Exploring Trust in Explainable AI Systems: The Role of Gender Dynamics and Alibi Choice

Marc de Zoeten
Johannes Gutenberg University Mainz
Mainz, Germany *mdezoet@uni-mainz.de*

Claus-Peter H. Ernst
Hochschule RheinMain
Wiesbaden, Germany *Claus-Peter.Ernst@hs-rm.de*

Abstract

This study investigates how gender dynamics and alibi choice affect trust in explainable AI systems. In an experiment with 188 participants engaged in a travel group matching task, we manipulated the displayed gender of evaluated individuals and provided some participants with a feigned (alibi) choice over the scope of the system's explanation. The experiment did not reveal significant differences in trust or its antecedents based on gender dynamics suggesting users' over-reliance on AI decisions. Additionally, the provision of the alibi choice slightly worsened perceptions of the system, indicating that superficial control options may trigger increased scrutiny rather than enhancing trust.

Keywords: Explainable AI, Gender, Trust, Risk, XAI.

1. Introduction

In today's world, individuals are constantly affected by the output of AI-based systems that rate them or make decisions about them [4]. While some individual decisions may have a negligible impact, such as recommendations for posts on social media, other decisions or ratings may have life-altering implications, such as a decision about a being hired [39]. The systems making such decisions typically do not provide insight into their decision-making logic, making them black-box systems [7]. This lack of insight often hinders adoption and usage of such systems [28], [45]. To address this shortcoming, explainable AI has been developed to provide insights into the decision-making logic of the system [4]. It is generally accepted that trust positively influences technology acceptance, usage, and adoption [33], [25], [54]. Hence, explainable AI seeks to enhance trust into AI systems, usually by improving its known antecedents, e.g., explainability, transparency, fairness, performance, and (perceived usage) risk [20]. However, multiple aspects in explainable AI's role in trust formation are as of yet underexplored. One key aspect that warrants further investigation is the role of gender bias.

In the field of AI, many systems are based on machine learning (ML). In the context of decision support, ML systems are usually trained using existing datasets that contain historical real-world decisions and/or pre-existing data [15], [55]. This data may, however, be substantially biased [1], because decision makers in the past have been shown to exhibit bias when making decisions. For example, judges may exhibit bias against certain age groups [47], and recruiters may exhibit bias against certain ethnicities [27]. This bias is then potentially replicated by the ML-based systems as it learns from this biased data [15]. While such systems can exhibit bias against various groups, research has shown that women, in particular, often face significant disadvantages by these decisions. For example, they are less likely to be shown adequate ads for jobs in social network sites [32], and after applying, they are less likely to be considered for a job [14], [57], [63]. The potentially increased vulnerability and risk of women, who are being evaluated automatically by ML-based systems, should cause individuals to be more critical of the corresponding decisions

when women are affected. In fact, we believe that when women are evaluated by an ML-based system, people generally perceive them as more vulnerable and at risk, leading to more negative perceptions of the system. Moreover, we believe women will have additional concerns about such systems evaluating other women, as they have an improved understanding on being discriminated against as a woman [11].

Another underexplored aspect of explainable AI's role in trust perception is the possibility to let people feel in control again [26]: With the increasing automation of decisions, affected individuals (both decision makers and evaluated individuals) may feel progressively left out of non-transparent decision-making processes [9], [38] and they lack any option to challenge decisions they perceive as incorrect. This leads to dissatisfaction [12]. Providing individuals with a means to take action – such as not only offering explanations for the system's decisions but also granting them even a feigned (alibi) choice regarding the scope of these explanations—may help enhance their overall perception of the system.

To explore the impacts of gender dynamics and alibi choice on perceptions of AI-based systems, we performed an experiment in the context of matching individuals that seek to form travel groups in which our first treatment was a variation of the displayed gender of the individuals evaluated by a mock-up system. The gender of the evaluator is not manipulated by us but instead it is the participant's gender. For our second treatment we vary the participants' choice, by offering some participants an alibi choice over the scope of the system's explanation (but no influence on the system's actual decision). After each treatment, we measure participants' trust perceptions (and known antecedents, e.g., explainability, transparency, fairness, performance, and two measures of risk) about the mock-up AI-based system and compare them.

This paper is organized as follows: first, we discuss the related work for our experiment and present our research model. Subsequently, we describe our methodology. Next, we present and briefly discuss our results. Lastly, we draw a short conclusion, point out limitations of our study and show avenues for future research.

2. Trust and Trust Antecedents in Explainable AI Systems

Within the field of Information Systems (IS), trust can be defined as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" [40]. Trust has been investigated across various systems, including AI-based systems and explainable AI systems and has been found to be crucial for a variety of stakeholders' usage and adoption of such systems [4]. Multiple factors influence trust in IS in general, and additional ones affect trust within the field of explainable AI:

While performance (1) and risk (2) are widely recognized as key trust antecedents [64], research in explainable AI has additionally identified explainability (3), transparency (4), and fairness (5) as additional factors influencing trust [33]. Moreover, trust in explainable AI systems is also shaped by whether individuals experience an expectation violation (6) either while using the system or being evaluated by it [20], [35].

