machine learning & knowledge extraction

MDPI

Article

Fairness and Explanation in AI-Informed Decision Making

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

[1] Medical Informatics, Statistics and Documentation, Medical University Graz, 8036 Graz, Austria; alessa.angerschmid@human-centered.ai (A.A.); andreas.holzinger@human-centered.ai (A.H.)
[2] Human-Centered AI Lab, University of Natural Resources and Life Sciences, 1190 Vienna, Austria
[3] Human-Centered AI Lab, University of Technology Sydney, Sydney, NSW 2007, Australia; fang.chen@uts.edu.au
[4] Doctoral School of Computer Science, Graz University of Technology, 8010 Graz, Austria; kevin.theuermann@egiz.gv.at
[5] xAI Lab, Alberta Machine Intelligence Institute, University of Alberta, Edmonton, AB T5J 3B1, Canada
[*] Correspondence: jianlong.zhou@uts.edu.au

Abstract: AI-assisted decision-making that impacts individuals raises critical questions about transparency and fairness in artificial intelligence (AI). Much research has highlighted the reciprocal relationships between the transparency/explanation and fairness in AI-assisted decision-making. Thus, considering their impact on user trust or perceived fairness simultaneously benefits responsible use of socio-technical AI systems, but currently receives little attention. In this paper, we investigate the effects of AI explanations and fairness on human-AI trust and perceived fairness, respectively, in specific AI-based decision-making scenarios. A user study simulating AI-assisted decision-making in two health insurance and medical treatment decision-making scenarios provided important insights. Due to the global pandemic and restrictions thereof, the user studies were conducted as online surveys. From the participant's trust perspective, fairness was found to affect user trust only under the condition of a low fairness level, with the low fairness level reducing user trust. However, adding explanations helped users increase their trust in AI-assisted decision-making. From the perspective of perceived fairness, our work found that low levels of introduced fairness decreased users' perceptions of fairness, while high levels of introduced fairness increased users' perceptions of fairness. The addition of explanations definitely increased the perception of fairness. Furthermore, we found that application scenarios influenced trust and perceptions of fairness. 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 and the degree of fairness introduced, but also the scenarios in which AI-assisted decision-making is used.

Keywords: AI explanation; AI fairness; trust; perception of fairness; AI ethics

check for updates

Citation: Angerschmid, A.; Zhou, J.; Theuermann, K.; Chen, F.; Holzinger, A. Fairness and Explanation in AI-Informed Decision Making. *Mach. Learn. Knowl. Extr.* **2022**, *4*, 556–579. https://doi.org/10.3390/make4020026

Academic Editor: Federico Cabitza

Received: 14 May 2022
Accepted: 14 June 2022
Published: 16 June 2022

1. Introduction

Artificial Intelligence (AI) informed decision-making is claimed to lead to faster and better decision outcomes. It has been increasingly used in our society from decision-making of daily lives such as recommending movies and books to making more critical decisions such as medical diagnosis, credit risk prediction, and shortlisting talents in recruitment. In 2020, the EU proposed the European approach to excellence and trust with their White Paper on AI [1]. They stated that AI will change lives by improving not only healthcare but also increasing the efficiency of farming and contributing to climate change mitigation. Thus, their approach is to improve lives, while respecting rights. 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 [2,3]. The black-box nature of AI models makes it hard for users to understand why a decision is made or how the data are processed for the decision-making [4–6]. Thus, trustworthy AI has experienced a significant surge in interest from the research community in 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 [7–11].

1.1. AI Explanation

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 [12]. For instance, Zhou et al. [13] showed that the explanation of influences of training data points on predictions significantly increased the user trust in predictions. Alam and Mueller [14] 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 research, different explanation approaches such as local and global explanations, as well as feature importance-based and example-based explanations are proposed [6]. As a result, besides the explanation presentation styles such as visualisation and text [14,15], it is also critical to understand how different explanation approaches affect user trust in AI-informed decision-making. In addition, Edwards [16] 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.

1.2. AI Fairness

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 [17,18]. In addition, 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 [19].

Duan et al. [20] argue 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. Thus, it is crucial to consider the human perception of AI in general, and to what extent users would be willing to use such systems [21]. Considerable research on perceived fairness has evidenced its links to trust such as in management and organizations [22,23].

1.3. Aims

Dodge et al. [24] 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 of these demonstrate the inter-connection relations between explanation and fairness in AI-informed decision-making. Despite the proliferation of investigations of the effects of AI explanation on trust and perception of fairness, or effects of introduced fairness on trust and perception of fairness, it is crucial to understand how AI explanation and introduced fairness concurrently affect user trust and perception of fairness since AI explanation and fairness are common partners in AI-informed decision-making.

Our aim in this paper is to understand user trust under both different types of AI explanations and different levels of introduced fairness. This paper also aims to understand the perception of fairness by humans under both AI explanations and introduced fairness (see Figure 1). In particular, two commonly used explanation approaches 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 benefit human's trust and perceived fairness respectively and, if so, which explanation type or fairness level benefits more than others. An online user study with 25 participants is conducted by simulating AI-informed decision-making in two scenarios of health insurance decision-making and medical treatment decision-making through manipulating AI explanations and introduced fairness levels. Statistical analyses are performed to understand the effects of AI explanations and introduced fairness on trust and perception of fairness.

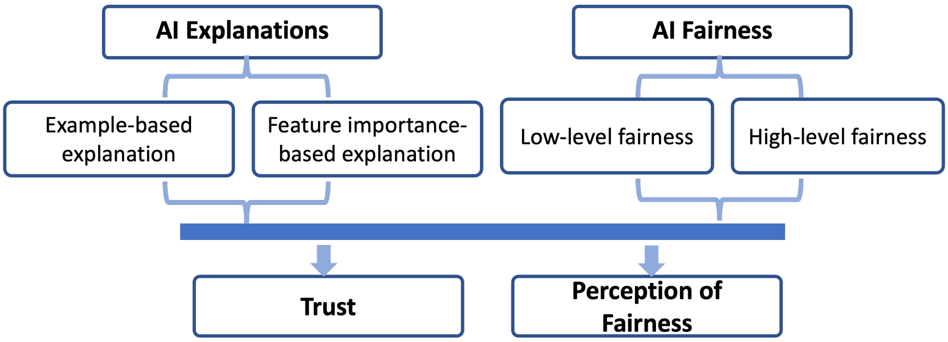


Figure 1. Effects of AI explanation and fairness on user trust and perception of fairness.

