that developed the AI models and the algorithms. Some of the end-user participants considered that the companies had purposefully induced bias for their own benefits.

> "The AI HR system that is biased toward males does not really surprise me. When the whole country or culture is inclined to favor certain groups, the company will naturally give them more weight in data collection or in algorithm design, making them easier to stand out for qualification." (U24, M)

Internal sources: To err is human? There were also some interviewees who did not realize the existence of biased AI results and instead put the responsibility on themselves. They thought they had to improve their own usage to fit the AI applications. Perhaps not surprisingly, almost no engineer participants held negative internal attribution when they experienced biased AI results.

> "I do get some irrelevant or unexpected results when I search for images, or just general searching. But then I think it's my fault that I did not use the correct keywords. The search engine nowadays is pretty advanced." (U01, M)

Facing biased results, some end-users even blamed themselves for not having enough critical thinking and logical reasoning abilities when using AI.

> "Users don't want to know how the results were derived. We just use it. If the results are biased and the users are not aware of them...it's we who should take responsibility. We need to improve ourselves. As users, we also need to work on our own judgment and logical reasoning because AI is just not that advanced yet." (U21, M)

Our data showed that users attributed external and internal sources when they spotted AI biases. For the internal sources, end-users most likely blamed themselves for not having advanced literacy to know how AI works. When biases emerged, it was most likely that they did not use the technology right. In other words, biases may thrive if users take themselves to blame. On the other hand, the external attribution of bias related to AI involves the following sources: 1) perceived erroneous algorithms, 2) engineers' lack of domain knowledge, and 3) companies that launch AI-enabled services. We would like to note that there are palpable differences regarding the extent to which end-users and engineers elaborated how they derived the attributions, most likely due to their different familiarity with and literacy about AI. Those differences further create challenges for an explainable AI, which we discuss in the next section.

4.2 In Search of Explainability

Our second research question addresses the challenges emerging during sense-making, and the data suggest that there are misaligned perceptions between end-users and engineers regarding AI biases. We elaborate on how explainability is construed with three themes: 1) users' desired information disclosure about AI and potential information overload, 2) users' verification process, and 3) ways of communicating explainability.

4.2.1 Information disclosure vs. overload. While it may be good to have a clear understanding of how AI works for XAI designs, it could be tough to strike a balance between being open and clear about what a XAI application does and information overloading on

the users' part. This is particularly true when the user lacks a basic understanding of AI and its relevant processes.

> "Sure you can provide information about where the data come from and its demographic information and everything, but it could potentially cause information overload. If you do this over and over again, people would just ignore it, in which case the information you provide becomes useless." (U08, F)

The overloading issue may come at a social cost for end-users. The line between what to disclose and what not can be tricky.

> "I don't think I would want the bank to tell me why they rejected my credit card application by saying something like 'sorry, your application is rejected because you changed your job recently.' It feels like I am looked down upon and they are making an effort to make sure that I know." (U15, F)

While it is ideal that the AI models and the AI applications are made transparent to the users, our results suggest that users may feel overwhelmed by such glaring transparency. Over time, they may even ignore such disclosure. Sometimes, such disclosure may not be desired if the AI algorithms use socioeconomic or demographic variables as parameters but ignore the actual users' perceptions. When dealing with biases and enhancing AI explainability, researchers and practitioners need to consider the fine line both cognitively and socially.

4.2.2 Users' verification process. While there may be limited explainability features in most AI-enabled services or applications, the end-user participants used several strategies to infer and verify how AI may work, including using their prior experience as a benchmark or relying on third-party organizations.

Our end-user participants pointed out that they evaluated AI model performance based on their prior interactions with the same AI. When inconsistency came up, they could not but have to resort back to their own judgements.

> "I will use my past experience to make a judgment about the AI application. If what the AI provides matches my past experience, and it solves the current issue, then I will deem the AI correct and of good performance. But if one of them is inconsistent, I will rely on my own experience and judgment." (U10, M)

If the user has no relevant experience, they may feel dubious about using the AI application because the use scenarios go beyond their prior understanding. It happens in some innovative or groundbreaking scenarios.

> "I am still suspicious of the AI jury thing, because it goes beyond what I have experienced. And it's natural that you feel suspicious about something you have never encountered in real life." (U19, M)

In addition to using personal experience and the outcome as verification measures to evaluate an AI application or to make sense of biases, our interviewees who do not possess the required knowledge to understand AI mechanisms and develop proper expectations towards AI relied on external sources, such as other trustworthy parties.

> "There may be problems that AI or the ones creating and maintaining the AI haven't discovered, so there has to be someone from the third parties [e.g., professionals], someone external to audit and expose the problem. In the case of gender inequality, maybe someone who is an expert in this field to actively investigate potential problems that the AI experts are not able to discover on their own." (U19, M)

In addition to third-parties, users mentioned that crowd-sourced reviews from peer users may also be of help.

> "Apart from reviews or evaluations from experts, I think reviews from other users who have used the system may also work." (U14, M)

One user commented that it may be possible to leverage the power and knowledge of professional communities to do the audit and make the algorithm transparent.

> "Sometimes neither the user nor the engineer could tell the issue, so it could be helpful if the algorithm is made public for those who understand to examine it." (U08, F)

To summarize, our users either rely on their own experience of interacting with AI for verification or other professional organizations and peer groups. Relying on one's own experience comes with the limitations that the users may not be able to deal with biased or unexpected situations should the AI exceeds the routine or they may feel difficult to make sense of novel applications with which they have no experience before. Alternatively, reliable third parties or peer groups may provide assistance.

