4-24-2023

DESIGNING FAIR AI SYSTEMS: HOW EXPLANATION SPECIFICITY INFLUENCES USERSâ€™ PERCEIVED FAIRNESS AND TRUSTING INTENTIONS

Yiliao Song
RMIT University, yiliao.song@rmit.edu.au

Tingru Cui
University of Melbourne, tingru.cui@unimelb.edu.au

Feng Liu
University of Melbourne, fengliu.ml@gmail.com

DESIGNING FAIR AI SYSTEMS: HOW EXPLANATION SPECIFICITY INFLUENCES USERS' PERCEIVED FAIRNESS AND TRUSTING INTENTIONS

Research in Progress

Yiliao Song
RMIT University
Melbourne, VIC 3000, Australia
yiliao.song@rmit.edu.au

Tingru Cui
University of Melbourne
Parkville, VIC 3010, Australia
tingru.cui@unimelb.edu.au

Feng Liu
University of Melbourne
Parkville, VIC 3010, Australia
feng.liu1@unimelb.edu.au

Abstract

Artificial intelligence (AI) is revolutionizing the way we make decisions, but it is rarely perfect, and human-centric AI calls for a thorough empirical understanding of how the the theoretical fairness notion translates into perceptions of fairness in practical scenarios. Drawing upon the explainable artificial intelligence literature and elaboration likelihood model, we investigate the interaction effects of explanation specificity of AIs and issue involvement of users. We used a 3x2 experiment design with 456 participants to verify the proposed research model. We found that for individuals of low issue involvement, AI with global explanation is more effective, while AI feature-based explanation is more effective in influencing high issues involved individuals on their fairness perceptions of AI decisions, consequently leading to their trusting intentions towards AI decision-making systems. This study significantly contributes to the theoretical landscape of AI fairness and human-AI interaction, and provide important practical contributions for AI designers.

Keywords: AI, decision support, explanation, perceived fairness.

1 Introduction

Modern lives are increasingly shaped by data-driven decisions, often made by systems that use Artificial Intelligence (AI) algorithms. These systems have the potential to augment human well-being in many ways (Rahwan *et al.*, 2019). While AI leads to more efficient and optimal decision outcomes (Lepri *et al.*, 2018), it often includes a downside: due to biased input data or faulty algorithms, unfair AI decision-making systems may potentially reinforce racial or gender stereotypes, marginalize minorities, or flat-out denigrate certain members of society (Teodorescu *et al.*, 2021). For example, the COMPAS algorithm disproportionately assigned a higher risk score of recidivism to black defendants than to white defendants (Chouldechova, 2017). Similarly, an algorithm of reviewing resume for Amazon penalizes resumes that included the words associated with women (Dastin, 2018).

There has been an increasing focus in the research community on understanding and improving the fairness of AI decision-making systems (e.g., Jobin *et al.*, 2019; Pastaltzidis *et al.*, 2022). However, the theoretical discrimination-aware AI approaches often lack behavioral studies investigating how people perceive fairness of AI decision-making systems (Veale *et al.*, 2018; Teodorescu *et al.*, 2021). People's fairness perception can be complicated and nuanced. Designing and implementing human-centric AI calls for a thorough empirical understanding of how the the theoretical fairness notion is translated into perceptions of fairness in practical scenarios (Barabas *et al.*, 2020). One approach proposed is increasing user understanding by novel explainable artificial intelligence (XAI) methods (Dodge *et al.*, 2019; Miller, 2019). Nevertheless, a user perspective research in XAI faces multiple challenges: First, it is unclear how users actually assess and understand the explanations (Miller, 2019). Second, how users appraise explanations affects decision outcomes, such as trust, is unclear (Erlei *et al.*, 2020). Overall, we have limited understanding about how users evaluate explanations of AI decision-making systems under different contexts. Therefore, this paper aims to answer the following research question: *How do different XAI approaches and users' issue involvement jointly influence their perceived fairness of AI decisions as well as trusting intentions of AI decision-making systems?*

To address the above research gaps, this study differentiates AI' explanation specificity as more general global explanations and more specific feature-based explanations, and examines their varying effects on user response through elaboration likelihood model (ELM). The ELM suggests that the issue involvement conditions (low vs. high) of users (e.g., communication recipient) can shift users' encoding procedure of received information by adopting a central/peripheral route, and subsequently affect users' final attitudes toward the communication (Petty *et al.*,1981). Accordingly, we propose that AI's explanation specificity and users' issue involvement may jointly influence users' perceived fairness of AI decisions, which has a carry-over effect on their trusting intentions of AI decision-making systems. Specifically, we suggest that AI with more general globle explanation is more effective in interacting users with low issue involvement, whereas AI with more specific feature-based explanation is more effective for those with high issue involvement. In this research-in-progress paper, we used a 3x2 experiment design to verify the proposed effects.