In summary, our study demonstrates important findings:

- The introduced fairness and explanations affect user trust in AI-informed decision-making. Different fairness levels and explanation types affect user trust differently.
- Similarly, the introduced fairness and explanations affect human's perceived fairness in AI-informed decision-making. However, the effects of fairness levels and explanation types on the perception of fairness are different from their effects on user trust.
- The application scenarios affect user responses of trust and perception of fairness under different explanation types and introduced fairness levels in AI-informed decision-making.

2. Related Work

2.1. Perception of Fairness

Current machine learning outlines fairness in the context of different protected attributes (race, sex, culture, etc.) receiving equal treatments by algorithms [25–27]. Definitions of fairness are formalised ranging from statistical bias, group fairness, and individual fairness, to process fairness, and others. Various metrics are proposed to quantify the unfairness (bias) of algorithms [28–30].

The research on the perception of fairness can be categorised into the following dimensions [19]: First, algorithmic factors study how the technical design of an AI system affects people's fairness perceptions. For example, Lee et al. [31,32] investigated people's perception of fairness regarding the allocation of resources based on equality, equity, or efficiency. They found that people had many variations in the preferences for the three fairness metrics (equality, equity, efficiency) impacted by the decision. Dodge et al. [24] found that people's perception of fairness is evaluated primarily based on features that are used and not used in the model, algorithm errors, and errors or flaws in input data. Secondly, human factors investigate how human-related information affects the perception of fairness. For example, Helberger et al. [33] found that education and age affected both perceptions of algorithmic fairness and people's reasons for the perception of AI fairness. Thirdly, comparative effects investigate how individuals react in fairness to humans compared to algorithmic decision-makers. For example, Helberger et al. [33] found that people believe that AI makes fairer decisions than human decision-makers. Some studies found the opposite results in the criminal justice system [34]. Fourthly, the consequence of the perception of fairness aims to investigate the impact of the perception of fairness on AI-informed decision-making. For example, Shin and Park [35] investigated the effects of perception of fairness on satisfaction and found that people's perception of fairness has a positive impact on satisfaction with algorithms. Moreover, Shin et al. [36] argued that the algorithmic experience is inherently related to the perception of fairness, transparency and the underlying trust. Zhou et al. [3] investigated the relationship between induced algorithmic fairness and its perception in humans. It was found that introduced fairness is positively related to the perception of fairness, i.e., the high level of introduced fairness resulted in a high level of perception of fairness.

People's perception of fairness has close relations with AI explanations. Shin [37] looked at explanations for an algorithmic decision as a critical factor of perceived fairness, and it was found that explanations for an algorithmic decision significantly increased people's perception of fairness in an AI-based news recommender system. Dodge et al. [24] found that case-based and sensitivity-based explanations effectively exposed fairness discrepancies between different cases, while demographic explanations (offering information about the classification for individuals in the same demographic categories) and input influence (presenting all input features and their impact in the classification) enhanced fairness perception by increasing people's confidence in understanding the model. Binns et al. [38] examined people's perception of fairness in AI-informed decision-making under four explanation types (input influence, sensitivity, case-based, and demographic). It was found that people did consider fairness in AI-informed decision-making. However, depending on when and how explanations were presented, explanations had different effects on people's perception of fairness: (1) when multiple explanation types were presented, case-based explanations (presenting a case from the model's training data which is most similar to the decision being explained) had a negative influence on the perception of fairness. (2) When only one explanation type was presented to people, the explanation did not show effects on people's perception of fairness.

Besides explanation types, mathematical fairness inherently introduced by AI models and/or data (also refers to introduced fairness in this paper) can affect people's perceived fairness [3]. However, little work is found in understanding whether different explanation types and introduced fairness together affect people's perception of fairness. This paper investigates the effects of two types of explanations (feature importance-based and example-based) on the perception of fairness and understands how the two types of explanations together with different introduced fairness levels affect people's perception of fairness.

2.2. AI Fairness and Trust

User trust in algorithmic decision-making has been investigated from different perspectives. Zhou et al. [39,40] argued that communicating user trust benefits the evaluation of the effectiveness of machine learning approaches. Kizilcec [41] 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 [8,42,43].

Understanding relations between fairness and trust is nontrivial in the social interaction context such as marketing and services. Roy et al. [23] showed that perceptions of fair treatment of customers play a positive role in engendering trust in the banking context. Earle and Siegrist [44] found that the issue's 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. [45] 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. [46] investigated the perception of fairness in algorithmic decision-making and found that people's perception of a system's decision as 'not fair' affects the participants' trust in the system. Shin's investigations [27,37] showed that perception of fairness had a positive effect on trust in an algorithmic decision-making system such as recommendations. Zhou et al. [3] obtained similar conclusions that introduced fairness is positively related to user trust in AI-informed decision-making.

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.

2.3. AI Explanation and Trust

Explainability is indispensable to foster user trust in AI systems, particularly in sensible application domains. Holzinger et al. [47] introduced the concept of causability and demonstrated the importance of causability in AI explanations [48,49]. Shin [37] 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 [50].

The influence of training data points on predictions is one of the typical AI explanation approaches [51]. Zhou et al. [13] 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. [52] 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. [53] 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 the general explanation style, and truthful explanations showed no difference compared to the rest of the explanations. Wang et al. [54] 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. Materials and Methods

3.1. Scenarios

This research selected two application contexts for AI-informed decision-making: health insurance decision-making and medical treatment decision-making.

