Effects of Fairness and Explanation on Trust in Ethical AI

Alessa Angerschmid[2], Kevin Theuermann[3], Andreas Holzinger[2,3,4,5], Fang Chen[1], and Jianlong Zhou[1,2⋆]

[1] Human-Centered AI Lab, University of Technology Sydney, Australia
{jianlong.zhou,fang.chen}@uts.edu.au
[2] Human-Centered AI Lab, Medical University Graz, Austria
alessa.angerschmid@human-centered.ai,andreas.holzinger@human-centered.ai
[3] Graz University of Technology, Austria
kevin.theuermann@egiz.gv.at
[4] University of Natural Resources and Life Sciences Vienna, Austria
[5] xAI Lab, Alberta Machine Intelligence Institute, Canada

Abstract. AI ethics has been a much discussed topic in recent years. Fairness and explainability are two important ethical principles for trustworthy AI. In this paper, the impact of AI explainability and fairness on user trust in AI-assisted decisions is investigated. For this purpose, a user study was conducted simulating AI-assisted decision making in a health insurance scenario. The study results demonstrated that fairness only affects user trust when the fairness level is low, with a low fairness level reducing user trust. However, adding explanations helped users increase their trust in AI-assisted decision making. The results show that the use of AI explanations and fairness statements in AI applications is complex: we need to consider not only the type of explanations, but also the level of fairness introduced. This is a strong motivation for further work.

Keywords: AI explanation · AI fairness · trust · AI ethics

1 Introduction

Artificial Intelligence (AI) informed decision-making is claimed to lead to faster and better decision outcomes, and has been increasingly used in our society from the decision-making of daily lives such as recommending movies and books to making more critical decisions such as medical diagnoses, credit risk prediction, and shortlisting talents in recruitment. Among such AI-informed decision-making tasks, trust and perception of fairness have been found to be critical factors driving human behaviour in human-machine interactions [40,48]. Because of the black-box nature of AI models that make it hard for users to understand why a decision is made or how the data is processed for the decision-making [7,46,44], trustworthy AI has experienced a significant surge in interest from the

⋆ Corresponding author.

research community to various application domains, especially in high stake domains which usually require testing and verification for reasonability by domain experts not only for safety but also for legal reasons [42,43,19,36]. Explanation and trust are common partners in everyday life, and extensive research has investigated the relations between AI explanations and trust from different perspectives ranging from philosophical to qualitative and quantitative dimensions [30]. For instance, Zhou et al. [45] showed that the explanation of influences of training data points on predictions significantly increased the user trust in predictions. Alam and Mueller [3] investigated the roles of explanations in AI-informed decision-making in medical diagnosis scenarios. The results show that visual and example-based explanations integrated with rationales had a significantly better impact on patient satisfaction and trust than no explanations, or with text-based rationales alone. The previous studies that empirically tested the importance of explanations to users in various fields consistently showed that explanations significantly increase user trust. Furthermore, with the advancement of AI explanation development, different explanation approaches such as local and global explanations, and feature importance-based and example-based explanations are proposed [44]. As a result, besides the explanation presentation styles such as visualisation and text [3], it is also critical to understand how different explanation approaches affect user trust in AI-informed decision-making. In addition, Edwards [10] stated that the main challenge for AI-informed decision-making is to know, whether an explanation that seems valid is accurate. This information is also needed to ensure transparency and accountability of the decision.

Besides, the data used to train machine learning models are often historical records or samples of events. They are usually not a precise description of events and conceal discrimination with sparse details, which are very difficult to identify. AI models are also imperfect abstractions of reality because of their statistical nature. All these lead to imminent imprecision and discrimination (bias) associated with AI. As a result, the investigation of fairness in AI has been becoming an indispensable component for responsible socio-technical AI systems in various decision-making tasks such as allocation of social benefits, hiring, and criminal justice [5,12]. And extensive research focuses on fairness definitions and unfairness quantification. Furthermore, human's perceived fairness (perception of fairness) plays an important role in AI-informed decision-making since AI is often used by humans and/or for human-related decision-making [35]. Duan et al. [9] argued that AI-informed decision-making can help users make better decisions. Furthermore, the authors propose that AI-informed decisions will be mostly accepted by humans, when used as a support tool. Considerable research on perceived fairness has evidenced its links to trust such as in management and organizations [25,32].

In addition, Dodge et al. [8] argued that AI explanations can also provide an effective interface for the human-in-the-loop, enabling people to identify and address fairness issues. They also demonstrated the need of providing different explanation types for different fairness issues. All these demonstrate the inter-

connection relations between explanation and fairness in AI-informed decision-making. Despite the proliferation of investigations of effects of AI explanation on trust and perception of fairness, or effects of introduced fairness on trust and perception of fairness, it is critical to understand how AI explanation and introduced fairness concurrently affect user trust since AI explanation and fairness are common partners in AI-informed decision-making. Therefore, in this work, we aim to investigate the effects of both AI explanation and introduced fairness on user trust.

Our aim in this paper is to understand user trust under both different types of AI explanations and different levels of introduced fairness. In particular, two commonly used explanation approaches of example-based explanations and feature importance-based explanations are introduced into the AI-informed decision-making pipeline under different levels of introduced fairness. We aim to discover, whether AI explanations and introduced fairness with fairness statement benefit human's trust and if so, which explanation type or fairness level benefits more than others. A user study is designed by simulating AI-informed decision-making in health insurance through manipulating AI explanations and introduced fairness levels. Statistical analyses are performed to understand effects of AI explanations and introduced fairness on trust.