4.2.3 Ways of communicating explainability. In order to foster system transparency and users' understanding of it, our interviewees also discussed different ways of communicating explainability, which we categorized into the following strategies.

Providing approachable narratives. One way developers of AI applications employ to communicate the models/products to their clients who do not possess domain knowledge is to cater their narratives to suit the audience's understanding.

> "Some of my clients are not AI experts and cannot understand all those technical jargons, in which case our product managers have to come up with a story for them to make sense of our products. The stories are the only way to facilitate the clients to make a decision. So before they come up with the story, the product managers and consultants as well as the engineers need to put our heads together to make the story clear enough for the clients, but also make technical sense. " (E15, M)

Making the intangible tangible. Another way our participants mentioned may be of help to solve the black-box issue is to make AI model and its functions concrete. One participant made an analogy of the ways they interacted with people with the ways they employed AI based on interaction history or cues they collected along the interaction process. When non-verbal cues are used to develop

an understanding and trust towards other interactants, our participants mentioned that details about AI are currently unavailable but may be used to solve the intangible issues with AI.

> "I can make judgments on one's credibility based on body language and facial expressions, but this does not work with machines I think. I feel I can't trust it because of intangibility. I would rather the system tell me how they get the results based on such and such information they acquired from me, even if they used my private data. It's at least a win-win situation." (U09, F)

One solution to a tangible and explainable AI is through providing key metrics, such as but not exclusive to the parameters and their possible weighting in the algorithm, to the users.

> "I think it's more important to know what factors from the raw data that can have enough impact on the final results. We don't need to know the nitty-gritty of the model...For example, if the results highlight gender but not other factors, then we should reveal that despite all the other parameters. That to end-users is enough... what we should do is to allow users to reverse-engineer the results." (E13, F)

At other times, it is also important to make salient what is left off in the model for our users because it may be where the biases come from.

> "Compared to the 80% right things you do, the data you are able to explain, it matters more to explain the 20% that you are not able to process, things that may go wrong. The key here is to reduce the 20% unexplainable to, say, 10%. That's all it matters. " (E04, M)

In addition to what are included and what are left off, our engineer participants pointed out that providing comparisons and distinctions from competing models can also help end-users assess the validity of the AI model and its results.

> "...So this X product tells their users the most reliable way to record their maximal oxygen uptake is to go for an uphill run and use the data for the future prediction...But our product can track this data under the context of low speed and low heart rate yet still delivers precise predictions. We would tell our users why our product works differently than the X product. (E07, M)

Revealing conflict of interest. Our end-user participants especially pointed out the importance of revealing the contextual structure of organizational and commercial interests an AI may serve.

> "I need to know who is behind the AI system, like who owns the data, who created the program, etc. I would like to know who will benefit from this. There may be a conflict of interest, and I want to know it ." (U09, F)

In cases when the AI is commercially applied, our participants may connect such disclosure of interest to the level of control they may have over their data and how they may want to react to the AI suggestions.

> "I think most people want to take control of their lives. Say I want to buy a pair of earphones, I would rather search for the information on my own rather than be fed with all these recommendations that I do not like. If I had access to the rationale for the recommendation system, I may be able to tell it the specifications I am concerned with and tweak it so that the recommendations are more pertinent." (U22, M)

In sum, our participants pointed out three ways of embodying AI explainability, including 1) using an approachable narrative to display how the AI models work; 2) revealing key metrics that determine the results or what are left off from the models; providing comparisons across similar models may also work; and 3) being clear about the conflicts of interest.

4.3 Perceptions about Contexts in which Bias Matters

Our RQ2 addresses the biased scenarios with different levels of stake. When probing the negotiation between providing more data and reducing AI biased results, we discovered that the benchmark or concerns our participants have over AI applications and their potential biases are not held constant but varied across contexts and scenarios. Two principles emerged as salient for judgements about the appropriateness of AI applications and their biases, including personal boundary and the stakes of the use cases.

4.3.1 Personal boundary precedes over AI assistance. Our end-user participants reported that while AI provides much convenience and efficiency in terms of assistance with daily reminders or automatic data tracking, the priority of personal privacy boundaries precedes. They were not against data tracking and collection per se; in fact, they were aware that their data were collected when using smart applications and concurred that it could be convenient at times. However, they were not clear about the extent to which their data were collected. And they were reserved about how the data was used or the purposes of data collection. U08 discussed that she was open to employing AI applications at work, but when it came to personal space, the personal boundary was not negotiable. U17 talked about how data tracking was acceptable in functional assistance but not personal monitoring.

> "It's ok if it's for work … I am handling information about orders at work or what not with the email system. So it's ok [that the email system crawls the data and sends her notifications or reminders about work]. If Facebook does that to me, I'll get mad. If Facebook crawls my private messages with my friends and sends me date notifications, I'll be really mad. That's private. Technically, I think Facebook can still do it, but it won't actually remind me of it. If it does, I'll sue the company." (U08, F)
>
> "My personal information? It depends. If it [autonomous vehicle] wants to collect my whereabouts like what I do and where I go, it breaches my privacy.

> But if it wants to track how the vehicle functions and its related performance, such as battery life, I think it's ok." (U17, M)

On the other hand, our participants also mentioned that the higher the perceived personal relevance, the higher precision and the lower biases they demand from the AI model. For example, according to U04, AI-supported personalized recommendations may be perceived differently.