Performance (1) has been shown to generally be an important trust antecedent [25]. It is regularly split into three dimensions [59]: (a) reliability (if a system operates error-free [29], [42]), (b) competency (how effective a system is [48]), and (c) predictability (individuals' perceived ability to anticipate a system's output [40]).

Risk (2) can be defined as the probabilities of adverse outcomes combined with the consequences of those outcomes [62]. It can further be split into the general risk (a), and personal risk (b) [61]. While the general risk is the risk that refers to (all) others experiencing adverse outcomes, the personal risk is the risk of an adverse outcome for oneself [53]. For any individual, those two types of risk may differ, based on an individual's characteristics, with some characteristics leadings to an increased vulnerability for an adverse outcome and consequently to a higher personal risk [8], [53]. The perception of risk (both general and personal) is not necessarily accurate and heavily influenced by an individual's believes, (recent) personal experiences, and information about what might put an individual at an increased risk, based on known adverse outcomes

[17], [30], [60].

For the field of explainable AI particularly, explainability (3) (the understanding of the model's inner workings as well as how the models arrive at their outputs) is crucial [13]. Additionally of importance are transparency (4) (the level of insight provided by the system into its internal processes) [49], and fairness (5) (the degree to which the system is believed to operate without discrimination) [3]. These factors and their influences on trust have been thoroughly explored in the literature [22, 23], [33].

Expectation violations (6) (discrepancies between what an individual predicts will or should be a system's output and the system's actual output) can have detrimental effects for trust and its antecedents [34], [50]. However, expectation violations also act as a motivation to use cognitive resources to properly investigate the reasons for the perceived discrepancy [5]. In fact, the provision of explanations might have no or very limited impact at all, if an individual did not experience an expectation violation that sufficiently motivates them to investigate provided explanations [19].

3. Research Model

3.1. Gender Dynamics in Trust Perceptions of Explainable AI Systems

In the field of automated decision-support systems that make use of ML, women are more likely than men to experience adverse effects from biased decision making, as these systems reproduce bias that is often found in the datasets used to train such systems[14], [32], [57]. Due to the missing insights into theses systems' decision-making processes [7] both Explainability and Transparency can be expected to be lower. As women are more affected than men, the Fairness suffers as well. Moreover, the increased likelihood of biased decisions decreases the systems' Performance and, at the same time, increases Risk. With the discussed antecedents affected, Trust in AI-based systems is likely to be lower. We propose:

H1: When women are evaluated by an AI-based system, the perceived Trust (H1a), Explainability (H1b), Transparency (H1c), Fairness (H1d), Competency (H1e), Predictability (H1f), and Reliability (H1g) of the system are lower, while the perceived General Risk (H1h), and Personal Risk (H1i) are higher compared to when men are evaluated by the same system.

Vulnerable individuals may have in some form already experienced biased decision making in real life. For example, women are (more) aware about potential discrimination in decision-making against women and are more sympathetic to women who potentially experience such discrimination [11]. Such experiences might make them more aware of potential biases in ML-based systems and, based on this increased awareness, cause them to assess that women have higher requirements towards such systems. We propose that the overall effects outlined in H1 result from women's heightened concern about the decision-making process, whereas we do not expect to find any effects when men evaluate the system. We propose:

H2: When women are evaluated by an AI-based system and women evaluate the system, the perceived Trust (H2a), Explainability (H2b), Transparency (H2c), Fairness (H2d), Competency (H2e), Predictability (H2f), and Reliability (H2g) of the system are lower, while the perceived General Risk (H2h), and Personal Risk (H2i) are higher compared to when men are evaluated by the same system.

3.2. Alibi Choice in Trust Perceptions of Explainable AI Systems

Humans increasingly work with systems that are (partially) automated [36]. While for some of those systems, users' perceived control might still be high, especially for systems where user input is perceived to be crucial for the output or decision, some autonomous agents make and communicate decisions without any human input or oversight [9], which

can lead to feelings of helplessness and frustration [12] [38]. For example, in recommender systems, individuals can at least partially influence and understand the recommendations made for them since knowing that these are a result of their behavior or characteristics [43]. In contrast, in modern (high-frequency) algorithmic trading of shares and bonds, AI-based systems can assess if a trade would be viable and completely autonomously execute that trade without any human oversights [10]. Consequently, traders may feel like they have very limited to no control over these trades.

Choice (or the illusion of being given/having a choice) has proven beneficial in a variety of contexts, even with limited actual impact [37] [51]. However, the effect (of the illusion of) choice may heavily depend on the perceived impact of that choice [56]. More specifically, if the choice is meaningless or has no effect, the effect of that choice may be inconsequential [37]. This impact is based on an individual's perception of the impact of the choice and not its objective impact [58].