2 Related Work

2.1 AI Fairness and Trust

User trust in algorithmic decision-making has been investigated from different perspectives. Zhou et al. [41,47] argued that communicating user trust benefits the evaluation of effectiveness of machine learning approaches. Kizilcec [23] found that appropriate transparency of algorithms by explanation benefited the user trust. Other empirical studies found the effects of confidence score, model accuracy and users' experience of system performance on user trust [39,43,38].

Understanding relations between fairness and trust is nontrivial in the social interaction context such as marketing and services. Roy et al. [32] showed that perceptions of fair treatment on customers play a positive role in engendering trust in the banking context. Earle and Siegrist [11] found that the issue importance affected the relations between fairness and trust. They showed that procedural fairness did not affect trust when the issue importance was high, while procedural fairness had moderate effects on trust when issue importance was low. Nikbin et al. [28] showed that perceived service fairness had a significant effect on trust, and confirmed the mediating role of satisfaction and trust in the relationship between perceived service fairness and behavioural intention.

Kasinidou et al. [21] investigated the perception of fairness in algorithmic decision-making and found that people's perception of a system's decision as 'not fair' is affecting the participants' trust in the system. Shin's investigations [33,34] showed that perception of fairness had a positive effect on trust in an algorithmic decision-making system such as recommendations. Zhou et al. [48]

got similar conclusions that introduced fairness is positively related to user trust in AI-informed decision-making, i.e. the high level of introduced fairness resulted in the high level of user trust.

These previous works motivate us to further investigate how multiple factors such as AI fairness and AI explanation together affect user trust in AI-informed decision-making.

2.2 AI Explanation and Trust

Explainability is indispensable to foster user trust in AI systems, particularly in sensible application domains. Holzinger et al. [17] introduced the concept of causability and demonstrated the importance of causability in AI explanations [18], [20]. Shin [34] used causability as an antecedent of explainability to examine their relations to trust, where causability gives the justification for what and how AI results should be explained to determine the relative importance of the properties of explainability. Shin argued that the inclusion of causability and explanations would help to increase trust and help users to assess the quality of explanations, e.g. with the Systems Causability Scale [15].

The influence of training data points on predictions is one of typical AI explanation approaches [24]. Zhou et al. [45] investigated the effects of influence on user trust and found that the presentation of influences of training data points significantly increased the user trust in predictions, but only for training data points with higher influence values under the high model performance condition. Papenmerer et al. [29] investigated the effects of model accuracy and explanation fidelity, and found that model accuracy is more important for user trust than explainability. When adding nonsensical explanations, explanations can potentially harm trust. Larasati et al. [26] investigated the effects of different styles of textual explanations on user trust in an AI medical support scenario. Four textual styles of explanations including contrastive, general, truthful, and thorough were investigated. It was found that contrastive and thorough explanations produced higher user trust scores compared to general explanation style, and truthful explanation showed no difference compared to the rest of the explanations. Wang et al. [37] compared different explanation types such as feature importance, feature contribution, nearest neighbour, and counterfactual explanation from three perspectives of improving people's understanding of the AI model, helping people recognize the model uncertainty, and supporting people's calibrated trust in the model. They highlighted the importance of selecting different AI explanation types in designing the most suitable AI methods for a specific decision-making task.

These findings confirmed the impact of explanation and its types on users trust in AI systems. In this paper, we investigate how different explanation types such as example-based and feature importance-based explanations affect user trust in AI-informed decision-making by considering the effects of AI fairness concurrently.

3 Method

3.1 Case study

This research selected the health insurance decision-making as a case study for AI-informed decision-making. The decision of the monthly payment rate is a significant step in the health insurance decision-making process. It is often based on information about the age and lifestyle of applicants. For example, a 20-year old applicant, who does neither smoke nor drink and works out frequently, is less likely to require extensive medical care. Therefore, the insurance company most likely decides to put this applicant into the lower payment class with a lower monthly rate for insurance. The insurance will increase with the age of the applicant and pre-known illnesses or previous hospital admissions. AI is used to get faster results for these decisions while enhancing customer experience since AI allows the automatic calculation of key factors and guarantees an equal procedure for every applicant [22]. This decision-making process is simulated in the study by creating fake personas with different attributes and showing their prediction of a monthly insurance rate. The simulation determines the monthly rate based on the factors of age, gender, physical activities, as well as drinking and smoking habits.

The advisory organ of the EU on GDPR, Article 29 Working Party, added a guideline [4] with detailed descriptions and requirements for profiling and automated decision-making. They also state that transparency is a fundamental requirement for the GDPR. Two explanation approaches of example-based explanation and feature importance-based explanation with fairness conditions are introduced into the decision-making process to meet requirements for AI-informed decision-making by GDPR [2] and other EU regulations and guidelines [1,13].

3.2 Explanations

This study aims to understand how AI explanations affect user trust in decision-making. Two types of explanations are investigated in the experiment:

- Example-based explanation. Example-based explanation methods select particular instances of the dataset as similar or adverse examples to explain the behaviour of AI models. Examples are commonly used as effective explanations between humans for explaining complex concepts [31]. Example-based AI explanations have been used to help users gain intuition for AI that are otherwise difficult to explain through algorithms [6]. In this study, both similar and adverse examples are introduced into tasks to investigate user responses.
- Feature importance-based explanation. Feature importance is one of the most common AI explanations [43]. It is a measure of the individual contribution of a feature to AI outcomes. For example, a feature is "important" if changing its values increases the model error, as the model relied on the feature for the prediction. A feature is "unimportant" if changing its values leaves the model error unchanged. In this study, the importance of each feature on a specific AI prediction is presented to analyse user responses.