> "I can see the functionality of Netflix. It's simple. I can decide what I want to watch based on its recommendation. It saves my decision-making time. But if it's a dating app, I need to interact with the person to understand how precise the app recommendation is. it's hard to verify the recommendation precision beforehand. I don't think the recommendation is based on full information and I think the algorithm is just random." (U04, F)

Our participants were ambivalent about the personalized recommendations AI models strive to provide. They consider that a high matching of AI recommendations with personal relevance is a good indicator of AI performance. And they concur that for the sake of convenience and precision, they may decide how much they would like to give away their personal data. On the other hand, they recognize a downside that comes along, which is the restricted diversity and constrained exploration users may otherwise be exposed to were it not for AI suggestions. Our participants pointed out that AI recommendations, based on users' own behavioral data like digital footprints, can be limiting. While this application is good for marketing purposes for the business, it can be harmful to the customers.

> "It [AI algorithm] gives you a bunch of similar things based on your search results. I come to reflect on the need of information diversity and the reason why there is the phenomenon of echo chamber. I know that there are torrents of searchable information online. Facebook or YouTube, not knowing what to recommend, might as well provide things that we are already interested in… It is certainly useful commercially … creating customer attachment. But it harms cognition, for users." (U14, M)

U23 further mentioned that when AI fails to distinguish between public and private boundaries and gives seemingly personalized but actually hit-or-miss recommendations, users would develop negative perceptions about the applications.

> "It's not diverse enough. I need to purposely search for new information. Otherwise, everything is prearranged for me. And I use Gmail and Google at work. If I search for things related to work, when I go home and open my Facebook, all ads are work-related. The problem is, I don't want to see those things in my personal space. It's annoying. I don't choose to use these technologies at work. I just have to. But the results of this passive choice affect my daily life. I felt like my agency was taken away. I was controlled." (U23, M)

Our participants point out that the current practices of data tracking in many applications are not nuanced enough for users in terms of drawing lines between public and personal domains. No layers are provided for users to control how much data and in which contexts they are willing to share. One consequence is it can be hard for users to develop trust in AI.

> "So it's common that the privacy setting of some applications would ask me if I allow them to access my camera and audio. What really pisses me off is that if I don't grant its access, I can't use the application. It should let me choose and evaluate. I should still be able to use the app even if I don't allow camera or audio access. Then I may increase my trust in the application … Or, I may grant the application to track and analyze my data and provide recommendations regarding whether I want to eat cake and if so, which cake. But I don't want it to analyze my private conversation with my friends." (U09, F)

On the other hand, our engineer participants were not necessarily aware of the detailed concerns end-users may have for AI performance and potential biases. Most engineers talked about the user feedback they received, if ever, in a way that was reduced to a single dimension of model precision. Instead of engaging in user-driven modeling from the beginning, our engineer participants mentioned they experienced more of client-driven problem-solving in an ad-hoc manner.

> "They may complain that the results were lame…not precise at all. Then we would review the data and see why the model yielded results of low precision. And we would go find the reasons. If any value is below our expected performance, we would re-train the model…They [users] don't want to know why our model is under-performed. They only care about raising the precision level if it's low. It's like…make the model perform better, you don't have to give me all these excuses. " (E14, M)

While the line is clearly drawn for how much AI can be applied in end-user participants' personal space, they have more contextualized deliberations in other scenarios depending on the stake of the use cases with respect to AI appropriateness and biases, which we report in the next section.

4.3.2 Stake of the use cases: relevance of how and when biases matter. There are several dimensions emerged from our data associated with the stake of the application scenarios. First, our end-user participants pointed out that biases are intolerable in high-stake situations like medical applications that concern life, compared with gender bias in job application contexts or fintech applications. While in our interview these were both categorized as high-stake scenarios, our participants had their own considerations regarding what constitutes "high-stake."

> "it depends on which issue we're looking at to determine if it's [bias] a serious matter. If it's related to the medical field, it's very serious. It probably misdiagnoses symptoms… what if there's cancer but the system says there isn't? It's life and death…as for

> the HR gender bias case, yeah, it's unfair, but it just happened... the algorithm just made a misjudgement." (U10, M)

In general, our participants pointed out that when it comes to people's life, human rights, and civil affairs, they care more about the model precision, purposes, or application scenarios since the consequences of employing such systems can be immense. Biases in such situations also bring more serious consequences to users' eyes.

> "the face recognition system [for criminal identification] should tell us its precision rate, at which areas this system is applied, and its effects for me to understand and trust the system. As for face id (e.g., on the phone), it's ok. When the purpose of the system is about human rights, I think it should be more discrete." (U04, F)

In other words, it is not so much about AI techniques (e.g., image recognition) that our participants are concerned about but how and where the techniques are applied (e.g., car plate recognition vs. criminal identification). The major reason why these scenarios are more serious is the underlying inference-making process the AI model employs, which our participants feel at dark about. For example, U08 pointed out that gender bias mattered less was not because gender issue was not important. It is the process where AI systems give recommendations and users may make judgements or decisions based on these recommendations that concern our participants. Our participants were also concerned that complex and contextualized issues are not factored into the system for a fair judgement.

> "... there's no inference [in gender bias in online searching]. When you search for 'CEO' online, the system gives you the results ... but in the interview scenario, based on the data, the system suggests something new for you. During the process, it seems to involve thinking and judgement. In which case, the HR system needs to pay attention to this inference process, whereas in the searching scenario, I throw in a keyword and the system gives me the results. There's no additional judgement or inferences. " (U08, F)
>
> "It's not like computer vision that identifies cats or dogs and there is a correct answer. In the HR situation, you don't know which parameter is valued...it seems that the system favors males over females. But is it true? Or females are screened out because of other reasons? It can't be that the system specifically filters applicants based on gender. There should be multiple levels but we don't know and it's inexplainable." (E08, M)

Another reason is that civil-affairs usually have the quality of monopoly in which most, if not all, citizens are involved but no or few alternatives are provided. In these high-stake scenarios, our participants demand model precision and the margin of error is less tolerated.