We believe that especially in the case of an expectation violation (triggering a need for control, or information, or both) and with no direct option to oppose a system's decision, individuals have a strong requirement to be provided with a mechanism that lets them act and consequently feel less frustrated and helpless [cf. 6]. Even just the illusion of control can substantially increase perceptions of a system [46]. Furthermore, we believe that the act of choosing between substantially different alternatives is relevant, therefore even an alibi choice can be expected to have an impact comparable to that of a real choice for an individual. We propose:

H3: For individuals that are offered an alibi choice regarding an AI-based system's output, the perceived Trust (H3a), Explainability (H3b), Transparency (H3c), Fairness (H3d), Competency (H3e), Predictability (H3f), and Reliability (H3g) of the system are higher, while the perceived General Risk (H3h), and Personal Risk (H3i) are lower compared to when individuals are offered no alibi choice regarding the system's output.

4. Methodology

To evaluate our research model, we carried out an experiment at two universities in January of 2025. The experiment itself was in German and participants were required to be able to read and write German. In addition, participants confirmed that they were of legal age in Germany (18 years or older). For their voluntary participation, participants were compensated with 5€ for on average 14 minutes and 38 seconds spent in the experiment. This exceeds the German minimum wage of 12.82€ per hour.

Overall, 212 people took part in our study. Of these, only 188 passed our attention check. All analyses were performed on those remaining 188 participants. Given the university setting, the average age was young at 21.92 (SD: 2.731) years old. Participants had the option of not reporting their age, which 6 participants did. 50% of our participants were female and 48.9% were male. Two individuals chose not to disclose their gender. Most of our participants were students (92.6%).

Within the experiment, there were four experimental conditions, with a 2x2 design, which are displayed in Table 1. Participants were randomly assigned to their experimental condition before answering any questions.

Table 1. Experimental Condition (Randomly Assigned).

	Displayed Gender Female	**Displayed Gender Male**	**Sum**
Choice for Scope of Explanation	41	46	87
No Choice for Scope of Explanation	48	53	101
Sum	89	99	188

At the beginning of the experiment, participants saw profiles of two people that were explained to be considering traveling to Southeast Asia but had no one to go with and did

not want to travel alone. Both profiles included various characteristics presented as tabular data, incorporating factors identified in the literature as influencing harmonious travel [e.g., 21] [e.g., 31] along with additional attributes: gender, age, preferred travel group size, preferred length of the journey, budget after cost for transport and housing, diet (e.g., omnivore or vegetarian), smoker, attitudes regarding food, parties, exploration of nature and cultural activities during the journey, how they intend to relax during the journey, get up time, and a field for additional comments.

We consider our setup to be a medium-risk setup for affected individuals. A wrong choice for a travel companion is usually not life-threatening [41] or altering for the individual making the choice (although given the distance and concerns about local security and medical infrastructure that might be on the minds of at least some participants), but for most individuals (especially in the group that took part in our experiment), the risk is not small either, as the financial strain caused by the trip would be substantial and not repeatable for quite a while. Our participants are technically not affected directly, which we account for by taking different approaches to measure risk [cf. 44](see Table 2), yet it still might influence their reactions.

Table 2. Constructs and Measurements.

Construct	Items	Cronbach's α	Based on
Trust	I consider recommendations for the formation of travel groups by Eastern Horizon AI to be trustworthy. I believe that recommendations for the formation of travel groups by Eastern Horizon AI are reliable. I trust the recommendations made by Eastern Horizon AI.	.928 (M1) .926 (M2)	[52]
Explainability	I perceive the recommendation of Eastern Horizon AI as easy to understand. I think that the recommendations provided by Eastern Horizon AI are explainable. I can understand the internal mechanisms of Eastern Horizon AI.	.766 (M1) .821 (M2)	
Transparency	I think that all the criteria used by Eastern Horizon AI for the recommendation are communicated clearly. Every Eastern Horizon AI recommendation can be explained to the people affected by the recommendation. I understand how Eastern Horizon AI generates recommendations for travel group formation from the information provided by the individuals.	.727 (M1) .848 (M2)	
Fairness	The AI-based system Eastern Horizon AI does not favor anyone and does not discriminate. The AI-based system Eastern Horizon AI is a fair system. The AI-based system Eastern Horizon AI is impartial and has no biases.	.851 (M1) .899 (M2)	
Competency	Eastern Horizon AI performs its task (recommendation if travel group should be formed) effectively.	-/-	[59]
Predictability	I can predict the recommendation of Eastern Horizon AI.	-/-	
Reliability	Eastern Horizon AI is error-free.	-/-	
Risk (general)	(Using) Eastern Horizon AI is risky.	-/-	[62]
Risk (personal)	When using Eastern Horizon AI, I have a personal risk.	-/-	[61]

Two out of our four experimental groups received the profiles of two females, while the other two experimental groups received the profiles of two males (treatment 1). The gender of the evaluated individuals was clearly communicated at the top of their respective profiles. The two experimental conditions differed by just the gender, the rest of the features and descriptions remained identical.