This research makes a contribution to research around human experiences of AI in the fields of algorithmic decision-making and human-AI interaction. Our work offers new insight on the effectiveness of distinct explanation types made by AI on users' fairness perception and trusting intentions. By doing so, our work highlights a gap in prior work, surfaces the importance of designing and evaluating AI that actively engages in explaining the algorithm's logic of fairness to their users, and calls for more research that examines XAI for different demographic groups. Practically, the results address the major challenge of people's reluctance to trust algorithms and design guidelines for fair AI systems, which will be beneficial for organizations aiming to make more accurate and informed decisions as well as for firms that develop and sell algorithmic tools.

2 Theoretical Foundation

2.1 AI Fairness

AI may enable efficient, optimized, and data-driven decisions, and this is one of main drivers of increasing adoption of AI for decision-making (Newell and Marabelli, 2015). However, the fact that these AI-made decisions may influence the perceptions of decisions, regardless of the qualities of the actual decision-outcomes (Dietvorst *et al.*, 2015). These perceptions may in turn influence people's trust in and attitudes toward AI decision-making systems, which are critical aspects of workplaces, communities, and societies that allow people to thrive.

In recent years, the concept of fairness has regained prominence as a core objective in designing AI (Jobin *et al.*, 2019). The term 'AI fairness' generally means AI-made decision should not produce unjust, discriminatory, or disparate consequences (Shin and Park, 2019). AI fairness is endorsed as one of the four main principles for trustworthy AI by policy institutions like the OECD (2019), and it has been featured in more than 80 percent of guidelines for AI ethics (Jobin *et al.*, 2019). There is a growing body of work that aims to improve fairness, accountability, and interpretability of machine learning algorithms (Pastaltzidis *et al.*, 2022; Starke *et al.*, 2022). However, Veale et al. (2018) found that these approaches and tools are often built in isolation of specific users and user context. Addressing the societal implications of (un)fair AI systems requires more than mere technological solutions (Sloane and Moss, 2019; Barabas *et al.*, 2020). Designing and implementing human-centric AI calls for a thorough empirical understanding of when and why users perceive AI systems to be (un)fair (Teodorescu *et al.*, 2021). For example, prior studies have examined users' fairness perception and reactions to AI hiring systems (e.g., Hunkenschroer and Lütge, 2021; Gonzalez *et al.*, 2022), employee management AI in organizations (e.g., Robert *et al.*, 2020), and fairness during interaction with a dialogue system (e.g., Janzen *et al.*, 2018). This study will echo this direction of discussion to theorize on and predict the effects of various XAI design characteristics on users' perceived fairness, consequently their trusting intentions.

2.2 Explainable Artificial Intelligence (XAI)

It is often assumed in the XAI literature that explanations can support users to understand the outcome of the underlying model (Diakopoulos and Koliska, 2017; Miller, 2019). However, the mere presence of explanations does not necessarily improve users' perceptions of AI decision-making systems (Arrieta *et al.*, 2020). As of today, there is no conclusive empirical evidence showing that explanations facilitate people's fairness perceptions towards AI models. For example, some prior work found that explanations increased perceptions of fairness (e.g., Wang, 2018; Lai and Tan, 2019; Chu *et al.*, 2020), while others observed that explanations had no significant or negative effect on AI fairness (e.g., (Kizilcec, 2016; Wang *et al.*, 2020; Shulner-Tal *et al.*, 2022). Furthermore, different explanation styles may play distinct roles in influencing people's fairness perceptions towards AI models (e.g., Binns *et al.*, 2018; Dodge *et al.*, 2019). For example, Binns et al. (2018) found that case-based explanations had a negative influence on perceived fairness, especially compared to sensitivity-based explanations.

In the meanwhile, emerging XAI literature suggests two main approaches of explanation: (1) Global explanation, and (2) Local explanation (e.g., Pedreschi *et al.*, 2019; Setzu *et al.*, 2021). They differ in the amount of information the explanation should convey, which is termed as *XAI specificity*, in this study. *Global explanation* provides an overview of what an algorithm is doing as a whole (Pedreschi *et al.*, 2019). The aim of this type of explanation is to convey to a human what the algorithm is doing rather than explain the process that led to a specific prediction or decision. These methods often include summarized information about how a model uses features to produce prediction, or a simplified approximation of a black-box model (Pedreschi *et al.*, 2019). Another type of global explanation this

includes transparency aboud how the model was trained, the type of data that was used, or even simply reporting model performance statistics (Ben David *et al.*, 2021). *Local explanation*, on the other hand, provides a more detailed description of how the model came up with a specific prediction (Zhang *et al.*, 2020), e.g., *feature-based explanation* (Bauer *et al.*, 2021; Ben David *et al.*, 2021). In general, local explanation is more costly in time and resources since they are computed on a case-by-case basis rather than globally for the entire system.

In spite of recent attempts to categorize explanations, there remains a lack of clear guidelines regarding how much information is necessary and which type of explanation should be used depending on the context, task, type of end user (Kulesza *et al.*, 2013; Sokol and Flach, 2020). For example, Perez Vallejos et al. (2017) found that young people demanded algorithm-level information to perceive fairness. However, providing too much information about the algorithm might also reduce the perceived fairness (Kizilcec, 2016). Hence, this work aims to conclude a holistic view of scattered findings at the intersection of XAI and fair AI decision-making. It will develop better understanding how explanation specificity and the user characteristics joinly impact user's fairness perceptions of AI systems.