> "It matters to everyone. If the results are problematic, it can't be used ... say if a system has higher precision with Caucasians but lower with Asians. It cannot be used. It has to work with everyone...A person's rights shouldn't be violated under any circumstance. Even the model has 100% precision in Asia but 80% for Caucasians and 99 percent for African Americans in the U.S., it's still not good enough. It's obvious racial discrimination." (U14, M)

Along the line with the previous point that high-stake scenarios concern a wide population, our participants pointed out that more stakeholders' perspectives should be involved. However, the current AI system may not integrate multi-party voices into the model processing or training. Also, our participants pointed out that meta-knowledge regarding the application scenario should be incorporated into the model to identify potential biases. But it can be difficult to define a comprehensive list of cases and scenarios.

> "I feel like that currently, the AI system just works on its own. You feed it data and it does the learning and gives us the results. 5 There should be dialogue at every stage...like in the AI judicial system...the system should talk to the suspect and the litigator as well as the judge or the prosecutor. It should collaborate with all stakeholders involved to have a comprehensive view. Then the system can come up with a report or suggestions for the judge. But that's not the end of the story. The judge should be able to interact with the system to verify and deliberate. But right now, it's like...'there you go, I come up with stuff from a black box and you take it.'" (U17, M)

In sum, our participants were concerned about AI model performance and whether there was potential bias involved especially when the application is of high stakes, such as one in the medical field or the judicial system. Since these applications may concern people's life, human rights, or civil rights, our participants point out that tolerance for errors or biases under the circumstances is low. In such cases, more stakeholders' involvement is important.

5 DISCUSSION

We discuss the implications of our results in response to our research questions.

5.1 Explainability Through AI Literacy

Regarding our RQ1 and RQ2 that investigated stakeholders' sensemaking and perceived challenges about AI and sociotechnical AI biases in contexts of different stakes, the findings from 4.1 found that both our end-user and engineer participants mentioned that due to the data collection process and application scenarios, AI biases are a reflection of social reality. When there was a biased result, our participants attributed the biases to algorithms, people who programmed the AI, the company that implemented the AI, or themselves. We found that our end-user participants could not pinpoint what went wrong but took AI-related factors and themselves to blame. Either way, this suggests that when people do not have enough knowledge to understand the AI mechanisms, they

may develop partial perceptions and raise suspicions or negative attributions when biases show up.

According to the attribution theory [19], people try to connect causes and effects even when there is none. Research has already shown that negative attribution towards AI can lead to distrust [16]. According to our results, our participants assigned both external/situational attributions (e.g., erroneous algorithms or flawed datasets) and internal/dispositional ones (e.g., themselves) when there was bias. The bias attributions may further harm users' trust in AI or even delay and obstruct user acceptance of smart and innovative services. It also limits users' educated and informed adaptation to biased results. Future XAI research may leverage the current study to further untangle which type of attribution can help enhance model explainability and user trust.

Meanwhile, the participants considered that there is no bias-free AI and with time, biases may evolve as society progresses or as the public concerns change. Therefore, AI should evolve too. Our data suggest that even the same type of AI bias may yield different consequences, depending on where and when the AI applications are launched and who the users are. The same bias, such as gender inequality, may be perceived with less negative implications in machine translation but more seriously in organization hiring. We also found that our participants who may be benefited from the biased results viewed biased AI less of a problem. Taken together, it suggests that developing awareness of bias and how to tackle biases derived from AI are separate matters.

As AI biases are a sociotechnical phenomenon, our results suggest that it takes both engineers' and end-users' efforts to cultivate AI-related literacy so that biases can be addressed from both technical and social lenses. In our data, not every engineer participant reported to beware that biases come from both technical and social levels. Also, some of the dataset-related biases reflect engineers' own limited experiences. For engineers, AI literacy consists of awareness and competency for identifying potential issues in the development stages like data collection or labeling as well as understanding what end-users may concern about so as to communicate model explainability to them.

For end-users, AI literacy should equip them with knowledge about how AI is developed and sensitivity to potential technical and social biases embedded in the process. For an interactive human-AI interaction, it is not feasible to passively wait for AI models to catch up with social trends or rely solely on engineers' awareness to solve this issue. Users should be encouraged to actively engage with AI for informed AI use, which is in line with another finding that users desire to retain their agency in deciding where and when AI can be in use. To this end, some interviewees even pointed out that it may be a blessing that AI fleshes out these biases that are otherwise invisible or hard to find evidence to corroborate in daily life. With biases made prominent by AI, it opens room for thorough discussion and improvement.

However, much of the above discussion is based on the assumption that AI literacy is a key factor in the deployment of AI in everyday life. While we acknowledge the significance of AI literacy, we also recognize that solely relying on it could be infeasible. For one, there is no consensus on what AI literacy should entail. And AI literacy may be built upon the protean forms of AI applications, the challenges of which are highly dependent on the goals, functions, and target user groups. For another, the actual, contextualized use of AI applications and their performance may not be covered by a rigid checklist of AI literacy items. For instance, our study fleshed out users' sense-making and expectations regarding bias-related situations. In other situations like performance-driven, ethically challenging, or morally ambiguous, the appropriate purview of AI literacy may be subject to change.