3.1.1. Decision-Making of Health Insurance Payment

The decision on 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 neither smokes nor drinks 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 obtaining 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 [55]. 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 [56] 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 [57] and other EU regulations and guidelines [58,59].

3.1.2. Decision-Making of Medical Treatment

The decision-making of medical treatment belongs to high-stake decisions and needs physicians to comprehensively evaluate the relevant data concerning the current status of the patient for the corresponding treatment, either alone or together with the patient [60]. Physicians also need to inform the patient about the relevant factors which influence the treatment decision. AI can improve this process by providing a predicted decision to the physician and the patient, which could potentially reduce the extensive human errors in medical practices [61]. In this study, this AI-informed decision-making process is simulated by showing predicted medical treatment decisions with different simulated patients based on factors of gender, age, past medical expenses and pre-existing illnesses of first and second-degree family members. Although past medical expenses do not seem like a relevant factor, Obermeyer et al. [62] showed that this was used as a proxy for the seriousness of a medical condition. The decisions from the simulated AI include two options: immediate treatment or waiting for the next free appointment. These help to prioritize and order the patients according to their urgency, similar to the algorithm proposed by Pourhomayoun and Shakibi [63].

Two types of AI explanations and two levels of introduced fairness are introduced into this decision scenario and manipulated to understand their effects on trust and perception of fairness, respectively.

3.2. Explanations

This study aims to understand how AI explanations affect the perception of fairness and 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 to explain complex concepts [64]. Example-based AI explanations have been used to help users gain an intuition for AI that are otherwise difficult to explain through algorithms [65]. 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 [8]. 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. Hence, fairness represents the topic of gender discrimination in this study. Two levels of introduced fairness are used in the study:

- Low fairness. At this level, the decisions are completely biased toward 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 Euros more than female customers." are used to show the least fairness of the AI system.
- High 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 in user responses in decision-making with and without the fairness information.

3.4. Task Design and Experiment Setup

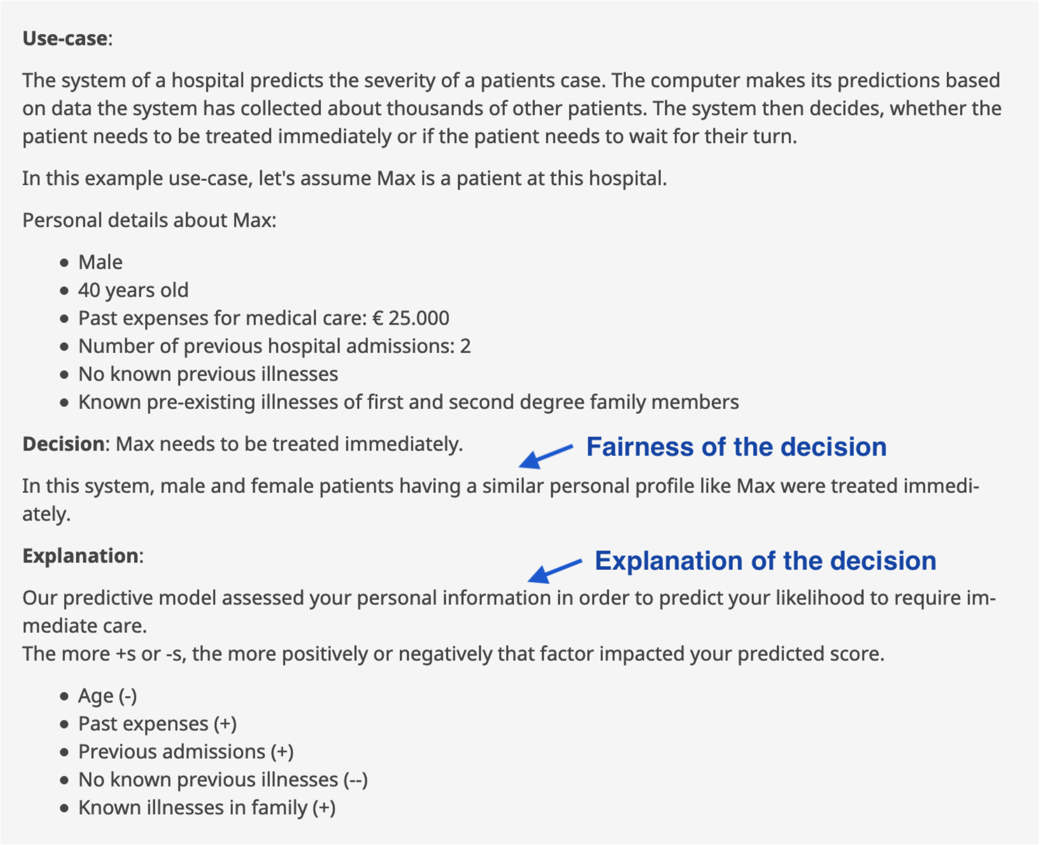
According to application scenarios as described above, we investigated the decisions made by participants under both explanation and fairness conditions (3 explanation conditions by 3 fairness conditions, see Table 1). All together, each participant conducts 18 tasks (3 explanation conditions × 3 fairness conditions × 2 scenarios = 18 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.

Table 1. Experiment task conditions.

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

During the experiment, each participant firstly signed the consent form and agreed to conduct the experiment. Two training tasks were then conducted by the participant to become familiar with the experiment before formal tasks. In each formal task, an application scenario was firstly presented to the participants. AI models then automatically recommended a decision based on the use case. The participant was asked to accept or reject this decision under the presentation of different explanation and fairness conditions. Figure 2 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 were asked to rate users' trust in AI models and perception of fairness in decision-making.

Due to social distancing restrictions and lockdown policies during the COVID-19 pandemic, this experiment was implemented and deployed on a cloud server online in Medical University Graz, Austria. The deployed application link was then shared with participants through emails and social networks to invite them to participate in the experiment.



Use-case:

The system of a hospital predicts the severity of a patients case. The computer makes its predictions based on data the system has collected about thousands of other patients. The system then decides, whether the patient needs to be treated immediately or if the patient needs to wait for their turn.