In addition, tasks without any specific explanations (called control condition in this study) are also introduced to see if the explanation is indeed helpful or provides a better understanding of the decision-making process.

3.3 Fairness

In this study, gender is used as a protected attribute in fairness investigations. Two levels of introduced fairness are used in the study:

- Low level of fairness. At this level, the decisions are completely biased to one gender. In this study, statements such as "male and female customers having a similar personal profile did receive a different insurance rate: male customers pay € 30 more than female customers." are used to show the least fairness of the AI system.
- High level of fairness. At this level, both males and females are fairly treated in the decision-making. In this study, statements such as "male and female having a similar personal profile were treated similarly" are used to show the most fairness of the AI system.

In addition, tasks without any fairness information (called control condition in this study) are also introduced to investigate the difference of user responses in decision-making with and without the fairness information.

3.4 Task Design

Table 1: Experiment task conditions.

		Fairness		
		Control	Low	High
Explanation	Control	T	T	T
	Example-Based	T	T	T
	Feature Importance	T	T	T

According to the application scenario as described above, we investigated the decisions made by participants under both explanation and fairness conditions. In each task, AI models automatically recommended a decision based on the use case. Participants were then asked to accept or reject this decision under the presentation of different explanation and fairness conditions (3 explanation conditions by 3 fairness conditions, see Table 1). Figure 1 shows an example of the use case statement, the decision recommended by AI models, as well as the presentation of fairness and explanation conditions. After the decision-making, different questions are asked to rate users' trust in AI models. All together, each participant conducts 9 tasks (3 explanation conditions $\times$ 3 fairness conditions = 9 tasks). The orders of tasks are randomised to avoid any bias introduced. In addition, 2 training tasks are conducted by each participant before formal tasks.

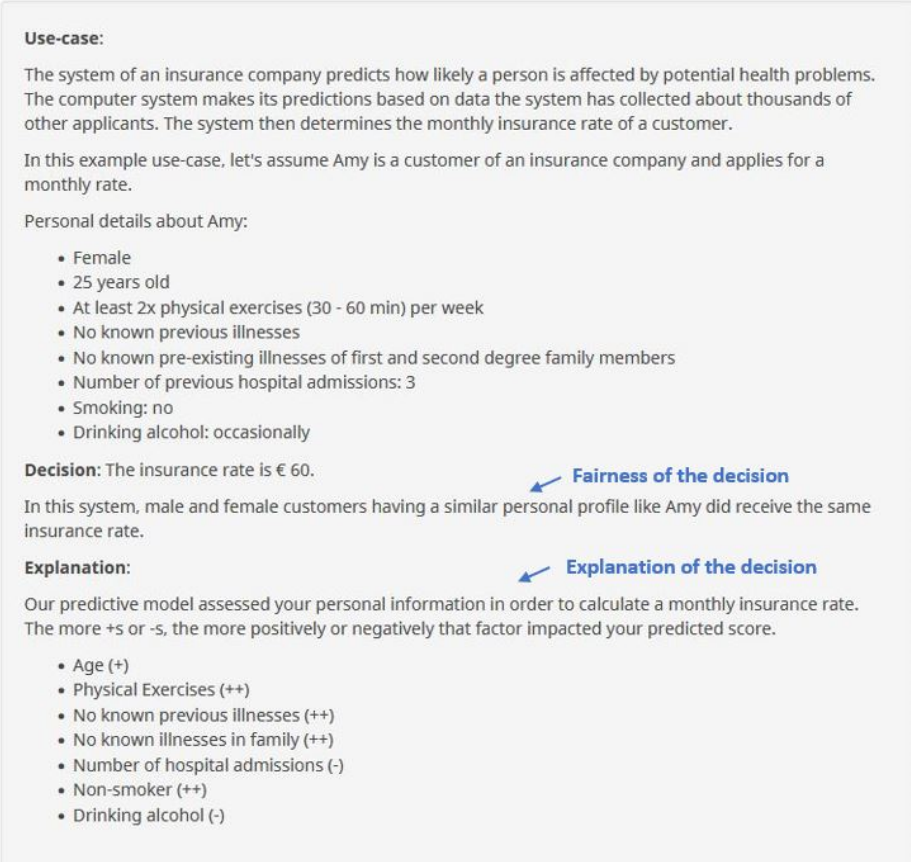

Fig. 1: An example of the experiment

3.5 Trust Scales

In this study, trust is assessed with six items using self-report scales following approaches in [27]. The scale is on a 4-point Likert-type response scale ranging from 1 (strongly disagree), 2 (disagree), 3 (agree), to 4 (strongly agree).

- I believe the system is a competent performer.
- I trust the system.
- I have confidence in the advice given by the system.
- I can depend on the system.
- I can rely on the system to behave in consistent ways.
- I can rely on the system to do its best every time I take its advice.

3.6 Experiment Setup

Due to social distancing restrictions and lockdown policies during the COVID-19 pandemic, this experiment was implemented and deployed on the cloud server

online. The deployed application link was then shared with participants through emails and social networks to invite them to participate in the experiment. Figure 1 shows the example of a task conducted in the experiment.