To complement prior conceptual works about guidelines for AI literacy and education [22, 46], our study provides more nuanced and contextualized extensions based on the empirical findings. While more high-level AI competencies or assessments about AI awareness are proposed based on systematic reviews of relevant works, such as Long and Magerko's [22] extensive list, it may be worthy of a more contextualized discussion about bias-associated AI literacy along the dimensions of the stakes of AI applications (low vs. high). According to our results, AI biases are not neutral but may vary depending on times, stakes of applications, and contexts like societies and cultures. The significance of stakes is not constant but varies depending on applications and contexts. Contextualized AI literacy in biased situations may help future works to form concrete ideas regarding how users perceive AI should work and how to respond when there is bias. It also allows AI literacy to be better assessed or incorporated into the curriculum. While we provide the first step to contextualize AI literacy, future works can further embody what AI literacy should construe by narrowing down dimensions like data types (e.g., text, image, or voice), applications (e.g., chatbots or autopilot cars) or contexts (e.g., in biases or about ethics).

5.2 Designing a Self-explanatory and Reflection-triggering System

Our results also uncovered the challenges our participants face when spotting, making sense of, and dealing with sociocultural AI biases. We propose correspondent designs to address these identified challenges.

While asking the general public to increase their AI literacy could be a way to address how to deal with AI biases and the demand for AI explainability, education is a long-term effort that may take years to reap the benefits. Additionally, with the burgeoning AI applications in real life, it is unrealistic and probably unnecessary to have users understand how *each* system works before using it. Just like a driver is able to drive a model that he/she has not driven before and probably still has no idea of how the engine system functions inside, AI system designs should have affordances that allow the general users to be able to understand their functions and reflect upon potential biases of the system even if it is the first time they use it.

To achieve this, our results from 4.2 suggest that AI systems that are used for inference-making, especially in high-stake use scenarios, should 1) disclose potential bias sources and explain how the results are derived in an approachable manner (e.g., through revealing where the data was collected, the categories included in the data, or how the data was handled); 2) highlight determining parameters for the inference making process; 3) be transparent about what factors are left off from the models; 4) provide comparisons of the performance across similar models; and 5) uncover potential

conflicts of interest (e.g., how much percentage the AI-mediated recommendations on the platforms is based on users' prior behaviors and how much percentage is based on product placement advertising). Clearly laying out these aspects may help users calibrate the results they get from the application and prompt them to reflect on the potential reasons for biases. In our interviews, when asked to reflect on the displayed biases in our proposed scenarios, even the end-users who possessed limited AI knowledge had suggested how the data, the engineers, and the social norms and values could have contributed to the biased results. Our results also showed that end-users desired to learn about such information, which helps them assess the AI system's performance and develop trust in the system. In its broadest sense, including the sources of biases and prompting the users to actively reflect on them are part of the AI literacy education process. Embedding such information in the system design can help evoke contextualized sense-making and reflection. This approach can also help users adjust when sociocultural norms and expectations change.

Alternatively, our results also indicate that the design of an XAI-enabled system could leverage the power of peer users and third-party organizations who are also using the same or a similar application to offer a more user-centric and audited explanation. For example, inviting impartial third parties such as professionals or NGOs for auditing model performance in high-stake AI applications like those in the medical, financial, or judicial domains may be a way to enhance trust [33]. The third-parties may act as representatives of the end-user groups to reflect community-specific or localized interests to the company and provide reports about how to make sense of the AI models to the end-users. Also, in addition to expert users such as doctors, bankers, or judges who rely on AI-mediated systems for decision-making, it is valuable for AI explainability to include the perspectives from the general end-users, who are at the mercy of such systems. They may have neither domain knowledge nor knowledge about how AI works, which puts them in a vulnerable position. Last, XAI can include what peer end-users care about and make the crowdsourced evaluations salient because this may help other users navigate information given by the AI systems or the third-party experts and make informed decisions.

The above indications and explanations can come in a variety of forms, not necessarily a rigid table where each piece of information, regardless of its significance, is outlined and elaborated in a standardized manner. The level of stake and the timings/contexts users value that our research highlights can be used as benchmarks to decide the granularity of explainability. While increasing an understanding about AI biases is desired, our results reveal a balance between giving sufficient information and avoiding information overload is needed, or in Long and Magerko's words [22], unveiling gradually. We do not argue that every AI application needs to come with a thick manual.

Our results showed that the role of a product manager that some of our engineer interviewees brought up already translated the knowledge from the designers and engineers into something sensible for the non-technician customers to understand and sent user feedback to the product designers to elicit more details and/or to tweak the system. The most important quality of such a product manager is user-centered interactivity and responsiveness. While it is not feasible or scalable to have this role present for every system and in every use context when bias comes up, a mediated role in an explainable artificial intelligence system can be a service chatbot or an animated agent, such as Microsoft Word's Clippy, activated when users need it. Such design also harks back to Long and Magerko's design consideration [22] of creating an AI learning experience that supports social interaction and collaboration. Such an interactive assistant could retrieve a manual/tutorial that is built upon engineers' knowledge of the system and leverage tips and comments from other peer users or previous user queries as well.

5.3 Paying Attention to Application Contexts

According to the results from 4.3, where AI is applied and the level of stake embedded in the application scenarios influence their perceived needs of a transparent and explainable model as well as their tolerance of biases. Our participants placed the highest reservation towards AI applications that may encroach on their personal space. They were also sensitive about the applications that may muddle the boundaries of public and private contexts, especially if the applications collect data from users' practices in public space (e.g., workplaces) and push the results in the forms of recommendations or reminders into their private space (e.g., interaction with friends).