In this example use-case, let's assume Max is a patient at this hospital.

Personal details about Max:

- Male
- 40 years old
- Past expenses for medical care: € 25.000
- Number of previous hospital admissions: 2
- No known previous illnesses
- Known pre-existing illnesses of first and second degree family members

Decision: Max needs to be treated immediately.

In this system, male and female patients having a similar personal profile like Max were treated immediately.

Explanation:

Our predictive model assessed your personal information in order to predict your likelihood to require immediate care.
The more +s or -s, the more positively or negatively that factor impacted your predicted score.

- Age (-)
- Past expenses (+)
- Previous admissions (+)
- No known previous illnesses (--)
- Known illnesses in family (+)

Figure 2. An example of the experiment.

3.5. Trust and Perception of Fairness Scales

In this study, trust is assessed with six items using self-report scales following approaches in [66]. 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.

Furthermore, this study measures the perception of fairness with a single scale that focuses on the global perception of appropriateness [67]. The questionnaire on fairness is "Overall, female and male applicants are treated fairly by the system". The scale is on a 4-point Likert-type response scale ranging from 1 (strongly disagree), 2 (disagree), 3 (agree), to 4 (strongly agree).

3.6. Participants and Data Collection

In this study, 25 participants were recruited to conduct experimental tasks via various means of communication such as emails and social media posts. They 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 the perception of fairness, trust, and satisfaction in the AI-informed decision-making, respectively.

3.7. Analysis Methods

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

- The effects of explanation and introduced fairness on user trust;
- The effects of explanation and introduced fairness on user perception of fairness.

Before statistical analysis, trust and perception of fairness 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 or perception of fairness rating values, respectively, from the participant i, V_i^{min} and V_i^{max} are the minimum and maximum of trust or perception of fairness rating values respectively from the participant i in all of his/her tasks.

4. Results

This section explores the results of the user study. The analysis of the effects of fairness conditions and explanations on trust and the perception of fairness are firstly presented. The results from two scenarios are also compared in order to gain further insight into the relationship between explanations and fairness conditions.

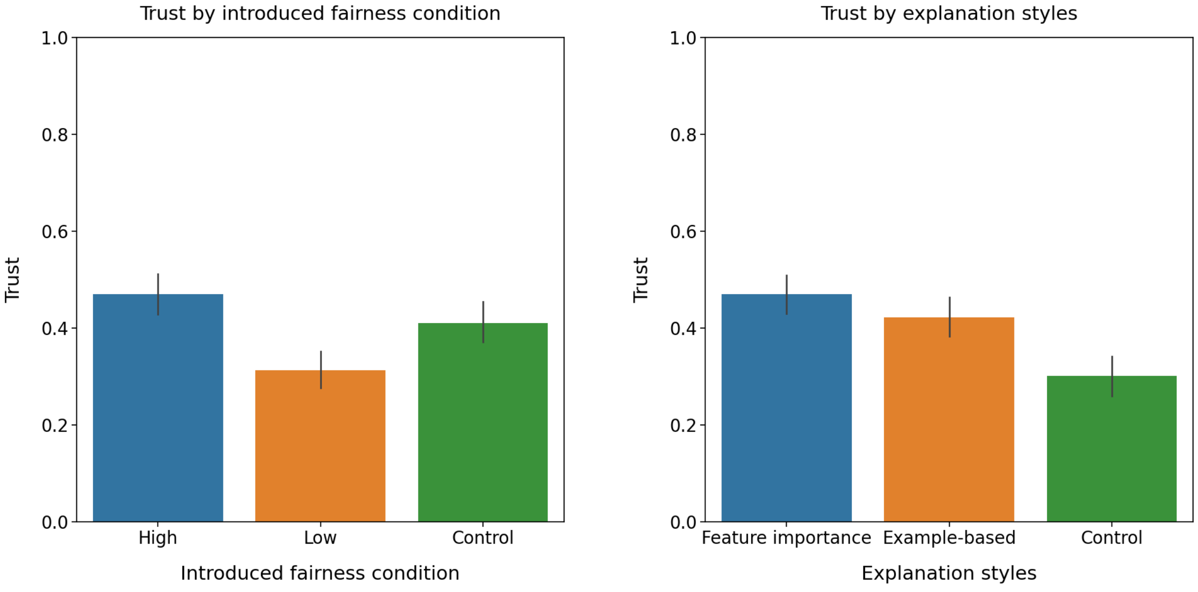
4.1. Trust

This subsection analyses the effects of explanation and fairness on trust in AI-informed decision-making.

4.1.1. Effects of Fairness on Trust

Figure 3a 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 $F(2, 447) = 14.300, p < 0.000$. A further post-hoc comparison with *t*-tests (with a Bonferroni correction under a significance level set at $\alpha < 0.017$) was conducted to find pair-wise differences in user trust between three fairness conditions. The adjusted significance alpha level of 0.017 was calculated by dividing the original alpha of 0.05 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 control condition of fairness (no fairness information presented) compared to the low fairness condition ($t = 3.289, p < 0.001$). Moreover, it was found that participants also had a statistically significant higher level of trust under the high fairness condition compared to that under the low fairness condition ($t = 5.428, p < 0.000$). However, there was not a statistically significant difference found in user trust between the introduced high fairness condition and the control condition ($t = 1.953, p < 0.052$).

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.



(**a**) (**b**)

Figure 3. User trust over fairness and explanations. (**a**) User trust under introduced fairness conditions. (**b**) User trust under explanation types.

4.1.2. Effects of Explanation on Trust

Figure 3b shows normalised trust values over various explanation conditions. A one-way ANOVA test revealed statistically significant differences in user trust under different explanation types $F(2, 447) = 17.325, p < 0.000$. Then, post-hoc tests with the aforementioned Bonferroni correction were conducted. It was found that participants had a statistically significant lower level of trust under the control condition (no explanation presented) than that under the feature importance-based explanation ($t = 5.656, p < 0.000$) and example-based explanation ($t = 4.080, p < 0.000$), respectively. There was not a significant difference in user trust between feature importance-based explanation and example-based explanation ($t = 1.619, p < 0.107$).