3.7 Participants and Data Collection

In this study, 25 participants were recruited to conduct experimental tasks via various means of communications such as emails and social media posts who were university students with an average age of 26 and 10 of them were females. After each task was displayed on the screen, the participants were asked to answer ten questions based on the task on trust and satisfaction in the AI-informed decision-making.

4 Results

Since two independent factors of explanation and fairness were introduced to investigate their effects on user trust in this study, two-way ANOVA tests were first conducted to examine whether there were interactions between explanation and introduced fairness on trust. We then performed one-way ANOVA tests, followed by a post-hoc analysis using t-tests (with a Bonferroni correction) to analyze differences in participant responses of trust under different conditions.

Before statistical analysis, trust values were normalised with respect to each subject to minimize individual differences in rating behavior (see Equation 1):

$$V_i^N = (V_i - V_i^{min})/(V_i^{max} - V_i^{min}) \tag{1}$$

where V_i and V_i^N are the original and normalised trust ratings respectively from the participant i, V_i^{min} and V_i^{max} are the minimum and maximum of trust values respectively from the participant i in all of his/her tasks.

4.1 Effects of Fairness on Trust

Figure 2 shows normalised trust values over the introduced fairness conditions. A one-way ANOVA test was performed to compare the effect of introduced fairness on user trust. The one-way ANOVA test found that there were statistically significant differences in user trust among three introduced fairness conditions $F(2, 222) = 8.446, p < .000$. A further post-hoc comparison with t-tests (with a Bonferroni correction under a significance level set at $\alpha < .017$) was conducted to find pair-wise differences in user trust between three fairness conditions. The adjusted significance alpha level of .017 was calculated by dividing the original alpha of .05 by 3, based on the fact that we had three fairness conditions. It was found that participants had a statistically significant high level of trust under the high level of fairness compared to the low fairness condition ($t = 4.185, p < .000$). Moreover, it was found that participants also had a statistically significant higher level of trust under the control condition (no fairness information presented)

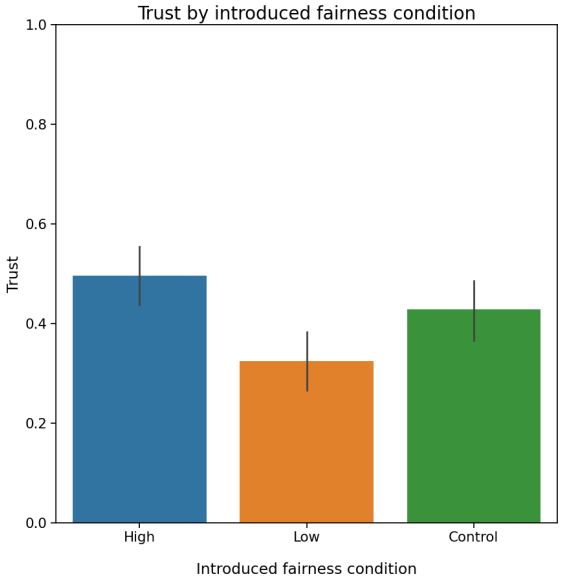

Fig. 2: User trust under introduced fairness conditions

than that under the low fairness condition ($t = 2.433, p < .016$). However, there was not a statistically significant difference found in user trust between the introduced high fairness condition and control condition ($t = 1.602, p < .111$).

These findings imply that the introduced fairness condition did affect user trust in AI-information decision-making only under the low fairness condition, where introduced fairness decreased user trust in AI-informed decision-making.

4.2 Effects of Explanation on Trust

Figure 3 shows normalised trust values over various explanation conditions. A one-way ANOVA test revealed statistically significant differences in user trust under different explanation types $F(2, 222) = 11.226, p < .000$. Then post-hoc tests with the aforementioned Bonferroni correction were conducted. It was found that participants had statistically significant lower level of trust under the control condition (no explanation presented) than that under the feature importance-based explanation ($t = 4.645, p < .000$) and example-based explanation ($t = 2.455, p < .015$) respectively. There was not a significant difference in user trust between feature importance-based explanation and example-based explanation ($t = 2.329, p < .021$).

The results showed that explanations did help users increase their trust significantly in AI-informed decision-making, but different explanation types did not show differences in affecting user trust.

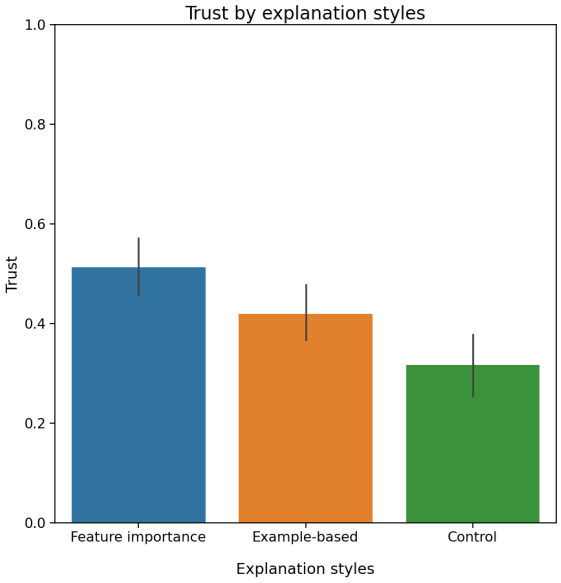

Fig. 3: User trust under explanation types

4.3 Effects of Fairness and Explanation on Trust

This subsection further analyses the effects of fairness on trust under different given explanation types, and the effects of explanation on trust under different given fairness levels.