The recent breakthroughs in audio and image data processing in AI training and modeling suggest that users' verbal and non-verbal behaviors, such as conversations or movements, in both public and private contexts, can be captured by sensing or audio-visual technologies [6]. Then the multi-modal data can be processed and analyzed for a wide range of AI-mediated applications. The scale can be leveled up when more and more products adopt such techniques. Users may feel a lack of control over their own data and how their behavioral data is going to be used in the context of omnipresent data collection. A lack of sensitivity to how and where AI can be applied may lead to users' rejection or mistrust towards the model.

In other words, privacy is a delicate issue that AI applications should pay attention to regarding data collection and model application in different contexts. Our study reveals that this is a kind of biased application scenario where users may give consent to or have the awareness that their behavioral data are being collected in work-related or other public contexts; however, such consent may not be extended to other contexts nor the consent is given to receiving services enabled by AI (e.g., advertisement recommendations) in a different situation. Future AI models should take users' privacy into consideration. While collecting as much data from diverse contexts for model training can enhance overall model performance, inadvertently pushing AI services can lead to backfire effects on users, such as what we found in our results: perceived violation of privacy, perceived restricted diversity by algorithmic recommendations, or distrust in the AI model.

Our results also suggest that in high-stake situations like medical, judicial, or other civil contexts, our participants hold higher standards for biases and needs for explainability. These contexts are important because they may concern people's lives or fundamental human rights based on AI's inferences. Also, these applications may hold the quality of monopoly where the majority of people are concerned but few or no alternative systems are provided. According

to our participants, such AI systems have not yet involved enough stakeholders' perspectives. And oftentimes, AI applications provide inferential suggestions in these situations (e.g., which organs may be infected or how long a culprit's sentence is going to be) but how the inference is derived is less clear. In such cases, our participants consider that data collection and model application should cover all users/stakeholders to rid of potential biases of gender, age, or race.

Taken together, our results suggest that for the designs of XAI, the following contexts are of importance to enhance model transparency and explainability as well as lower potential biases: 1) the level of relevance to users' personal boundaries, 2) the level of contextual awareness depending on users' switching between public and private spaces, and 3) the level of the stake associated with the application scenario.

When collecting user data and deploying models, the application should inform users about how, where, and what types of user data are collected and how the data is used to optimize model performance for explainability and transparency. In other words, in addition to keeping human in the loop for data processing and model development [18, 45], users should also be involved to decide the trade-off of model applications (e.g., AI service precision vs. privacy) [11]. Also, AI systems should consider including context awareness and sensing designs to avoid disrupting users' public and private spaces, especially when the system collects data from users' public practices and apply the model learning results in their private ones. AI systems should provide nuanced explainability in high-stake applications regarding potential biases related to sociocultural norms, ethics, and human rights. Last, it is also important to give users options about how much they want to be informed. For applications with lower personal relevance and stakes, a flexible granularity of explainability information allows users to avoid information overload.

6 LIMITATIONS AND FUTURE DIRECTIONS

It is our intention to include the general users, not necessarily tech-savvy ones, in our study because they are the majority who come to interact with AI in their day-to-day routines and receive the consequences. In order to understand general users' understanding about AI or potential underlying biases, our interview did not focus on a specific domain or application in case our end-user participants had no experience with it or our engineer participants were in a different domain. It is worth further examining a specific domain, such as health or law, or a specific application, such as chatbot or deepfake, to pinpoint users' perceptions about and solutions to domain-specific biases. Also, some marginalized user groups should be recognized and included in future studies to promote XAI design solutions for empowerment because they may be more vulnerable than others, such as AI novices [41] or those who work with constrained cognition [35]. Other sociocultural or demographic characteristics, such as age, gender, tech-savviness, or culture, may be considered to identify potentially marginalized user groups who are at the mercy of AI decisions. Then, despite our efforts, we only recruited one female engineer in our sample, which may limit perspectives regarding a specific bias, such as gender, from this under-represented engineer group. While we did not specifically focus on gender bias, future studies can look into how the gender of an under-represented group, such as female engineers in our sample, affects members' perceptions about gender bias. Also, our results pointed out that some biases are sociocultural. Given that our study was conducted in Taiwan (with three engineers working in the U.S.) with Mandarin native speakers, it is possible for future studies to draw on a cross-cultural perspective to examine how the same biases are perceived and dealt with in different sociocultural contexts. Other possible research directions include using quantitative research methods, such as a survey that broadly investigates a wide variety of users' perceptions and attitudes towards AI and related biases for clearer correlations among key variables. Or lab experiments can be conducted to identify the causal relationships of specific AI designs and users' trust. Also, our results did not cover how participants' own biases influenced their potentially biased interpretations about a non-biased AI, which can also be further explored in future studies. Last, despite our efforts to include as diverse roles in the AI development and deployment pipeline as possible in the study, there are many other key actors we were unable to recruit in this study, such as UI/UX designers, quality control engineers, or legal consultants. In addition, our engineer participants pointed out that their clients were oftentimes companies or organizations that had the needs of AI applications and they did not directly face end-users. How explainability may be lost in translation in the process of B to B and B to C should be further explored. It is important to deepen the knowledge about XAI from the perspective of connected and relevant stakeholders in future studies.