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.

4.1.3. Effects of Fairness and Explanation on Trust

A two-way ANOVA test was performed to analyse the effect of introduced fairness and explanation types on user trust in AI-informed decision-making. The two-way ANOVA test showed that there were no statistically significant interactions between fairness conditions and explanation types on trust, $F(4, 441) = 1.233, p < 0.296$. 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 4a 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, $F(2, 147) = 8.735, p < 0.000$. 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 ($t = 3.893, p < 0.000$). Moreover, user trust was significantly higher under the control condition (no fairness information presented) than that under the low introduced fairness level ($t = 3.372, p < 0.001$). However, there was not a significant difference in trust between the high introduced fairness and the control condition (t = 0.456, $p < 0.649$).

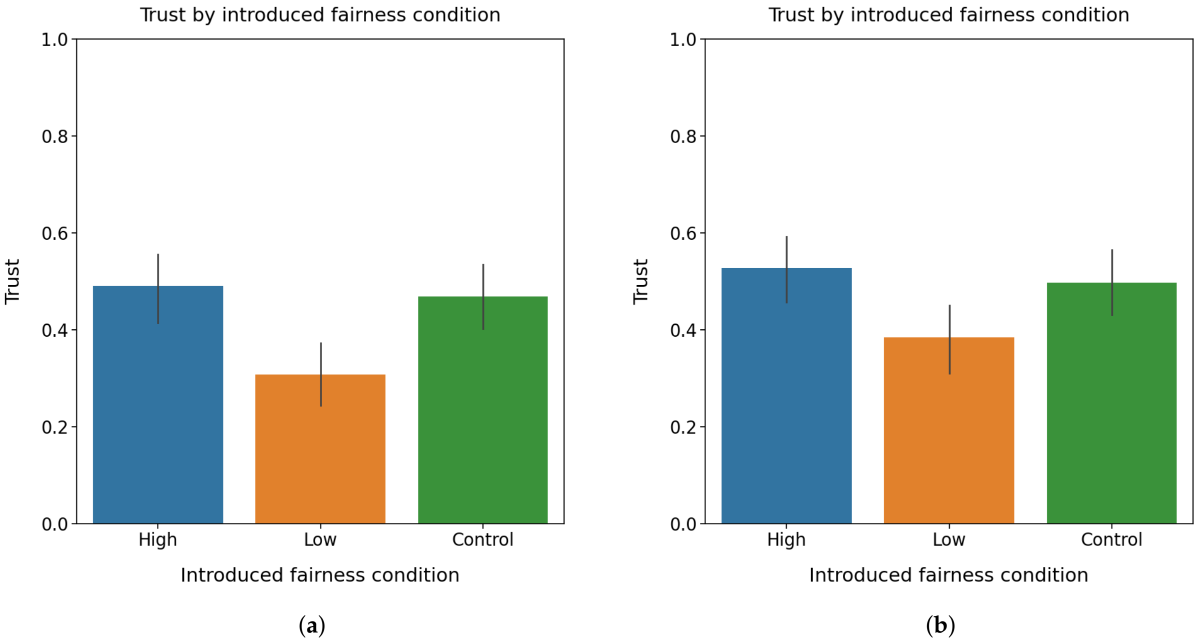


Figure 4. Effects of fairness on user trust under different explanations. (**a**) effects of fairness on user trust under the example-based explanation; (**b**) effects of fairness on user trust under feature importance-based explanations.

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 4b shows the normalized trust levels for introduced fairness levels under feature importance-based explanation. A one-way ANOVA test found significant differences in trust in different introduced fairness levels under the feature importance-based explanation, $F(2, 147) = 4.676, p < 0.011$. The further post-hoc t-tests found that user trust was statistically significant lower under the low introduced fairness level than that under the high introduced fairness level ($t = 2.794, p < 0.006$) and the control condition (no fairness information presented) ($t = 2.330, p < 0.022$), respectively (considering only two actual introduced fairness conditions (high and low), the adjusted alpha can be re-adjusted to $\alpha = 0.025$ $(0.05/2)$). However, no significant difference was found between the control condition and the high level of introduced fairness ($t = 0.603, p < 0.548$).

From the results, we can see that, under the feature importance-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. This is similar to the conclusion under the example-based explanation condition we obtained previously.

Effects of Explanation on Trust under Low Level Introduced Fairness

Figure 5a 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, $F(2, 147) = 4.149, p < 0.018$. The further t-test found that participants showed a significantly higher level of trust under feature importance-based explanation than that under the control condition (no explanation presented) ($t = 2.814, p < 0.006$). However, there were no significant differences found in user trust between the control condition and example-based explanation ($t = 1.336, p < 0.185$). There was also no significant difference in user trust between the two explanation types ($t = 1.567, p < 0.120$).