Effects of fairness on trust under example-based explanations Figure 4 shows normalised trust values over various fairness conditions under the example-based explanation condition. A one-way ANOVA test was conducted to compare the effect of introduced fairness on user trust under the example-based explanation. The test found a statistically significant difference in trust between introduced fairness levels, $F(2, 72) = 8.146, p < .001$. Further post-hoc t-tests (with Bonferroni correction) were then conducted to find differences in trust among different fairness levels. Participants showed a significant higher trust level under high introduced fairness than that under the low introduced fairness level ($t = 3.887, p < .000$). Moreover, user trust was significantly higher under the control condition (no fairness information presented) than that under the low introduced fairness level ($t = 3.266, p < .002$). However, there was not a significant difference in trust between the high introduced fairness and the control condition ($t = .436, p < .665$).

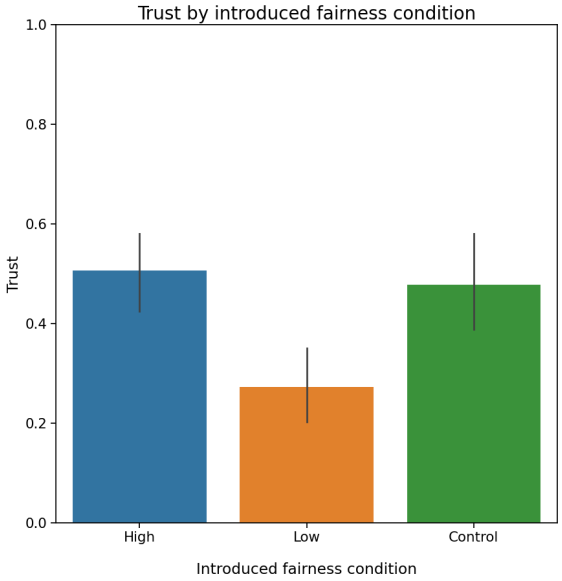

Fig. 4: Effects of fairness on user trust under the example-based explanation

The results showed that under the example-based explanation condition, the low level of fairness statement significantly decreased the user trust in decision-making but the high level of fairness statement did not affect user trust.

Effects of fairness on trust under feature importance-based explanations Figure 5 shows the normalized trust levels for introduced fairness levels under feature importance-based explanation. A one-way ANOVA test found no significant differences in trust in different introduced fairness levels under the feature importance-based explanation, $F(2, 72) = 2.353, p < .102$.

From the results, we can see that under the feature importance-based explanation condition, no fairness information seems to influence the user's trust.

Effects of explanation on trust under low level introduced fairness Figure 6 shows the normalized trust values with different explanation types under low-level introduced fairness. A one-way ANOVA test found statistical significant differences in trust among explanation types under the low level of introduced fairness, $F(2, 72) = 3.307, p < .042$. The further t-test found that participants showed no significant higher level of trust under feature importance-based explanation than that under the control condition (no explanation presented) ($t = 2.248, p < .046$). Moreover, there were neither significant differences

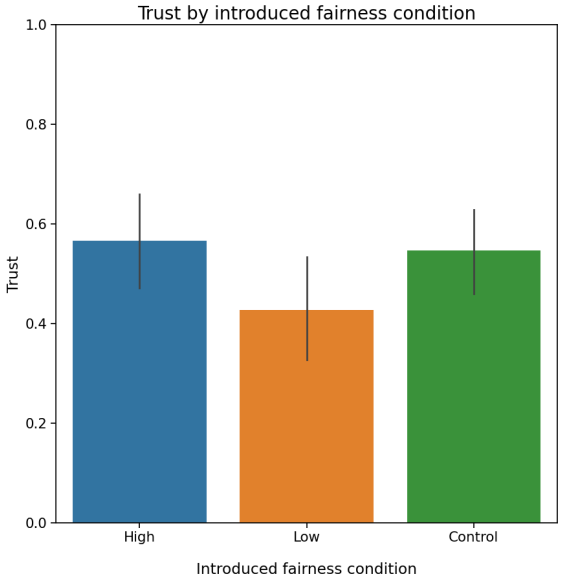

Fig. 5: Effects of fairness under feature importance-based explanations on user trust

found in user trust between the control condition and example-based explanation ($t = .035, p < .972$), nor between the two explanation types ($t = 2.296, p < .026$).

Therefore, we can say that under the low level of introduced fairness, neither explanation type did significantly increase user's trust in the decision-making process.

Effects of explanation on trust under high level of introduced fairness Figure 7 shows the normalised trust values in different explanation types under the high level of introduced fairness. A one-way ANOVA test revealed no statistical significant differences in user trust among explanation types under the high level of introduced fairness, $F(2, 72) = 2.369, p < .101$.

The explanation type under the high level of introduced fairness had no influence on user's trust.