After reviewing the profiles, participants were asked to either recommend for the two individuals presented in the profiles to form a travel group or to not form a travel group – 141 of our participants recommended to form a travel group, 47 recommended not to do

so.

Subsequently, participants were briefly introduced to the mock-up system "Eastern Horizon AI" by explaining that it was developed to match individuals for journeys to South-East Asia. Regardless of the participants' recommendations, Eastern Horizon AI always provided a recommendation that contradicts the participants' recommendation to cause an expectation violation thus motivating participants to critically assess the system and the explanation [19]. Such limited manipulations can be considered normal when performing experiments [16]. Our participants then answered several questions regarding their perceptions about the system (measurements M1). The constructs, items and sources for our measurements are displayed in Table 2. All measurements were made using a 7-point Likert-type scale.

Next, two of our experimental groups were asked about the scope of the explanation that they would like to be provided with four options: (1) a small, basic explanation, (2) a medium explanation with higher details, (3) a large explanation with extended details on the decision, or (4) a random scope for the explanation (treatment 2). Table 3 indicates the decision scope chosen by our participants, only for the two experimental groups that had the option to choose.

Table 3. Choice for Explanation Scope (by Participants).

Explanation scope	Small	Medium	Large	Random
Chosen by	8	32	44	3

However, the scope of the explanation shown was always identical and did not depend on participants' choices: First, it was explained that Eastern Horizon AI creates a "Traveler Score" to decide whether to recommend the two individuals travel together. Then, participants saw an enhanced version of the original table with the individuals' characteristics. This updated table included the impact value of each characteristic on the "Traveler Score" and a short verbal explanation next to each impact value.

Afterwards, participants were once again asked questions about their perception of the system (measurements M2, cf. Table 4). Finally, we collected demographic information. Figure 1 provides an overview of our experimental setting.

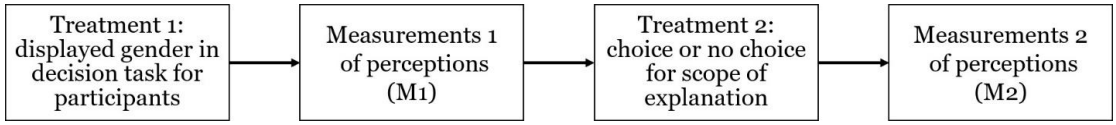

Figure 1. Treatments and Measurements of Participants' Perceptions

5. Results & Discussion

We made use of SPSS 29.0.2.0 for all our analyses. Table 4 displays descriptive statistics for our variables, for both points in the analysis (M1 and M2).

First, we investigated the gender dynamics in trust perceptions of explainable AI systems. For hypotheses H1a-i (which, overall, propose that the system is perceived as less favorable if a woman is evaluated compared to when men are evaluated), we find that, although our predictions regarding the direction of differences hold true for every variable except Predictability, none of the differences are significant, as displayed in Table 5.

We then proceeded to investigate H2a-i, (which propose that when women are evaluated by an AI-based system and women evaluate the system they are more critical towards the system compared to when men are evaluated). None of these hypotheses could be confirmed, as there were no statistically significant differences for the group of the female evaluators. Still, while female participants were on average more critical of the system for almost all variables (overall and split by displayed gender) just Reliability was significantly different ($p = .002$; since we did not formulate a corresponding hypothesis a priori, it is not displayed in any results table).

Table 4. Descriptive Statistics.

	M1: N = 188				M2: N = 188			
	Mean	SD	Min	Max	Mean	SD	Min	Max
Trust	2.95	1.24	1	6	3.67	1.34	1	7
Explainability	3.86	1.39	1	7	4.96	1.29	1	7
Transparency	3.64	1.37	1	7	4.92	1.31	1	7
Fairness	4.52	1.22	1	7	4.93	1.39	1	7
Competency	4.21	1.50	1	7	4.70	1.49	1	7
Predictability	3.17	1.53	1	7	3.98	1.47	1	7
Reliability	2.07	1.25	1	6	2.23	1.28	1	6
Risk (general)	3.83	1.59	1	7	3.65	1.71	1	7
Risk (personal)	3.88	1.74	1	7	3.53	1.80	1	7

Table 5. The Impact of Displayed Gender and Participant's Gender (M1). Unpaired T-Tests. * = p<.05; ** = p<.01; *** = p<.001.