2.3 Elaboration Likelihood Model

The elaboration likelihood model (ELM; Petty and Cacioppo, 1986) provides a comprehensive framework for understanding the basic processes of organizing and developing effective persuasive communications. The ELM framework posits two routes to persuasion, namely central and peripheral routes. The central route to persuasion is more likely to occur under a person's careful and thoughtful consideration of information presented (Petty and Cacioppo, 1986). The peripheral route to persuasion, however, occurs as a result of some simple cues without necessitating scrutiny of information presented (Petty *et al.*, 1981). Empirical studies have consistently shown that the two determinants, attributes of the message deliverer and characteristics of the message recipient, can jointly influence the effectiveness of communication (Kitchen *et al.*, 2014).

Consistent with the ELM, our focus of the XAI explanation specificity is a key attribute of message, which should interact with the characteristics of users to influence the perception of AI-made. In the current research-in-progress paper, we concentrate on an essential characteristic of users - *Issue Involvement*, defined as the extent to which users perceive a message topic to be personally important or relevant (Petty and Cacioppo, 1990). Given the recognition that under low versus high issue involvement conditions, users take two distinct persuasion routes respectively (Petty and Cacioppo, 1986), we attempt to identify whether different levels of explanation specificity would match the information processing styles by central versus peripheral routes and then leads to variation in users' perceived fairness (Kitchen *et al.*, 2014).

3 Hypothesis Development

This study proposes that users' preferences for XAI specificity are context-dependent, in a way that is consistent with ELM. It investigates how explanations at two different levels of XAI specificity can determine users' perceived fairness and trusting intentions depending on users' levels of involvement with the issue. The research model for this study is shown in Figure 1.

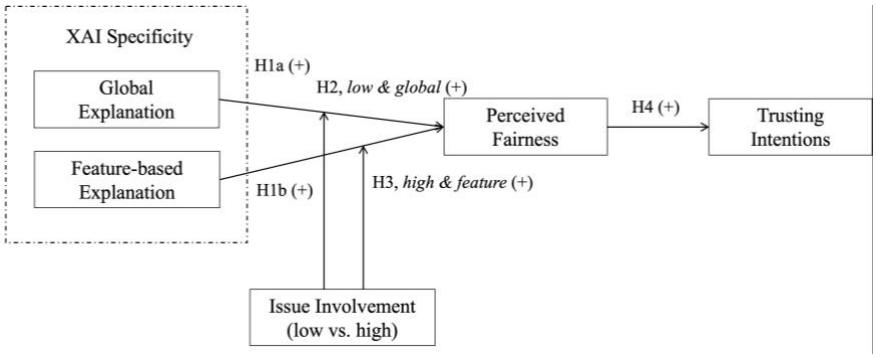

Figure 1. Research model.

3.1 Effects of XAI

In organizational context, prior studies showed that transparency of the decision-making process has an important impact on employees' perceived fairness(Wanberg *et al.*, 1999). In the context of AI decision-making, explanation also influences people's perceived fairness. Researchers have shown that some level of transparency of an algorithm would increase users' trust of the system even when the user's expectation is not aligned with the algorithm's outcome (Kizilcec, 2016; Shin, 2021). Qualifying this finding, Lee et al. (2019) found that outcome explanations had a significant influence on people's perceptions of AI fairness, but the direction of the effects largely depended on the context.

In particular, explanations help to increase knowledge-based trust in the system (Wang and Benbasat, 2016; Wang and Wang, 2019). Both global and feature-based explanations enhance the transparency of the AI system (Dodge *et al.*, 2019; Schoeffer *et al.*, 2021). They help users validate the AI system's decisions and increase the perceived understandability of AI systems (Bussone *et al.*, 2015; Meske and Bunde, 2020). When they understand the AI system's decision-making logic and process, people will judge an AI decision to be more fair. Therefore, we propose that both global and feature-based explanations increase users' perceived fairness in AI decisions:

H1a-b: *(a) Global explanation, and (b) feature-based explanation positively influence users' perceived fairness in AI decisions.*

3.2 XAI Specificity and Low Issue Involvement

We posit that low level of XAI specificity, i.e., globle explanation has a greater impact on users' fairness perception under the low issue involvement condition. The ELM suggests that low issue involvement renders individuals to assess the validity of a persuasive message through the peripheral route (Petty and Cacioppo, 1986). When the topic of message is less personally important or relevant, people tend to follow the interest of cognitive economy and focus less on message content itself to make decision (Petty and Cacioppo, 1990). Prior persuasion and marketing research has examined information specificity in the online advertising context of retargeting and suggested that ads shown early in the purchase process are more effective when they present abstract information, while ads shown later in the process are more effective when they present more specific information (e.g., Lambrecht and Tucker, 2013).