5 Discussion

Explanation and fairness, along with robustness [16,14], are among the indispensable components of AI-based decision making for trustworthy AI. AI-informed decision-making and automated aids have been becoming much popular with the advent of new AI-based intelligent applications. Therefore, we opted to study the

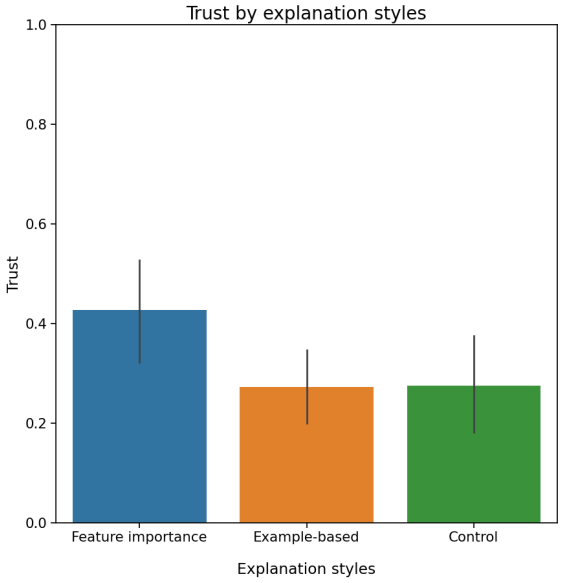

Fig. 6: Effects of explanation on user trust under low introduced fairness level

effects of both AI explanations and fairness on human-AI trust in a specialised AI-informed decision-making scenario.

The study found that the fairness statement in the scenario did affect user trust in AI-information decision-making only under the low level of fairness condition, where the low-level fairness statement decreased user trust in AI-informed decision-making. However, the addition of explanations helped users increase their trust significantly in AI-informed decision-making, and different explanation types did not show differences in affecting user trust. We then drilled down into the effects on trust under specific conditions. From the explanation's perspective, it was found that under the example-based explanation condition, the low level of fairness statement significantly decreased the user trust in decision-making but the high level of fairness statement did not affect user trust. Nevertheless, the level of fairness under the feature importance-based explanation condition did not show any impact on the user trust. Furthermore, from the introduced fairness' perspective, it revealed that under the low level of introduced fairness, neither explanation type significantly increased user trust in decision-making. The high level of introduced fairness, on the other hand, showed no effects at all on user trust. It also implies that the introduced fairness levels did not affect user trust too much.

These findings suggest that the deployment of AI explanation and fairness statements in real-world applications is complex: we need to not only consider

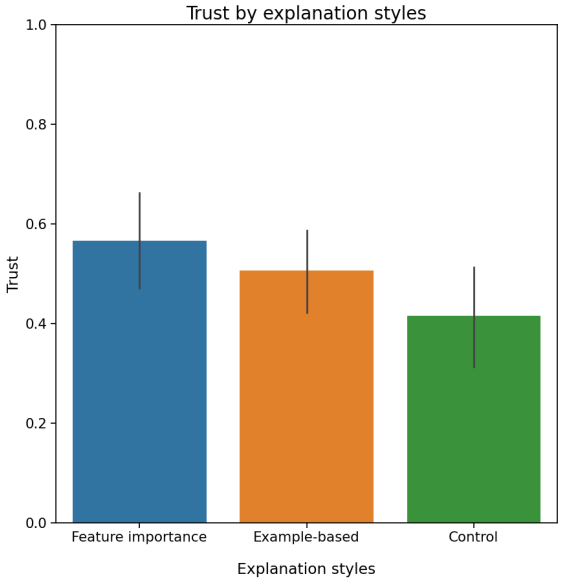

Fig. 7: Effects of explanation on trust under high introduced fairness level

explanation types, but also levels of introduced fairness. In order to maximise user trust in AI-informed decision-making, the explanation types and the level of fairness statement can be adjusted in the user interface of intelligent applications.

6 Conclusion and Future Work

This paper investigated the effects of introduced fairness and explanation on user trust in AI-informed decision-making. A user study by simulating AI-informed decision-making through manipulating AI explanations and fairness levels found that the introduced fairness affected user trust in AI-informed decision-making only under the low level of fairness condition. It was also found that the AI explanations increased user trust in AI-informed decision-making, and different explanation types did not show differences in affecting user trust. The future work of this study will focus on the effects of explanation and introduced fairness on perception of fairness as well as accountability of AI.

Acknowledgements

This work does not raise any ethical issues. Parts of this work have been funded by the Austrian Science Fund (FWF), Project: P-32554 explainable Artificial

Intelligence; and by the Australian UTS STEM-HASS Strategic Research Fund 2021.