	Overall N = 188			Participant Female N = 94			Participant Male N = 92		
Evaluated Gender	Female	Male	Sig.	Female	Male	Sig.	Female	Male	Sig.
N	89	99		46	48		43	49	
Trust	2.78	3.01	.087	2.73	3.02	.236	2.85	3.24	.132
Explainability	3.75	3.95	.302	3.51	3.86	.225	4.00	4.04	.888
Transparency	3.63	3.66	.892	3.60	3.53	.818	3.66	3.74	.772
Fairness	4.50	4.53	.835	4.58	4.36	.386	4.41	4.73	.221
Competency	4.03	4.36	.232	3.93	4.31	.232	4.14	4.41	.390
Predictability	3.22	3.12	.846	2.93	2.88	.846	3.53	3.35	.555
Reliability	1.99	2.15	.616	1.87	1.75	.616	2.12	2.59	.078
Risk (general)	3.92	3.67	.455	3.93	3.77	.619	3.91	3.63	.409
Risk (personal)	3.97	3.80	.510	4.15	3.97	.621	3.76	3.53	.525

The findings so far are relevant: Women are (until better mechanisms will have been developed to combat bias in existing datasets or new unbiased datasets have been generated) at an increased risk when potentially biased ML-based systems become even more widespread and impactful. The lack of awareness about the potential bias inherent in such systems might potentially cause overconfidence in such systems, resulting in worse outcomes both for the organization employing the system and affected individuals. making them potentially too trusting in applications that might discriminate based on certain characteristics or against certain groups. In fact, individuals may see AI-based system as an authority for a decision-task and consequently not sufficiently challenge its potential downsides [24]. This may lead to an overreliance on such systems [36]. This may be particularly concerning, as our participants are young students who are increasingly subjected to evaluations and decisions made by automated ML-based systems.

Finally, we proceeded to investigate the impact of alibi choice for the scope of the explanation. We investigate H3a-i, (which overall propose that individuals that have an alibi choice will have improved perceptions of the system). We find that overall, personal risk is perceived as significantly higher, if individuals can choose the scope of their explanation. Likewise, both personal and general risk is perceived to be higher by the male participants if they are given the choice. Consequently, we cannot confirm any of our hypotheses as shown in Table 6.

In fact, we proposed an opposite effect. However, the literature provides multiple possibles reasons for our results: (1) the choice for the scope of the explanations caused participants to imagine a specific explanation. These expectations about the explanation could then in turn be violated, which would cause perceptions of the system to worsen [18]. (2) With existing systems that our participants are likely highly familiar with, such as ChatGPT, the control exercised by humans (versus the AI-based system) is potentially much higher. Therefore, the choice itself was not sufficiently large to let them feel sufficiently in charge. (3) The provision of choice had a significant impact (mainly on risk)

but a consistent adverse (even if mostly insignificant) impact on all other variables but two (Transparency & Predictability), which leads us to believe that the choice has triggered participants to examine the system more closely (just like an expectation violation does) and consequently led them to identify potential problems [2]. This is in line with the choice and transparency paradoxes, respectively.

Overall, we believe that our results suggest that choice may play a role for these systems. However, the exact type of choice and where it is placed may be highly critical. Finally, our explanation of the system's decision was fairly effective at improving participants' perceptions of the system. More specifically, improvements for all variables except Personal Risk and Reliability were highly significant ($p < 0.001$, descriptive statistics are available in Table 4; no prior hypothesis had been provided).

Table 6. The Impact of Alibi-Choice (M2).
Unpaired T-Tests. * = p<.05; ** = p<.01; *** = p<.001

	Overall N = 188			Participant Female N = 94			Participant Male N = 92		
Choice	No	Yes	Sig.	No	Yes	Sig.	No	Yes	Sig.
N	101	87		51	43		50	42	
Trust	3.84	3.47	.054	3.72	3.39	.196	3.97	3.63	.246
Explainability	5.03	4.88	.424	5.07	4.72	.240	5.00	5.07	.773
Transparency	4.90	4.94	.831	4.95	4.91	.867	4.84	4.99	.575
Fairness	4.98	4.88	.626	4.90	5.04	.442	5.05	4.75	.317
Competency	4.92	4.45	.030*	5.18	4.65	.072	4.66	4.29	.245
Predictability	3.97	4.00	.890	3.65	3.88	.447	4.30	4.12	.539
Reliability	2.31	2.15	.403	2.14	1.93	.400	2.48	2.43	.429
Risk (general)	3.46	3.89	.088	3.94	3.60	.350	2.96	4.05	.001**
Risk (personal)	3.26	3.85	.024*	3.55	3.86	.425	2.96	3.74	.028*

6. Conclusion

In this article, we explored the impacts of gender dynamics and alibi choice on individuals' perceptions of a system (Trust, Explainability, Transparency, Fairness, Performance, Risk) in the context of explainable AI. To this end, we conducted an experiment involving the matching of individuals that seek to form travel groups in which we varied the gender of individuals evaluated by a mock-up AI-based system and measured our participants' perceptions after they were subjected to the system's decision. Next, each participant received an explanation for the decision, with some participants being given an alibi choice regarding the scope of the provided explanation upfront. Finally, we measured all participants' perceptions once again.