7 CONCLUSION

As more and more technologies and services are mediated by AI, it is important for general users to understand how AI works in order to leverage the results to make informed practices. This is especially true when AI algorithms produce biased results that cause socio-cultural consequences, like biases towards a certain gender, occupation, age, race, or culture. With much effort in computational and algorithmic ways to enhance AI explainability (XAI) for goals like fair, accountable, and transparent AI models and applications, it can still be difficult for general users to make sense of the explanations. To contribute to the existing XAI literature from a user-centered perspective, our work focuses on how general users make sense of AI and AI related sociocultural biases as well as what they want to know about AI in terms of how the results are derived when there are biases in contexts of high and low stakes. We conducted an in-depth interview study with 24 end-users and 15 engineers to address the issues. We identified users' sense-making processes of AI biases and their attributions of such biases. We reported people's desired levels, timings, and contexts of explainable AI and AI transparency. Based on our analysis, we found that explainability may need to come with users' AI literacy. However, more importantly, to achieve explainability and transparency, future AI systems should consider implementing features that disclose potential bias sources, highlight important parameters used in the AI model, recognize parameters that are omitted in the model, provide comparisons across similar models, and uncover potential conflicts of interests. Alternatively, the user interface should

consider functions that support users to reflect and learn about the model and the results when they need it. Last, our findings also showed that users desire to have control over when and where the AI-mediated systems provide services to them because unsolicited services can reduce user trust.

REFERENCES

[1] Samira Abbasgholizadeh Rahimi, Michelle Cwintal, Yuhui Huang, Pooria Ghadiri, Roland Grad, Dan Poenaru, Genevieve Gore, Hervé Tchala Vignon Zomahoun, France Légaré, and Pierre Pluye. 2022. Application of artificial intelligence in shared decision making: scoping review. *JMIR Medical Informatics* 10, 8 (2022), e36199.

[2] Ashraf Abdul, Jo Vermeulen, Danding Wang, Brian Y Lim, and Mohan Kankanhalli. [n. d.]. Trends and trajectories for explainable, accountable and intelligible systems: An hci research agenda. In *Proceedings of the 2018 CHI conference on human factors in computing systems*. 1–18.

[3] Amina Adadi and Mohammed Berrada. 2018. Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access* 6 (2018), 52138–52160.

[4] Chelsea Barabas, Madars Virza, Karthik Dinakar, Joichi Ito, and Jonathan Zittrain. 2018. Interventions over predictions: Reframing the ethical debate for actuarial risk assessment. In *Conference on Fairness, Accountability and Transparency*. PMLR, 62–76.

[5] Andrea Brennen. 2020. What Do People Really Want When They Say They Want" Explainable AI?" We Asked 60 Stakeholders.. In *Extended Abstracts of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–7.

[6] Francesco Camastra and Alessandro Vinciarelli. 2015. *Machine learning for audio, image and video analysis: theory and applications*. Springer.

[7] Maartje MA De Graaf and Bertram F Malle. 2017. How people explain action (and autonomous intelligent systems should too). In *2017 AAAI Fall Symposium Series*.

[8] Ashley Deeks. 2019. The Judicial Demand for Explainable Artificial Intelligence. *Columbia Law Review* 119, 7 (2019), 1829–1850.

[9] Jonathan Dodge, Q Vera Liao, Yunfeng Zhang, Rachel KE Bellamy, and Casey Dugan. 2019. Explaining models: an empirical study of how explanations impact fairness judgment. In *Proceedings of the 24th international conference on intelligent user interfaces*. 275–285.

[10] John Fox, David Glasspool, Dan Grecu, Sanjay Modgil, Matthew South, and Vivek Patkar. 2007. Argumentation-based inference and decision making–A medical perspective. *IEEE intelligent systems* 22, 6 (2007), 34–41.

[11] Batya Friedman. 1996. Value-sensitive design. *interactions* 3, 6 (1996), 16–23.

[12] Timnit Gebru, Jamie Morgenstern, Briana Vecchione, Jennifer Wortman Vaughan, Hanna Wallach, Hal Daumé III, and Kate Crawford. 2018. Datasheets for datasets. *arXiv preprint arXiv:1803.09010* (2018).

[13] Katy Ilonka Gero, Zahra Ashktorab, Casey Dugan, Qian Pan, James Johnson, Werner Geyer, Maria Ruiz, Sarah Miller, David R Millen, Murray Campbell, et al. 2020. Mental Models of AI Agents in a Cooperative Game Setting. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–12.

[14] Riccardo Guidotti, Anna Monreale, Dino Pedreschi, and Fosca Giannotti. 2021. Principles of explainable artificial intelligence. In *Explainable AI within the digital transformation and cyber physical systems*. Springer, 9–31.

[15] David Gunning. 2017. Explainable artificial intelligence (xai). *Defense Advanced Research Projects Agency (DARPA), nd Web* 2, 2 (2017).

[16] Jess Hohenstein and Malte Jung. 2020. AI as a moral crumple zone: The effects of AI-mediated communication on attribution and trust. *Computers in Human Behavior* 106 (2020), 106190.

[17] Anna Jobin, Marcello Ienca, and Effy Vayena. 2019. The global landscape of AI ethics guidelines. *Nature Machine Intelligence* 1, 9 (2019), 389–399.

[18] Fabrice Jotterand and Clara Bosco. 2020. Keeping the "human in the loop" in the age of artificial intelligence. *Science and Engineering Ethics* 26, 5 (2020), 2455–2460.

[19] Harold H Kelley. 1967. Attribution theory in social psychology.. In *Nebraska symposium on motivation*. University of Nebraska Press.

[20] Jean-Baptiste Lamy, Boomadevi Sekar, Gilles Guezennec, Jacques Bouaud, and Brigitte Séroussi. 2019. Explainable artificial intelligence for breast cancer: A visual case-based reasoning approach. *Artificial Intelligence in Medicine* 94 (2019), 42–53.