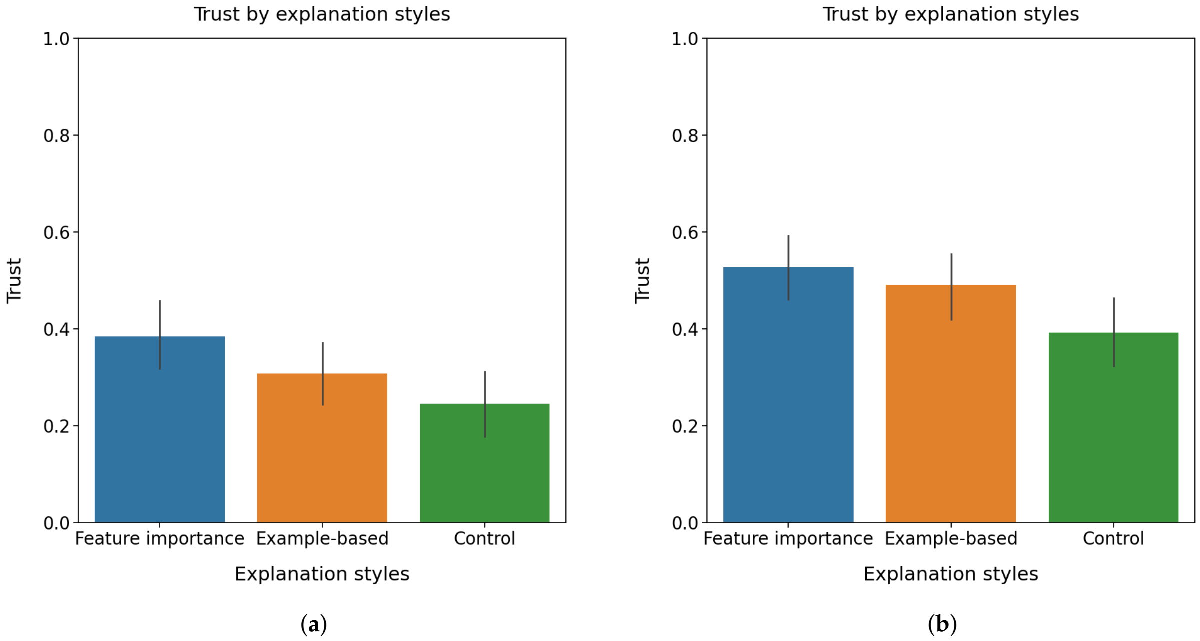


Figure 5. Effects of explanation on user trust under different fairness levels. (**a**) effects of explanation on user trust under low introduced fairness level; (**b**) effects of explanation on trust under high introduced fairness level.

Therefore, we can say that, under the low level of introduced fairness, the feature importance-based explanation significantly increased user trust in decision-making, but the example-based explanation did not.

Effects of Explanation on Trust under High Level of Introduced Fairness

Figure 5b shows the normalised trust values in different explanation types under the high level of introduced fairness. A one-way ANOVA test revealed statistical significant differences in user trust among explanation types under the high level of introduced fairness, $F(2, 147) = 3.855, p < 0.023$. The further post-hoc t-tests showed that user trust was significantly higher under feature importance-based explanation than that under the control condition (no explanation presented) ($t = 2.626, p < 0.010$). However, there was neither a significant difference in user trust between the control condition and example-based explanation ($t = 1.971, p < 0.052$), nor between example-based and feature importance-based explanation ($t = 0.723, p < 0.471$).

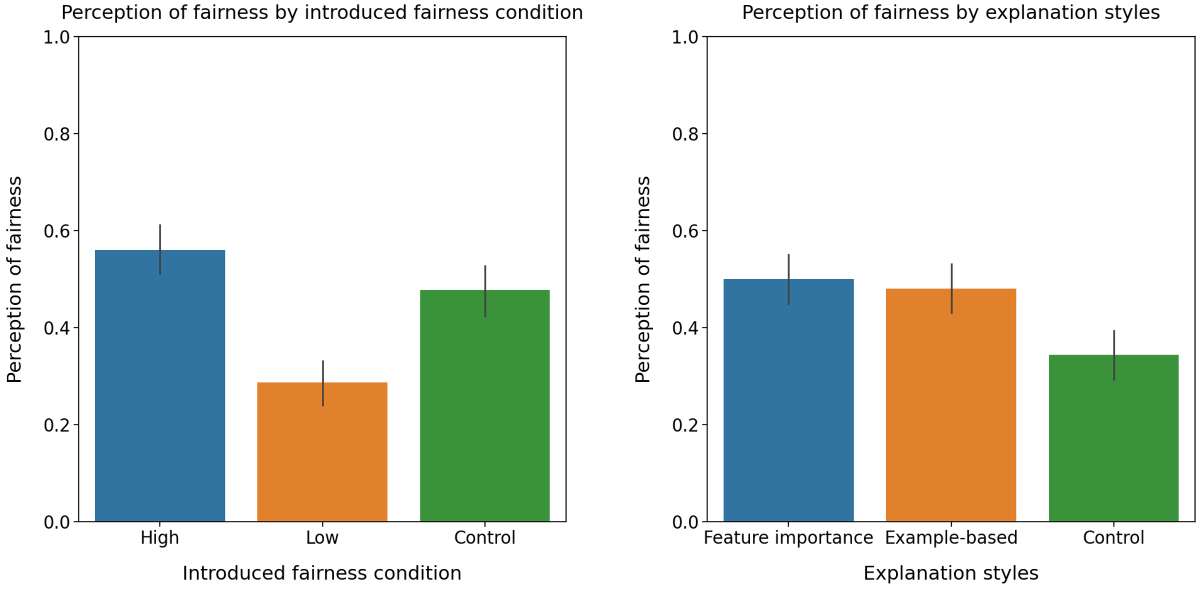
The high level of introduced fairness showed similar effects on user trust as the low level of introduced fairness: the feature importance-based explanation significantly increased user trust, while the example-based explanation did not.

4.2. Perception of Fairness

This subsection analyses the effects of explanation and introduced fairness on the perception of fairness.

4.2.1. Effects of Introduced Fairness on Perception of Fairness

Figure 6a shows normalised values of perception of fairness under different introduced fairness conditions. There were statistically significant differences found in the perception of fairness under different introduced fairness conditions with a one-way ANOVA test, $F(2, 147) = 31.435, p < 0.000$. The further post-hoc t-tests with Bonferroni correction as described previously (with a significance level $\alpha < 0.017$) were applied. The results showed that the low-level introduced fairness condition resulted in a statistically significant lower level of perception of fairness than that under the high level of introduced fairness ($t = 7.774, p < 0.000$). In addition, participants had a statistically significant lower level of perception of fairness under the low introduced fairness condition compared to the control condition (no fairness information presented) ($t = 5.477, p < 0.000$). There was no significant difference in the perception of fairness between the high-level introduced fairness and the control condition ($t = 2.281, p < 0.023$). However, considering only two actual introduced fairness conditions (high and low), the adjusted alpha can be re-adjusted to $\alpha = 0.025$ $(0.05/2)$. Therefore, the post-hoc test indicated a statistically significant higher level of the perception of fairness under the high-level introduced fairness than that under the control condition ($t = 2.281, p < 0.023$).