Similarly, the globle explanation of AI includes summarized information about how a model produces decision, such as the data or algorithm used. Individuals assess these abstract and high level information to infer the merits of decision advised. For example, Ramon et al. (2021) found that participants better understand model's decision making when presented with higher-level representation of features, instead of original features. Thus, when individuals feel that the topic of AI decision-making is less personally relevant, such as product pricing in the early purchase process, they would heuristically focus

on and tend to be influenced by the globle explanation of AI. Globle explanation hence can help users weigh the system's accuracy against its bias and determine that AI system has provided accurate support, consequently enhancing their fairness perception (Setzu *et al.*, 2021).

At the same time, explanation with high level of specificity, i.e., feature-based explanation may not work effectively for the peripheral route to persuasion under the low issue involvement condition. Cognitive psychology studies have indicated that low issue involvement decreases message recipients' motivation to devote extra cognitive resources to process detailed information of a message (Petty and Cacioppo, 1986). Feature-based explanation usually require additional cognitive efforts to process. However, the original features used by the model might not always lead to the most comprehensible explanations. For example, Poursabzi-Sangdeh et al. (2021) analyzed human-AI decision-making for the case of house price estimation and found that performance did not increase in the presence of local explanations which is likely due to information overload. As a result, individuals with low issue involvement may not want to devote the necessary cognitive energy to systematically consider all available feature information (DeBono and Harnish, 1988). Thus, we hypothesize:

H2: *For users under low issue involvement, AI with global explanation is more effective than feature-based explanation in positively influencing users' perceived fairness in AI decisions.*

3.3 XAI Specificity and High Issue Involvement

We also propose that high level of XAI specificity, i.e., feature-based explanation has a greater impact on users' fairness perception under the high issue involvement condition. According to the ELM, when a decision-making situation is personally involved for users, they tend to engage in more systematic or central route processing (Petty and Cacioppo, 1986). In other words, they have an increased need to understand the details of the decision-making process and are more likely to focus on encoding and interpreting more detailed linguistical arguments presented (Petty and Cacioppo, 1990). Consequently, simple summarized information of the model would not be targeted as the key information for users and thus have limited impact on ther comprehension of the explanation (Petty *et al.*, 1981).

In contrast, feature-based explanations provide high levels of transparency in an algorithm, which can afford people a sense of personalization. Specific and accountable information affords users a sense of confidence, which, in turn, promotes a sense of satisfaction and assurance (Kizilcec, 2016). As a result, users would like to pay more attention to elaborate and evaluate the explanation delivered by AI. This activation is consistent with the central route to persuasion in the ELM (Petty and Cacioppo, 1990) meaning that the effects of feature-based explanations from AIs should be more salient in positively influencing users' perceived fairness when they are highly involved. Consistent with this perspective, persuasive marketing literature has shown that when consumers are more involved with the advertised product, they favor more specific and detailed information, rather than simple general information (e.g., Xue, 2014). In the context of AI, prior studies also have observed the positive effects of more detailed verbal social cues from conversational agents on persuasion when the message topic is personally relevant (e.g., Bickmore *et al.*, 2009). Hence, with more transparency via feature-based explanation, highly involved users are able to understand the logic of an AI system and this leads to assurance in the AI decisions (Renjith *et al.*, 2020; Zhang *et al.*, 2020). This further enhances their perceptions of the decisions to be more personalized and trustworthy (Shin, 2021). We conclude our third hypothesis as:

H3: *For users under high issue involvement, AI with feature-based explanation is more effective than global explanation in positively influencing users' perceived fairness in AI decisions.*

3.4 Perceived Fairness and Trusting Intentions

Next, we consider the relationship between perceived fairness and users' trusting intentions in the AI decision-making systems. In line with prior justice literature, when individuals feel they are being treated

fairly, they tend to develop favorable impressions toward the recommendation (Greenberg, 2011; Pérez-Rodríguez, Topa and Beléndez, 2019). When users generate greater perceived fairness of the AI-made decisions, they experience positive cognitions and affects, consequently their trusting intentions will further shift toward the AI systems (Shin, 2021). Understanding relations between fairness and trust is nontrivial in the social interaction context such as marketing and services. For example, Roy et al. (2015) showed that customers' perceptions of fair treatment play a positive role in engendering trust in the banking context. In the context of AI, trust is considered as the belief that an AI's services or reported results are reliable and trustworthy so that AI can fulfill obligations in an exchange relationship with the user (Shin and Park, 2019). Kasinidou et al. (2021) found that people's perception of a 'not fair decision' affects the participants' trust in an AI decision-making system. Similarly, Shin (2021) obtained similar conclusions that perception of fairness had a positive effect on trust in an AI decision-making system. These motivate us to propose the following hypothesis:

H4: *Users' perceived fairness positively influences their trusting intentions in AI decision-making systems.*

4 Research Method

4.1 Experimental Design

To investigate our hypotheses regarding the joint influence of XAI specificity and issue involvement, we adopted a 3 (baseline no explanation vs. global vs. feature-based explanation) × 2 (low vs. high issue involvement) between-subject design with random assignment. Specifically, for the explanation specificity, in the *baseline no explanation* group, the AI simply makes decision without additional information. In the *global explanation* condition, participants were given very general information, only the type and extent of data. For example, a global explanation is "based on data from apartment rental over several years, the algorithm recommendation is $350 per night." While in the *feature-based explanation* condition, participants were given a detailed account about the features and their importance for specific prediction. For example, a feature-based explanation is "based on data from apartment rental over several years, previous rentals in last month, and current market demand, the algorithm recommendation is $350 per night."