[21] Q Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: informing design practices for explainable AI user experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–15.

[22] Duri Long and Brian Magerko. 2020. What is AI literacy? Competencies and design considerations. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–16.

[23] Michael A Madaio, Luke Stark, Jennifer Wortman Vaughan, and Hanna Wallach. 2020. Co-designing checklists to understand organizational challenges and opportunities around fairness in ai. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*. 1–14.

[24] Prashan Madumal, Tim Miller, Liz Sonenberg, and Frank Vetere. 2019. A grounded interaction protocol for explainable artificial intelligence. *arXiv preprint arXiv:1903.02409* (2019).

[25] Sherin Mary Mathews. 2019. Explainable artificial intelligence applications in NLP, biomedical, and malware classification: a literature review. In *Intelligent Computing-Proceedings of the Computing Conference*. Springer, 1269–1292.

[26] Christian Meske, Enrico Bunde, Johannes Schneider, and Martin Gersch. 2022. Explainable artificial intelligence: objectives, stakeholders, and future research opportunities. *Information Systems Management* 39, 1 (2022), 53–63.

[27] Tim Miller. 2019. " But why?" Understanding explainable artificial intelligence. *XRDS: Crossroads, The ACM Magazine for Students* 25, 3 (2019), 20–25.

[28] Tim Miller. 2019. Explanation in artificial intelligence: Insights from the social sciences. *Artificial intelligence* 267 (2019), 1–38.

[29] Margaret Mitchell, Simone Wu, Andrew Zaldivar, Parker Barnes, Lucy Vasserman, Ben Hutchinson, Elena Spitzer, Inioluwa Deborah Raji, and Timnit Gebru. 2019. Model cards for model reporting. In *Proceedings of the conference on fairness, accountability, and transparency*. 220–229.

[30] Brent Mittelstadt. 2019. Principles alone cannot guarantee ethical AI. *Nature Machine Intelligence* 1, 11 (2019), 501–507.

[31] Joseph Happy Okoh. 2022. 10 Impacts of Artificial Intelligence on Our Everyday Life.

[32] Franziska Pilling, Haider Ali Akmal, Joseph Lindley, Adrian Gradinar, and Paul Coulton. 2022. Making AI Infused Products and Services more Legible. *Leonardo* (Oct 2022), 1–11. https://doi.org/gq9nf8

[33] Inioluwa Deborah Raji, Peggy Xu, Colleen Honigsberg, and Daniel Ho. 2022. Outsider oversight: Designing a third party audit ecosystem for ai governance. In *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*. 557–571.

[34] Marco Tulio Ribeiro, Sameer Singh, and Carlos Guestrin. 2016. " Why should i trust you?" Explaining the predictions of any classifier. In *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining*. 1135–1144.

[35] Justus Robertson, Athanasios Vasileios Kokkinakis, Jonathan Hook, Ben Kirman, Florian Block, Marian F Ursu, Sagarika Patra, Simon Demediuk, Anders Drachen, and Oluseyi Olarewaju. 2021. Wait, but why?: assessing behavior explanation strategies for real-time strategy games. In *26th International Conference on Intelligent User Interfaces*. 32–42.

[36] Stuart Russell and Peter Norvig. 2002. Artificial intelligence: a modern approach. (2002).

[37] Wojciech Samek, Thomas Wiegand, and Klaus-Robert Müller. 2017. Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. *arXiv preprint arXiv:1708.08296* (2017).

[38] Andrew D Selbst, Danah Boyd, Sorelle A Friedler, Suresh Venkatasubramanian, and Janet Vertesi. 2019. Fairness and abstraction in sociotechnical systems. In *Proceedings of the conference on fairness, accountability, and transparency*. 59–68.

[39] Kacper Sokol. 2019. Fairness, accountability and transparency in artificial intelligence: A case study of logical predictive models. In *Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society*. 541–542.

[40] Harini Suresh and John V Guttag. 2019. A framework for understanding unintended consequences of machine learning. *arXiv preprint arXiv:1901.10002* (2019).

[41] Maxwell Szymanski, Martijn Millecamp, and Katrien Verbert. 2021. Visual, textual or hybrid: the effect of user expertise on different explanations. In *26th International Conference on Intelligent User Interfaces*. 109–119.

[42] Prasanna Tambe, Peter Cappelli, and Valery Yakubovich. 2019. Artificial intelligence in human resources management: Challenges and a path forward. *California Management Review* 61, 4 (2019), 15–42.

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

[44] Danding Wang, Qian Yang, Ashraf Abdul, and Brian Y Lim. 2019. Designing theory-driven user-centric explainable AI. In *Proceedings of the 2019 CHI conference on human factors in computing systems*. 1–15.

[45] Fabio Massimo Zanzotto. 2019. Human-in-the-loop Artificial Intelligence. *Journal of Artificial Intelligence Research* 64 (2019), 243–252.

[46] Brahim Zarouali, Tom Dobber, Guy De Pauw, and Claes de Vreese. 2020. Using a Personality-Profiling Algorithm to Investigate Political Microtargeting: Assessing the Persuasion Effects of Personality-Tailored Ads on Social Media. *Communication Research* (2020), 009365022096196.

[47] Tal Zarsky. 2016. The trouble with algorithmic decisions: An analytic road map to examine efficiency and fairness in automated and opaque decision making. *Science, Technology, & Human Values* 41, 1 (2016), 118–132.

[48] Carlos Zednik. 2019. Solving the black box problem: a normative framework for explainable artificial intelligence. *Philosophy & Technology* (2019), 1–24.