We could not identify any relevant differences caused by gender dynamics. This takes into account both the genders of individuals being evaluated by an AI-based system as well as the gender of the persons observing that evaluation. This may be particularly concerning since our participants' perceptions do not reflect the increased risk that women find themselves at when being evaluated by AI-based systems. They seem to be unaware of the potential risks associated with these systems—which stem from biased training data—and they seem to over-rely on AI decisions, potentially treating AI as an authority [24, 36].

For the impact of alibi choice, we found that participants' perceptions are not improved by the provision of an alibi choice. This finding is important, as organizations cannot easily manipulate perceptions of their systems positively by merely offering token concessions that have no real value to users. In fact, our results point to worsening overall perceptions of the system, if an alibi choice is offered. This goes against our expectations. We discussed possible explanations from the literature, such as stronger expectations towards the explanation or increased mental resources directed at how the system works [18].

Our study faces some limitations. First, no mixed-gender experimental condition was provided and all provided conditions were based on only two individuals (as opposed to larger groups). Additionally, we did not gather information about our participants' experiences and knowledge with gender-specific discrimination by AI. We did also not include potential influences of other demographic factors such as age. Finally, for the

impact of choice, we only offered a small alibi choice that supposedly affected the scope of the explanation, but not the actual decision.

We plan to address these limitations by conducting additional experiments that take into account different group sizes as well as groups that have mixed gender. For the lack of information of individuals' knowledge on AI, we plan to ask them about their perceived knowledge as well as adapt short quizzes from the literature that give a more objective measure of AI-knowledge and potential discrimination based on existing datasets. Finally, we plan to vary the scope of the choice, shifting from an alibi choice to one that influences the actual decision.

References

1. Akter, S., McCarthy, G., Sajib, S., Michael, K., Dwivedi, Y.K., D'Ambra, J., Shen, K.N.: Algorithmic Bias in Data-driven Innovation in the Age of AI. International Journal of Information Management 60, 102387 (2021)
2. Alter, A., Oppenheimer, D., Epley, N., Eyre, R.: Overcoming Intuition: Metacognitive Difficulty Activates Analytic Reasoning. Journal of experimental psychology: General 136, 569–575 (2007)
3. Angerschmid, A., Zhou, J., Theuermann, K., Chen, F., Holzinger, A.: Fairness and Explanation in AI-informed decision making. MAKE 4, 556–579 (2022)
4. Arrieta, B.A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., et al.: Explainable Artificial Intelligence (XAI): Concepts, Taxonomies, Opportunities and Challenges toward Responsible AI. Information Fusion 58, 82–115 (2020)
5. Aubert-Teillaud, B., Bran, A., Vaidis, D.C.: Expectation Violation and Cognitive Dissonance Theory: Proposal for an Epistemic Inconsistency Management Model. European Journal of Social Psychology 53, 1663–1679 (2023)
6. Babiker, A., Alshakhsi, S., Al-Thani, D., Montag, C., Ali, R.: Attitude Towards AI: Potential Influence of Conspiracy Belief, XAI Experience and Locus of Control. International Journal of Human–Computer Interaction, 1–13 (2024)
7. Bauer, K., Zahn, M. von, Hinz, O.: Expl(AI)ned: The Impact of Explainable Artificial Intelligence on Users' Information Processing. Information Systems Research 34, 1582–1602 (2023)
8. Beldad, A., Jong, M. de, Steehouder, M.: I trust not therefore it must be risky: Determinants of the perceived risks of disclosing personal data for e-government transactions. Computers in Human Behavior 27, 2233–2242 (2011)
9. Benjamins, R.: A Choices Framework for the Responsible Use of AI. AI Ethics 1, 49–53 (2021)
10. Beverungen, A., Lange, A.-C.: Cognition in High-Frequency Trading: The Costs of Consciousness and the Limits of Automation. Theory, Culture & Society 35, 75–95 (2018)
11. Blodorn, A., O'Brien, L.T., Kordys, J.: Responding to Sex-based Discrimination: Gender Differences in Perceived Discrimination and Implications for Legal Decision Making. Group Processes & Intergroup Relations 15, 409–424 (2012)
12. Burger, J., Cooper, H.: The Desirability of Control. Motivation and Emotion 3, 381–393 (1979)
13. Chamola, V., Hassija, V., Sulthana, A.R., Ghosh, D., Dhingra, D., Sikdar, B.: A Review of Trustworthy and Explainable Artificial Intelligence (XAI). IEEE Access 11, 78994–79015 (2023)
14. Chen, Z.: Ethics and Discrimination in Artificial Intelligence-enabled Recruitment practices. Humanities and Social Sciences Communications 10 (2023)
15. de Cremer, D., de Schutter, L.: How to use Algorithmic Decision-making to Promote Inclusiveness in Organizations. AI Ethics 1, 563–567 (2021)
16. de Melo, C., Marsella, S., Gratch, J.: People Do Not Feel Guilty About Exploiting Machines. ACM Transactions on Computer-Human Interaction 23 (2016)
17. de Wit, J., Das, E., Vet, R.: What Works Best: Objective Statistics or a Personal