To identify tasks with varying levels of issue involvement, we adapted dynamic product price-setting and loan application tasks used in prior AI decision-making studies (e.g., Ramon *et al.*, 2021; Chen *et al.*, 2022). In the *low involvement* context of dynamic product price-setting, the user only needs to choose one desired product without much inputs. In the *high involvement* context of loan application, user input more personal details to apply for a loan, and therefore they are more involved in the application.

4.2 Experimental Procedure and Measurements

We recruited a total of 456 participants to ensure a sufficient statistical power of 0.80 (226 women, *age mean* = 38.12, *SD* = 10.25) from Amazon Mechanical Turk for a 15-minute online experiment (presented as helping us to train an AI tool), remunerated with US$5. Four inclusion criteria were used: location (i.e., the United States), language (i.e., English), not in the top 4% of workers in terms of volume (i.e., people who complete surveys almost professionally), and above a 98% worker approval rate (to ensure high quality). Careless participants were excluded based on failed attention checks (e.g., key characteristic of AI, key function the AI in this study). Participants were randomly assigned to treatment groups and no significant differences (all $p > 0.05$) on gender and age across all 6 groups.

In the experiment, participants assessed one dynamic product price-setting (either lower priced or higher priced) or completed one loan application (either approved or rejected). The outomes are randomized across all participants to avoid any confounding effect. Specifically, for dynamic product price-setting,

participants are asked to learn the price changes over time to meet their imagined demand of a rental car, and the AI-decided price were lower or higher than average by 50%. For loan application, participants were asked to input their personal information to apply for a loan with either approve or reject decision made by AI. Explanations revealed how price or loan application decision was made.

After completing the task, participants were asked to complete a survey. We first asked participants to indicate their perceived fairness (adapted from Franke *et al.*, 2013; Lee and Baykal, 2017). We also asked participants about their trusting intentions (McKnight *et al.*, 2002) by inquiring if they would recommend using the AI systems by e-commerce or bank employees. Afterward, we collected demographic information (age, gender and education level), familiarity with price-setting or loans on three items (adapted from Gefen, 2000), familiarity with AI (Logg *et al.*, 2019) and disposition to trust (Gefen, 2000) as control variables. All items are measured on a 7-point Likert scale.

5 Preliminary Results

The manipulation of explanation specificity was assessed by comparing the explanation they have been exposed being specific or general. A t-test ($t = 15.83$, $p < 0.05$) showed that subjects in feature-based explanation condition felt the AI-made explanation more specific (mean=5.58, std=1.20) than those in globle explanation condition (mean=2.12, std=1.21). Similarly, the manipulation of issue involvement was assessed by comparing their personal relevance with the experiment task. A t-test ($t = 6.741$, $p < 0.05$) showed that subjects in high involvement condition felt the focal task more personally relevant (mean=5.56, std=1.42) than those in low issue involvement condition (mean=3.50, std=1.89).

All statistical tests were carried out at a 5% level of significance. Exploratory factor analysis (EFA) was conducted to test the instrument's convergent and discriminant validity for perceptual constructs. First, we found a five-factor structure with eigenvalues greater than 1.0 and loadings above 0.7, significantly higher than the cross-loadings. Cronbach's alphas (0.88-0.96) and composite reliability (0.89-0.95) for all constructs were above 0.7. Average variance extracted (AVE) for all constructs was above 0.5 (0.72-0.89). Lastly, the square roots of AVE were greater than 0.7 and the corresponding inter-construct correlations. All constructs had sufficient convergent and discriminant validity. Table 1 displays the descriptive statistics of all variables included in the measurement model.

	Means (S.D.)	**CR**	**Correlations & AVE in parenthesis**				
Variable			**1**	**2**	**3**	**4**	**5**
Perceived fairness (1)	4.22 (1.48)	0.92	(0.85)				
Trusting intentions (2)	4.56 (1.53)	0.91	0.36	(0.82)			
Familiarity with price-setting or loans (3)	4.39 (1.20)	0.89	0.32	0.32	(0.72)		
Familiarity with AI (4)	4.26 (1.33)	0.95	0.06	-0.04	0.13	(0.89)	
Propensity to trust (5)	4.36 (1.41)	0.90	0.02	0.08	0.15	0.28	(0.83)

Table 1. Descriptive statistics.

Six control variables were treated as covariates in ANCOVA. ANCOVA was conducted on the perceived fairness (see Table 2). No covariates had significant interaction with independent variables. The main effect of globle explanation on perceived fairness is significant ($F(1, 221) = 2.650$, $p < 0.01$), however, the effect of feature-based explanation is not signiticant ($F(1, 221) = 1.976$, $p > 0.05$). Hence, H1a was supported withle H1b was rejected. In support of H2, global explanation and low issue involvement have a positive significant interaction impact on the perceived fairness ($F(1, 107) = 7.024$, $p < 0.05$), Similarly, feature-based explanation and high issue involvement have a positive significant interaction impact on the perceived fairness ($F(1, 107) = 8.147$, $p < 0.01$). This supports H3. Regressions were conducted on the dependent variable, trusting intentions. After excluding the effects of all

manipulated factors and control variables, perceived fairness still had a significant positive effect on trusting intention (t=3.62, $p < 0.01$). Thus, H4 was supported.