(**a**) (**b**)

Figure 6. Perception of fairness. (**a**) perception of fairness under different introduced fairness conditions; (**b**) perception of fairness under different explanation types.

The results indicated that the introduced fairness did affect the user's perception of fairness, where the lower level of introduced fairness decreased the user's perception of fairness, while the higher level of introduced fairness increased the user's perception of fairness.

4.2.2. Effects of Explanation on Perception of Fairness

Figure 6b shows normalised values of perception of fairness under different explanation types. A one-way ANOVA test revealed statistically significant differences in the perception of fairness under different explanation types, $F(2, 147) = 10.508, p < 0.000$. Then, post-hoc t-tests were applied to find pair-wise differences in the perception of fairness among explanation conditions. It was found that participants showed a statistically significant higher level of perception of fairness under example-based explanations ($t = 3.701, p < 0.000$) and feature importance-based explanations ($t = 4.167, p < 0.000$) respectively than that under control condition (no explanation presented). There were no significant differences found in the perception of fairness between example-based explanations and feature importance-based explanations ($t = 0.544, p < 0.587$).

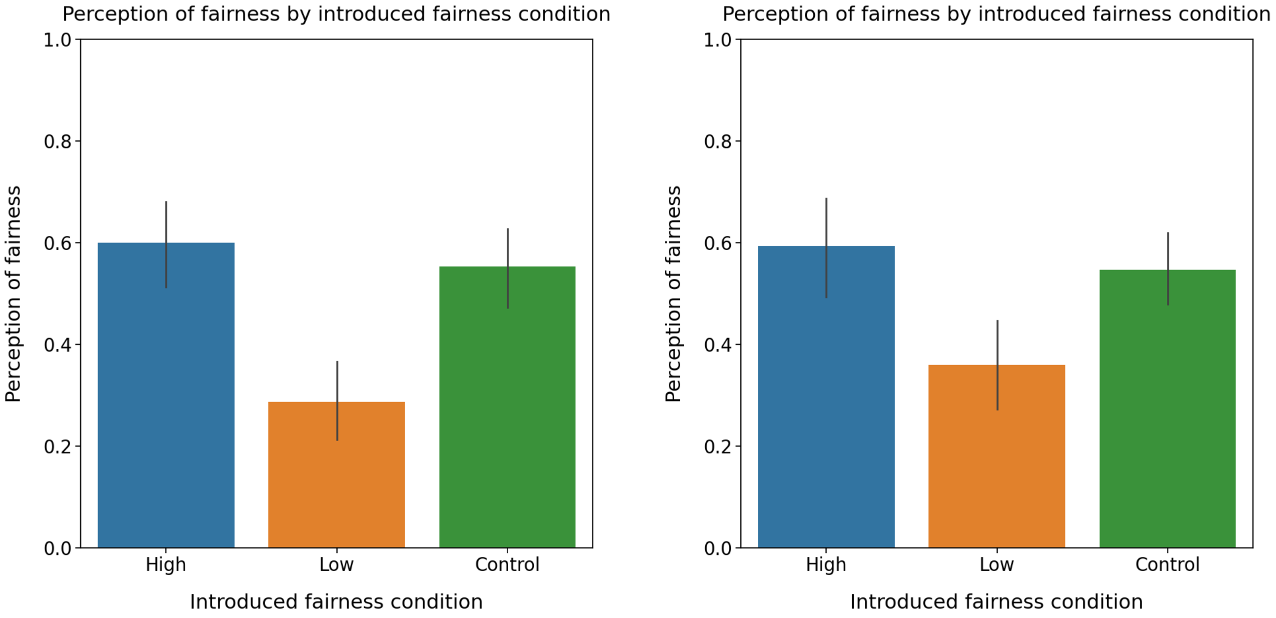
The results implied that explanations did benefit and increase the perception of fairness as we expected in AI-informed decision-making. However, different explanation types investigated in the study did not show differences in affecting the perception of fairness.

4.2.3. Effects of Explanation and Introduced Fairness on Perception of Fairness

A two-way ANOVA test did not find significant interactions between fairness levels and explanation types on the perception of fairness, $F(4, 441) = 1.005, p < 0.405$. This subsection further analyses the effects of introduced fairness on the perception of fairness under different given explanation types, and the effects of explanation on perception of fairness under different given introduced fairness levels.

Effects of Fairness on Perception of Fairness under Example-Based Explanations

Figure 7a shows the normalised values of perception of fairness with different introduced fairness under example-based explanations. A statistically significant difference in perceived fairness among different fairness conditions was found by performing a one-way ANOVA test, $F(2, 147) = 17.964, p < 0.000$. The further post-hoc t-tests showed that high introduced fairness resulted in a statistically significant higher level of perception of fairness than that under the low introduced fairness ($t = 5.408, p < 0.000$). Moreover, the analysis showed that the control condition (no fairness information presented) resulted in a higher perception of fairness than that under the low introduced fairness ($t = 4.903, p < 0.000$). However, there was not a significant difference in perceived fairness found between the high introduced fairness and control condition ($t = 0.821, p < 0.414$).



(**a**) (**b**)

Figure 7. Perception of fairness per introduced fairness. (**a**) perception of fairness per introduced fairness levels under example-based explanations; (**b**) perception of fairness per introduced fairness levels under feature importance-based explanations.

The results indicated that, under example-based explanations, the low level of introduced fairness decreased perception of fairness significantly, while the high level of introduced fairness did not.