Testimonial? An Assessment of the Persuasive Effects of Different Types of Message Evidence on Risk Perception. Health psychology 27, 110–115 (2008)
18. de Zoeten, M.: The Impact of an Explanation-Induced Expectation Violation in Explainable AI. In: Proceedings of the Thirtieth Americas Conference on Information Systems (2024)
19. de Zoeten, M., Ernst, C.-P. H., Rothlauf, F.: A Matter of Trust: How Trust in AI-Based Systems Changes During Interaction. In: Proceedings of the Twenty-ninth Americas Conference on Information Systems (2023)
20. de Zoeten, M., Ernst, C.-P. H., Rothlauf, F.: The Effect of Explainable AI on AI-Trust and its Antecedents over the Course of an Interaction. In: Proceedings of the Thirty-Second European Conference on Information Systems (2024)
21. Decrop, A.: Group Processes in Vacation Decision-Making. Journal of Travel & Tourism Marketing 18, 23–36 (2005)
22. Endsley, M.R.: Supporting Human-AI Teams:Transparency, Explainability, and Situation Awareness. Computers in Human Behavior 140, 107574 (2023)
23. Ferrario, A., Loi, M.: How Explainability Contributes to Trust in AI. In: Proceedings of the ACM Conference on Fairness, Accountability, and Transparency (2022)
24. Ghasemaghaei, M., Kordzadeh, N.: Understanding how algorithmic injustice leads to making discriminatory decisions: An obedience to authority perspective. Information & Management 61, 103921 (2024)
25. Glikson, E., Woolley, A.W.: Human Trust in Artificial Intelligence: Review of Empirical Research. ANNALS 14, 627–660 (2020)
26. Ha, T., Sah, Y.J., Park, Y., Lee, S.: Examining the Effects of Power Status of an Explainable Artificial Intelligence System on Users' Perceptions. Behaviour & Information Technology 41, 946–958 (2022)
27. Hangartner, D., Kopp, D., Siegenthaler, M.: Monitoring Hiring Discrimination through Online Recruitment Platforms. Nature 589, 572–576 (2021)
28. Hassija, V., Chamola, V., Mahapatra, A., Singal, A., Goel, D., Huang, K., Scardapane, S., Spinelli, I., Mahmud, M., Hussain, A.: Interpreting Black-Box Models: A Review on Explainable Artificial Intelligence. Cognitive Computation 16, 45–74 (2024)
29. Hoff, K.A., Bashir, M.: Trust in Automation: Integrating Empirical Evidence on Factors that Influence Trust. Human Factors 57, 407–434 (2015)
30. Horst, M., Kuttschreuter, M., Gutteling, J.M.: Perceived Usefulness, Personal Experiences, Risk Perception and Trust as Determinants of Adoption of E-Government Services in The Netherlands. Computers in Human Behavior 23, 1838–1852 (2007)
31. Hu, C., Hu, X., Liu, H.: Making Friends with Tourists before the Group Tour: A study of Guide-Tourist pre-tour Interactions based on the Social Situation Framework. Journal of Tourism Research 26 (2024)
32. Imana, B., Korolova, A., Heidemann, J.: Auditing for Discrimination in Algorithms Delivering Job Ads. In: Proceedings of the Web Conference 2021, pp. 3767–3778. ACM, New York, NY, USA (2021)
33. Kelly, S., Kaye, S.-A., Oviedo-Trespalacios, O.: What factors contribute to the acceptance of artificial intelligence? A systematic review. Telematics and Informatics 77, 101925 (2023)
34. Kim, T., Barasz, K., John, L.K.: Why am I Seeing this Ad? The Effect of Ad Transparency on Ad Effectiveness. Journal of Consumer Research 45, 906–932 (2019)
35. Kizilcec, R.F.: How much Information? Effects of Transparency on Trust in an Algorithmic Interface. In: 2016 CHI Conference on Human Factors in Computing Systems, pp. 2390–2395 (2016)
36. Klingbeil, A., Grützner, C., Schreck, P.: Trust and Reliance on AI — An experimental Study on the Extent and Costs of Overreliance on AI. Computers in Human Behavior 160, 108352 (2024)
37. Klusowski, J., Small, D.A., Simmons, J.P.: Does Choice Cause an Illusion of