Source		df	Mean Square	F	p
Covariates	Age	1	3.255	3.250	0.075
	Gender	1	1.564	2.005	0.325
	Education	1	0.285	0.230	0.631
	Familiarity with price-setting or loan application	1	2.453	1.650	0.104
	Familiarity with AI	1	3.566	2.965	0.088
	Propensity to trust	1	2.123	1.967	0.184
Main effect	Global explanation (GLO)	1	3.817	2.650	0.009**
	Feature-based explanation (FEA)	1	1.025	1.976	0.121
	Issue involvement (INV)	1	2.237	1.984	0.162
Interaction effect	GLO x low INV	1	4.832	7.024	0.023*
	FEA x high INV	1	5.331	8.147	0.006**
Notes. * $p < 0.05$; ** $p < 0.01$, *** $p < 0.001$					

Table 2. Results of ANCOVA (Dependent variable: perceived fairness).

6 Discussion and Conclusion

This research-in-progress paper offers novel insights that the effect of XAI depends on charactersitics of users while feature-based explanation may not directly affect users' fairness perception. Specifically, for users of low issue involvement, AI with global explanation is more effective, while AI with feature-based explanation is more effective in influencing high issues involved users on their fairness perceptions of AI decisions, consequently leading to their trusting intentions in AI decision-making systems. We will conduct more systematic data analyse and explore the impact of more granular explanation categories (e.g., model-specific versus model-agnostic, feature importance versus feature versus feature interactions) in the future research. This study makes three contributions. First, we empirically show how an IT artifact's persuasive design can shape individual perception and behavioral intention. Second, we contribute to the XAI literature by differentiating and examining the effects of global and feature-based explanations for more effective communication of AI. Third, it extends the ELM and adds on the persuasive technology design literature by identifying a novel and complete view on both the communication message as well as the user characteristics. This study also offers pragmatic insights for AI designers on how to use exaplanation features to improve AI adoption.

References

Arrieta, A.B. *et al.* (2020) 'Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI', *Information fusion*, 58, pp. 82–115.

Barabas, C. *et al.* (2020) 'Studying up: reorienting the study of algorithmic fairness around issues of power', in. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 167–176.

Bauer, K. *et al.* (2021) 'Expl (AI) n it to me–explainable AI and information systems research', *Business & Information Systems Engineering*, 63(2), pp. 79–82.

Ben David, D., Resheff, Y.S. and Tron, T. (2021) 'Explainable AI and Adoption of Financial Algorithmic Advisors: An Experimental Study', in. *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society*, pp. 390–400.

Bickmore, T., Schulman, D. and Shaw, G. (2009) 'DTask and LiteBody: Open source, Standards-Based tools for building Web-Deployed embodied conversational agents', in. *International workshop on intelligent virtual agents*, Springer, pp. 425–431.

Binns, R. *et al.* (2018) '"It's Reducing a Human Being to a Percentage" Perceptions of Justice in Algorithmic Decisions', in. *Proceedings of the 2018 Chi conference on human factors in computing systems*, pp. 1–14.

Bussone, A., Stumpf, S. and O'Sullivan, D. (2015) 'The role of explanations on trust and reliance in clinical decision support systems', in. *2015 international conference on healthcare informatics*, IEEE, pp. 160–169.

Chen, Z. *et al.* (2022) 'What's in a Face? An Experiment on Facial Information and Loan-Approval Decision', *Management Science* [Preprint].

Chouldechova, A. (2017) 'Fair prediction with disparate impact: A study of bias in recidivism prediction instruments', *Big data*, 5(2), pp. 153–163.

Chu, E., Roy, D. and Andreas, J. (2020) 'Are visual explanations useful? a case study in model-in-the-loop prediction', *arXiv preprint arXiv:2007.12248* [Preprint].

Dastin, J. (2018) 'Amazon scraps secret AI recruiting tool that showed bias against women', in *Ethics of Data and Analytics*. Auerbach Publications, pp. 296–299.

DeBono, K.G. and Harnish, R.J. (1988) 'Source expertise, source attractiveness, and the processing of persuasive information: A functional approach.', *Journal of Personality and social Psychology*, 55(4), p. 541.

Diakopoulos, N. and Koliska, M. (2017) 'Algorithmic transparency in the news media', *Digital journalism*, 5(7), pp. 809–828.

Dietvorst, B.J., Simmons, J.P. and Massey, C. (2015) 'Algorithm aversion: people erroneously avoid algorithms after seeing them err.', *Journal of Experimental Psychology: General*, 144(1), p. 114.

Dodge, J. *et al.* (2019) 'Explaining models: an empirical study of how explanations impact fairness judgment', in. *Proceedings of the 24th international conference on intelligent user interfaces*, pp. 275–285.