References

1. European parliament resolution of 20 october 2020 with recommendations to the commission on a framework of ethical aspects of artificial intelligence, robotics and related technologies, 2020/2012(inl). https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52020IP0275, (19.01.2022)
2. Regulation (eu) 2016/679 of the european parliament and of the council of 27 april 2016 on the protection of natural persons with regard to the processing of personal data and on the free movement of such data, and repealing directive 95/46/ec (general data protection regulation). https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:02016R0679-20160504 (2016)
3. Alam, L., Mueller, S.: Examining the effect of explanation on satisfaction and trust in AI diagnostic systems. BMC Medical Informatics and Decision Making **21**(1), 178 (June 2021). https://doi.org/https://doi.org/10.1186/s12911-021-01542-6
4. Article 29 Working Party: Guidelines on automated individual decision-making and profiling for the purposes of regulation 2016/679. https://ec.europa.eu/newsroom/article29/items/612053/en, (19.01.2022)
5. Berk, R., Heidari, H., Jabbari, S., Kearns, M., Roth, A.: Fairness in criminal justice risk assessments: The state of the art. Sociological Methods & Research p. 0049124118782533 (2018)
6. Cai, C.J., Jongejan, J., Holbrook, J.: The effects of example-based explanations in a machine learning interface. In: Proceedings of the 24th International Conference on Intelligent User Interfaces. p. 258–262. IUI '19 (2019). https://doi.org/10.1145/3301275.3302289
7. Castelvecchi, D.: Can we open the black box of AI? Nature News **538**(7623), 20 (October 2016)
8. Dodge, J., Liao, Q.V., Zhang, Y., Bellamy, R.K.E., Dugan, C.: Explaining models: An empirical study of how explanations impact fairness judgment. In: Proceedings of the 24th International Conference on Intelligent User Interfaces. p. 275–285. IUI '19 (2019)
9. Duan, Y., Edwards, J.S., Dwivedi, Y.K.: Artificial intelligence for decision making in the era of big data – evolution, challenges and research agenda. International Journal of Information Management **48**, 63–71 (2019). https://doi.org/10.1016/j.ijinfomgt.2019.01.021
10. Dwivedi, Y.K., Hughes, L., Ismagilova, E., Aarts, G., Coombs, C., Crick, T., Duan, Y., Dwivedi, R., Edwards, J., Eirug, A., Galanos, V., Ilavarasan, P.V., Janssen, M., Jones, P., Kar, A.K., Kizgin, H., Kronemann, B., Lal, B., Lucini, B., Medaglia, R., Le Meunier-FitzHugh, K., Le Meunier-FitzHugh, L.C., Misra, S., Mogaji, E., Sharma, S.K., Singh, J.B., Raghavan, V., Raman, R., Rana, N.P., Samothrakis, S., Spencer, J., Tamilmani, K., Tubadji, A., Walton, P., Williams, M.D.: Artificial intelligence (ai): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. International Journal of Information Management **57** (2021). https://doi.org/10.1016/j.ijinfomgt.2019.08.002
11. Earle, T.C., Siegrist, M.: On the relation between trust and fairness in environmental risk management. Risk Analysis **28**(5), 1395–1414 (October 2008)

12. Feldman, M., Friedler, S.A., Moeller, J., Scheidegger, C., Venkatasubramanian, S.: Certifying and removing disparate impact. In: Proceedings of KDD2015. pp. 259–268 (2015)
13. High-Level Export Group on Artificial Intelligence: Ethics guidelines for trustworthy ai. `https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai`, (19.01.2022)
14. Holzinger, A.: The next frontier: Ai we can really trust. In: Kamp, M. (ed.) Proceedings of the ECML PKDD 2021, CCIS 1524, pp. 1–14. Springer Nature (2021)
15. Holzinger, A., Carrington, A., Mueller, H.: Measuring the quality of explanations: The system causability scale (scs). comparing human and machine explanations. KI - Kuenstliche Intelligenz (German Journal of Artificial intelligence), Special Issue on Interactive Machine Learning, Edited by Kristian Kersting, TU Darmstadt **34**(2), 193–198 (2020). https://doi.org/10.1007/s13218-020-00636-z
16. Holzinger, A., Dehmer, M., Emmert-Streib, F., Cucchiara, R., Augenstein, I., Del Ser, J., Samek, W., Jurisica, I., Díaz-Rodríguez, N.: Information fusion as an integrative cross-cutting enabler to achieve robust, explainable, and trustworthy medical artificial intelligence. Information Fusion **79**(3), 263–278 (2022). https://doi.org/10.1016/j.inffus.2021.10.007
17. Holzinger, A., Langs, G., Denk, H., Zatloukal, K., Müller, H.: Causability and explainability of artificial intelligence in medicine. Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery **9**(4), 1–13 (2019). https://doi.org/10.1002/widm.1312
18. Holzinger, A., Malle, B., Saranti, A., Pfeifer, B.: Towards multi-modal causability with graph neural networks enabling information fusion for explainable ai. Information Fusion **71**(7), 28–37 (2021). https://doi.org/10.1016/j.inffus.2021.01.008
19. Holzinger, K., Mak, K., Kieseberg, P., Holzinger, A.: Can we trust machine learning results? artificial intelligence in safety-critical decision support. ERCIM News **112**(1), 42–43 (2018)
20. Hudec, M., Minarikova, E., Mesiar, R., Saranti, A., Holzinger, A.: Classification by ordinal sums of conjunctive and disjunctive functions for explainable ai and interpretable machine learning solutions. Knowledge Based Systems **220**, 106916 (2021). https://doi.org/10.1016/j.knosys.2021.106916
21. Kasinidou, M., Kleanthous, S., Barlas, P., Otterbacher, J.: 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. p. 690–700. FAccT '21 (2021). https://doi.org/10.1145/3442188.3445931
22. Kelley, K.H., Fontanetta, L.M., Heintzman, M., Pereira, N.: Artificial intelligence: Implications for social inflation and insurance. Risk Management and Insurance Review **21**(3), 373–387 (2018). https://doi.org/10.1111/rmir.12111
23. Kizilcec, R.F.: 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. CHI '16, Association for Computing Machinery (2016). https://doi.org/10.1145/2858036.2858402
24. Koh, P.W., Liang, P.: Understanding black-box predictions via influence functions. In: Proceedings of ICML 2017. pp. 1885–1894 (july 2017)
25. Komodromos, M.: Employees' perceptions of trust, fairness, and the management of change in three private universities in cyprus. Journal of Human Resources Management and Labor Studies **2**(2), 35–54 (July 2014)