Control? Psychological science 32, 159–172 (2021)
38. Landau, M.J., Kay, A.C., Whitson, J.A.: Compensatory Control and the Appeal of a Structured World. Psychological Bulletin 141, 694–722 (2015)
39. Laurim, V., Arpaci, S., Prommegger, B., Krcmar, H.: Computer, whom Should I Hire? – Acceptance Criteria for Artificial Intelligence in the Recruitment Process. In: Hawaii International Conference on System Sciences 2021 (2021)
40. Lee, J.D., See, K.A.: Trust in Automation: Designing for Appropriate Reliance. Human Factors 46, 50–80 (2004)
41. Leichtmann, B., Humer, C., Hinterreiter, A., Streit, M., Mara, M.: Effects of Explainable Artificial Intelligence on Trust and Human Behavior in a High-risk Decision Task. Computers in Human Behavior (2023)
42. Li, X., Hess, T.J., Valacich, J.S.: Why do we Trust new Technology? A Study of Initial Trust Formation with Organizational Information Systems. The Journal of Strategic Information Systems 17, 39–71 (2008)
43. Lin, Z.: An Empirical Investigation of User and System Recommendations in E-Commerce. Decision Support Systems 68, 111–124 (2014)
44. Lindell, M.K., Hwang, S.N.: Households' Perceived Personal Risk and Responses in a Multihazard Environment. Risk analysis : an official publication of the Society for Risk Analysis 28, 539–556 (2008)
45. Lockey, S., Gillespie, N., Holm, D., Someh, I.A.: A Review of Trust in Artificial Intelligence: Challenges, Vulnerabilities and Future Directions. In: Hawaii International Conference on System Sciences 2021 (2021)
46. Lukyanenko, R., Maass, W., Storey, V.C.: Trust in Artificial Intelligence: From a Foundational Trust Framework to Emerging Research Opportunities. Electronic Markets 32, 1993–2020 (2022)
47. Manning, K.L., Carroll, B.A., Carp, R.A.: Does Age Matter? Judicial Decision Making in Age Discrimination Cases. Social Science Quarterly 85, 1–18 (2004)
48. Mayer, R.C., Davis, J.H., Schoorman, F.D.: An Integrative Model of Organizational Trust. AMR 20, 709–734 (1995)
49. Meuwissen, M., Bollen, L.: Transparancy versus Explainability in AI. Unpublished (2021)
50. Park, S.-Y., Cho, M., Kim, S.: The effect of CSR expectancy violation: value from expectancy violation theory and confirmation bias. Journal of Marketing Communications 27, 365–388 (2021)
51. Romero Meza, L., D'Urso, G.: Navigating the Paradox of Choice in the Digital Age: A Scoping Review on the Potential Role of Recommender Systems. World Futures, 108–137 (2025)
52. Shin, D.: The effects of Explainability and Causability on Perception, Trust, and Acceptance: Implications for Explainable AI. International Journal of Human-Computer Studies 146 (2021)
53. Sjöberg, L.: The Different Dynamics of Personal and General Risk. Risk Management 5, 19–34 (2003)
54. Staegemann, D., Haertel, C., Daase, C., Pohl, M., Abdallah, M., Turowski, K.: A Review on Large Language Models and Generative AI in Banking. In: Proceedings of the 7th International Conference on Finance, Economics, Management and IT Business, pp. 267–278 (2025)
55. Staegemann, D., Volk, M., Jamous, N., Turowski, K.: Understanding-Issues-in-Big-Data-Applications-A-Multidimensional-Endeavor. In: Proceedings of the Twenty-fifth Americas Conference on Information Systems (2019)
56. Sturman, M., Hannon, J., Milkovich, G.: Computerized Decision Aids for Flexible Benefits Decisions: The Effects of an Expert System and Decision Support System on Employee Intentions and Satisfaction with Benefits. Personnel Psychology 49, 883–908 (1996)
57. Tilmes, N.: Disability, Fairness, and Algorithmic Bias in AI Recruitment. Ethics and Information Technology 24 (2022)
58. Tversky, A., Kahnemann, D.: Loss Aversion in Riskless Choice a Reference

Dependent Model. The Quarterly Journal of Economics, 1039–1061 (1991)
59. Uggirala, A., Gramopadhye, A.K., Melloy, B.J., Toler, J.E.: Measurement of Trust in Complex and Dynamic Systems Using a Quantitative Approach. International Journal of Industrial Ergonomics 34, 175–186 (2004)
60. van der Linden, S.: On the Relationship between Personal Experience, Affect and Risk Perception: The case of Climate Change. European Journal of Social Psychology 44, 430–440 (2014)
61. Walpole, H.D., Wilson, R.S.: Extending a broadly applicable measure of risk perception: the case for susceptibility. Journal of Risk Research 24, 135–147 (2021)
62. Wilson, R.S., Zwickle, A., Walpole, H.: Developing a Broadly Applicable Measure of Risk Perception. Risk analysis : an official publication of the Society for Risk Analysis 39, 777–791 (2019)
63. Yuan, C.W., Bi, N., Lin, Y.-F., Tseng, Y.-H.: Contextualizing User Perceptions about Biases for Human-Centered Explainable Artificial Intelligence. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, pp. 1–15. ACM, New York, NY, USA (2023)
64. Zhang, Y., Liao, Q.V., Bellamy, R.K.E.: Effect of Confidence and Explanation on Accuracy and Trust Calibration in AI-assisted Decision Making. In: Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, pp. 295–305. ACM, New York, NY, USA (2020)