Erlei, A. *et al.* (2020) 'Impact of algorithmic decision making on human behavior: Evidence from ultimatum bargaining', in. *Proceedings of the AAAI conference on human computation and crowdsourcing*, pp. 43–52.

Franke, N., Keinz, P. and Klausberger, K. (2013) '"Does this sound like a fair deal?": Antecedents and consequences of fairness expectations in the individual's decision to participate in firm innovation', *Organization science*, 24(5), pp. 1495–1516.

Gefen, D. (2000) 'E-commerce: the role of familiarity and trust', *Omega*, 28(6), pp. 725–737.

Gonzalez, M.F. *et al.* (2022) 'Allying with AI? Reactions toward human-based, AI/ML-based, and augmented hiring processes', *Computers in Human Behavior*, 130, p. 107179.

Greenberg, J. (2011) 'Organizational justice: The dynamics of fairness in the workplace.'

Hunkenschroer, A.L. and Lütge, C. (2021) 'How to improve fairness perceptions of AI in hiring: the crucial role of positioning and sensitization', *AI Ethics J*. [Preprint].

Janzen, S. *et al.* (2018) 'inSIDE Fair Dialogues: Assessing and Maintaining Fairness in Human-Computer-Interaction.', in. *Mensch & Computer Workshopband*.

Jobin, A., Ienca, M. and Vayena, E. (2019) 'The global landscape of AI ethics guidelines', *Nature Machine Intelligence*, 1(9), pp. 389–399.

Kasinidou, M. *et al.* (2021) 'I agree with the decision, but they didn't deserve this: Future Developers' Perception of Fairness in Algorithmic Decisions', in. *Proceedings of the 2021 acm conference on fairness, accountability, and transparency*, pp. 690–700.

Kitchen, P.J. *et al.* (2014) 'The elaboration likelihood model: review, critique and research agenda', *European Journal of Marketing* [Preprint].

Kizilcec, R.F. (2016) 'How much information? Effects of transparency on trust in an algorithmic interface', in. *Proceedings of the 2016 CHI conference on human factors in computing systems*, pp. 2390–2395.

Kulesza, T. *et al.* (2013) 'Too much, too little, or just right? Ways explanations impact end users' mental models', in. *2013 IEEE Symposium on visual languages and human centric computing*, IEEE, pp. 3–10.

Lai, V. and Tan, C. (2019) 'On human predictions with explanations and predictions of machine learning models: A case study on deception detection', in. *Proceedings of the conference on fairness, accountability, and transparency*, pp. 29–38.

Lambrecht, A. and Tucker, C. (2013) 'When does retargeting work? Information specificity in online advertising', *Journal of Marketing research*, 50(5), pp. 561–576.

Lee, M.K. *et al.* (2019) 'Procedural justice in algorithmic fairness: Leveraging transparency and outcome control for fair algorithmic mediation', *Proceedings of the ACM on Human-Computer Interaction*, 3(CSCW), pp. 1–26.

Lee, M.K. and Baykal, S. (2017) 'Algorithmic mediation in group decisions: Fairness perceptions of algorithmically mediated vs. discussion-based social division', in. *Proceedings of the 2017 acm conference on computer supported cooperative work and social computing*, pp. 1035–1048.

Lepri, B. *et al.* (2018) 'Fair, transparent, and accountable algorithmic decision-making processes', *Philosophy & Technology*, 31(4), pp. 611–627.

Logg, J.M., Minson, J.A. and Moore, D.A. (2019) 'Algorithm appreciation: People prefer algorithmic to human judgment', *Organizational Behavior and Human Decision Processes*, 151, pp. 90–103.

McKnight, D.H., Choudhury, V. and Kacmar, C. (2002) 'The impact of initial consumer trust on intentions to transact with a web site: a trust building model', *The journal of strategic information systems*, 11(3–4), pp. 297–323.

Meske, C. and Bunde, E. (2020) 'Transparency and trust in human-AI-interaction: The role of model-agnostic explanations in computer vision-based decision support', in. *International Conference on Human-Computer Interaction*, Springer, pp. 54–69.

Miller, T. (2019) 'Explanation in artificial intelligence: Insights from the social sciences', *Artificial intelligence*, 267, pp. 1–38.

Newell, S. and Marabelli, M. (2015) 'Strategic opportunities (and challenges) of algorithmic decision-making: A call for action on the long-term societal effects of "datification"', *The Journal of Strategic Information Systems*, 24(1), pp. 3–14.

OECD (2019) 'Recommendation of the Council on Artificial Intelligence'. Available at: https://legalinstruments.oecd.org/en/instruments/OECD-LEGAL-0449.

Pastaltzidis, I. *et al.* (2022) 'Data augmentation for fairness-aware machine learning: Preventing algorithmic bias in law enforcement systems', in. *2022 ACM Conference on Fairness, Accountability, and Transparency*, pp. 2302–2314.

Pedreschi, D. *et al.* (2019) 'Meaningful explanations of black box AI decision systems', in. *Proceedings of the AAAI conference on artificial intelligence*, pp. 9780–9784.

Perez Vallejos, E. *et al.* (2017) 'Young people's policy recommendations on algorithm fairness', in. *Proceedings of the 2017 ACM on web science conference*, pp. 247–251.