26. Larasati, R., Liddo, A.D., Motta, E.: The effect of explanation styles on user's trust. In: Proceedings of the Workshop on Explainable Smart Systems for Algorithmic Transparency in Emerging Technologies co-located with IUI 2020,. pp. 1–6 (2020)
27. Merritt, S.M., Heimbaugh, H., LaChapell, J., Lee, D.: I trust it, but i don't know why: Effects of implicit attitudes toward automation on trust in an automated system. Human Factors **55**(3), 520–534 (2013)
28. Nikbin, D., Ismail, I., Marimuthu, M., Abu-Jarad, I.: The effects of perceived service fairness on satisfaction, trust, and behavioural intentions. Singapore Management Review **33**(2), 58–73 (2011)
29. Papenmeier, A., Englebienne, G., Seifert, C.: How model accuracy and explanation fidelity influence user trust. In: IJCAI 2019 Workshop on Explainable Artificial Intelligence (xAI). pp. 1–7 (August 2019)
30. Pieters, W.: Explanation and trust: what to tell the user in security and AI? Ethics and Information Technology **13**(1), 53–64 (2011). https://doi.org/https://doi.org/10.1007/s10676-010-9253-3
31. Renkl, A., Hilbert, T., Schworm, S.: Example-based learning in heuristic domains: A cognitive load theory account. Educational Psychology Review **21**, 67–78 (03 2009). https://doi.org/10.1007/s10648-008-9093-4
32. Roy, S.K., Devlin, J.F., Sekhon, H.: The impact of fairness on trustworthiness and trust in banking. Journal of Marketing Management **31**(9-10), 996–1017 (2015)
33. Shin, D.: User perceptions of algorithmic decisions in the personalized ai system:perceptual evaluation of fairness, accountability, transparency, and explainability. Journal of Broadcasting and Electronic Media **64**(4), 541–565 (2020). https://doi.org/10.1080/08838151.2020.1843357
34. Shin, D.: The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable ai. International Journal of Human-Computer Studies **146**, 102551 (2021). https://doi.org/10.1016/j.ijhcs.2020.102551
35. Starke, C., Baleis, J., Keller, B., Marcinkowski, F.: Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature (2021)
36. Stoeger, K., Schneeberger, D., Kieseberg, P., Holzinger, A.: Legal aspects of data cleansing in medical ai. Computer Law and Security Review **42**, 105587 (2021). https://doi.org/10.1016/j.clsr.2021.105587
37. Wang, X., Yin, M.: Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making. In: Proceedings of 26th International Conference on Intelligent User Interfaces. p. 318–328. ACM (2021)
38. Yin, M., Vaughan, J.W., Wallach, H.: Does stated accuracy affect trust in machine learning algorithms? In: Proceedings of ICML2018 Workshop on Human Interpretability in Machine Learning (WHI 2018). pp. 1–2 (7 2018)
39. 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. p. 295–305. FAT* '20 (2020)
40. Zhou, J., Arshad, S.Z., Luo, S., Chen, F.: Effects of uncertainty and cognitive load on user trust in predictive decision making. In: Bernhaupt, R., Dalvi, G., Joshi, A., K. Balkrishan, D., O'Neill, J., Winckler, M. (eds.) Human-Computer Interaction – INTERACT 2017. pp. 23–39. Springer, Cham (2017)
41. Zhou, J., Bridon, C., Chen, F., Khawaji, A., Wang, Y.: Be informed and be involved: Effects of uncertainty and correlation on user's confidence in decision making. In: Proceedings of the 33rd Annual ACM Conference Extended Abstracts on Human Factors in Computing Systems. pp. 923—928. CHI EA '15, Association for Computing Machinery (2015). https://doi.org/10.1145/2702613.2732769

42. Zhou, J., Chen, F.: 2d transparency space—bring domain users and machine learning experts together. In: Human and Machine Learning: Visible, Explainable, Trustworthy and Transparent, pp. 3–19. Human–Computer Interaction Series, Springer International Publishing (2018)
43. Zhou, J., Chen, F. (eds.): Human and Machine Learning: Visible, Explainable, Trustworthy and Transparent. Human–Computer Interaction Series, Springer International Publishing (2018)
44. Zhou, J., Gandomi, A.H., Chen, F., Holzinger, A.: Evaluating the quality of machine learning explanations: A survey on methods and metrics. Electronics **10**(5) (2021)
45. Zhou, J., Hu, H., Li, Z., Yu, K., Chen, F.: Physiological indicators for user trust in machine learning with influence enhanced fact-checking. In: Machine Learning and Knowledge Extraction. pp. 94–113 (2019)
46. Zhou, J., Khawaja, M.A., Li, Z., Sun, J., Wang, Y., Chen, F.: Making machine learning useable by revealing internal states update — a transparent approach. International Journal of Computational Science and Engineering **13**(4), 378–389 (2016)
47. Zhou, J., Sun, J., Chen, F., Wang, Y., Taib, R., Khawaji, A., Li, Z.: Measurable decision making with gsr and pupillary analysis for intelligent user interface. ACM Transactions on Computer-Human Interaction **21**(6), 1–23 (01 2015). https://doi.org/10.1145/2687924
48. Zhou, J., Verma, S., Mittal, M., Chen, F.: Understanding relations between perception of fairness and trust in algorithmic decision making. In: Proceedings of the International Conference on Behavioral and Social Computing (BESC 2021). pp. 1–5 (October 2021)