Pérez-Rodríguez, V., Topa, G. and Beléndez, M. (2019) 'Organizational justice and work stress: The mediating role of negative, but not positive, emotions', *Personality and Individual Differences*, 151, p. 109392.

Petty, R.E. and Cacioppo, J.T. (1986) 'The elaboration likelihood model of persuasion', in *Communication and persuasion*. Springer, pp. 1–24.

Petty, R.E. and Cacioppo, J.T. (1990) 'Involvement and persuasion: Tradition versus integration.'

Petty, R.E., Cacioppo, J.T. and Goldman, R. (1981) 'Personal involvement as a determinant of argument-based persuasion.', *Journal of personality and social psychology*, 41(5), p. 847.

Poursabzi-Sangdeh, F. *et al.* (2021) 'Manipulating and measuring model interpretability', in. *Proceedings of the 2021 CHI conference on human factors in computing systems*, pp. 1–52.

Rahwan, I. *et al.* (2019) 'Machine behaviour', *Nature*, 568(7753), pp. 477–486.

Ramon, Y. *et al.* (2021) 'Can metafeatures help improve explanations of prediction models when using behavioral and textual data?', *Machine Learning*, pp. 1–40.

Renjith, S., Sreekumar, A. and Jathavedan, M. (2020) 'An extensive study on the evolution of context-aware personalized travel recommender systems', *Information Processing & Management*, 57(1), p. 102078.

Robert, L.P. *et al.* (2020) 'Designing fair AI for managing employees in organizations: a review, critique, and design agenda', *Human–Computer Interaction*, 35(5–6), pp. 545–575.
Roy, S.K., Devlin, J.F. and Sekhon, H. (2015) 'The impact of fairness on trustworthiness and trust in banking', *Journal of Marketing Management*, 31(9–10), pp. 996–1017.
Schoeffer, J., Machowski, Y. and Kuehl, N. (2021) 'A study on fairness and trust perceptions in automated decision making', *arXiv preprint arXiv:2103.04757* [Preprint].
Setzu, M. *et al.* (2021) 'Glocalx-from local to global explanations of black box AI models', *Artificial Intelligence*, 294, p. 103457.
Shin, D. (2021) 'The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI', *International Journal of Human-Computer Studies*, 146, p. 102551.
Shin, D. and Park, Y.J. (2019) 'Role of fairness, accountability, and transparency in algorithmic affordance', *Computers in Human Behavior*, 98, pp. 277–284.
Shulner-Tal, A., Kuflik, T. and Kliger, D. (2022) 'Fairness, explainability and in-between: understanding the impact of different explanation methods on non-expert users' perceptions of fairness toward an algorithmic system', *Ethics and Information Technology*, 24(1), pp. 1–13.
Sloane, M. and Moss, E. (2019) 'AI's social sciences deficit', *Nature Machine Intelligence*, 1(8), pp. 330–331.
Sokol, K. and Flach, P. (2020) 'Explainability fact sheets: a framework for systematic assessment of explainable approaches', in. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 56–67.
Starke, C. *et al.* (2022) 'Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature', *Big Data & Society*, 9(2), p. 20539517221115188.
Teodorescu, M.H. *et al.* (2021) 'Failures of Fairness in Automation Require a Deeper Understanding of Human-ML Augmentation.', *MIS Quarterly*, 45(3).
Veale, M. and Binns, R. (2017) 'Fairer machine learning in the real world: Mitigating discrimination without collecting sensitive data', *Big Data & Society*, 4(2), p. 2053951717743530.
Veale, M., Van Kleek, M. and Binns, R. (2018) 'Fairness and accountability design needs for algorithmic support in high-stakes public sector decision-making', in. *Proceedings of the 2018 chi conference on human factors in computing systems*, pp. 1–14.
Wanberg, C.R., Gavin, M.B. and Bunce, L.W. (1999) 'Perceived fairness of layoffs among individuals who have been laid off: A longitudinal study', *Personnel Psychology*, 52(1), pp. 59–84.
Wang, A. (2018) 'Procedural justice and risk-assessment algorithms', *Available at SSRN 3170136* [Preprint].
Wang, R., Harper, F.M. and Zhu, H. (2020) 'Factors influencing perceived fairness in algorithmic decision-making: Algorithm outcomes, development procedures, and individual differences', in. *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems*, pp. 1–14.
Wang, W. and Benbasat, I. (2016) 'Empirical assessment of alternative designs for enhancing different types of trusting beliefs in online recommendation agents', *Journal of management information systems*, 33(3), pp. 744–775.
Wang, W. and Wang, M. (2019) 'Effects of sponsorship disclosure on perceived integrity of biased recommendation agents: Psychological contract violation and knowledge-based trust perspectives', *Information Systems Research*, 30(2), pp. 507–522.
Xue, F. (2014) 'It looks green: Effects of green visuals in advertising on Chinese consumers' brand perception', *Journal of International Consumer Marketing*, 26(1), pp. 75–86.
Zhang, Y., Liao, Q.V. and Bellamy, R.K. (2020) '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.