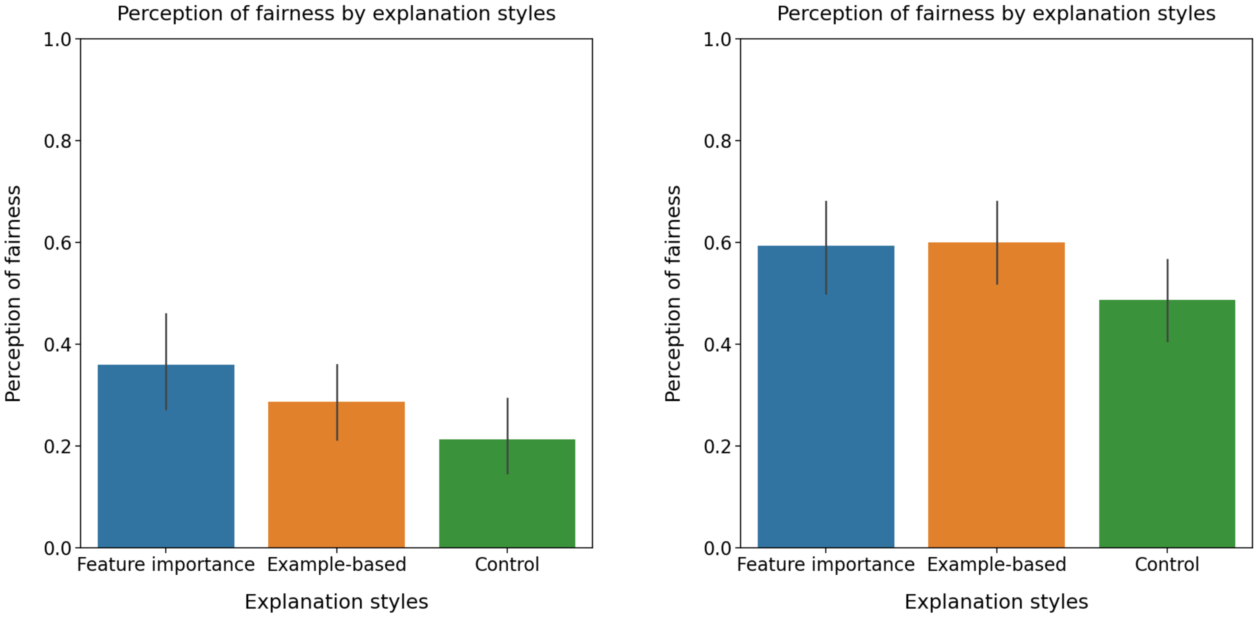
Effects of Fairness on Perceived Fairness under Feature Importance-Based Explanations

The same procedure was carried out for the feature importance-based explanation. Figure 7b shows the normalised values of perception of fairness with different introduced fairness under feature importance-based explanations. A one-way ANOVA test found statistically significant differences in perceived fairness among different introduced fairness conditions, $F(2, 147) = 7.892, p < 0.001$. The further post-hoc t-tests found that participants had a higher level of perception of fairness under the high-level introduced fairness, compared to the low level introduced fairness ($t = 3.532, p < 0.001$). Moreover, the tests found a significantly higher level of perception of fairness under the control condition (no fairness information presented) than that under the low-level introduced fairness ($t = 3.115, p < 0.002$). However, there were no significant differences in perception of fairness between the high level and control groups of introduced fairness ($t = 0.774, p = 0.441$).

Therefore, similar conclusions were obtained for the perception of fairness under feature importance-based explanations as that under example-based explanations: the low level of introduced fairness decreased the perception of fairness significantly, while the high level of introduced fairness did not.

Effects of Explanation on Perceived Fairness under Low Level of Introduced Fairness

Figure 8a shows the normalised perception of fairness with different explanation types under the low level of introduced fairness. A one-way ANOVA test found that there were statistically significant differences in perception of fairness among three explanation types, $F(2, 147) = 3.199, p < 0.044$. The further post-hoc t-tests showed that participants' perception of fairness was significantly higher under feature importance-based explanations than that under the control condition (no explanation presented) ($t = 2.478, p < 0.015$). However, there were no significant differences in the perception of fairness between example-based and feature importance-based explanations ($t = 1.205, p < 0.231$). Moreover, there was no significant difference in the perception of fairness between the control condition (no explanation presented) and the example-based explanation ($t = 1.366, p < 0.175$).



(**a**) (**b**)

Figure 8. Effect of explanation types on the perception of fairness. (**a**) effect of explanation types on perception of fairness under low level of introduced fairness; (**b**) effect of explanation types on the perception of fairness under the high level of introduced fairness.

The results showed that, under the low level of introduced fairness, feature importance-based explanations significantly increased participants' perception of fairness, while example-based explanations did not.

Effects of Explanation on Perceived Fairness under High-Level Introduced Fairness

Figure 8b shows the normalised values of perception of fairness with different explanation types under the high level of introduced fairness. A one-way ANOVA test revealed that there were no statistically significant differences found in the perception of fairness among different explanation types under the high level of introduced fairness, $F(2, 147) = 2.074, p < 0.129$.

The results indicated that different explanation types did not benefit a human's perception of fairness under the high level of introduced fairness.

4.3. Trust and Perception of Fairness in Different Scenarios

This subsection investigates the effect of scenarios (health insurance decision-making and medical treatment decision-making) on user trust and perception of fairness. Moreover, we also explore whether the scenarios combined with introduced fairness levels and explanation types affect user trust and perception of fairness.

A one-way ANOVA test did not find any statistically significant differences in user trust between two scenarios $F(1, 448) = 2.263, p < 0.133$. A one-way ANOVA test also did not find any statistically significant differences in perception of fairness between two scenarios $F(1, 448) = 0.593, p = 0.442$.

4.3.1. Explanation Types and Scenarios on Trust

To further investigate the effect of explanation types and scenarios on user trust, a two-way ANOVA test was performed, which did not find significant interactions in trust between introduced fairness and scenario $F(2, 444) = 1.287, p < 0.277$.

A further one-way ANOVA test revealed that explanation types in connection with the scenario of health insurance decision-making had a significant effect on user trust, $F(2, 222) = 11.226, p < 0.000$. Figure 9a shows the normalised values of user trust for the scenario of health insurance decision-making. The post-hoc t-tests showed a statistically significant higher level of user trust under feature importance-based explanations than that under example-based explanations ($t = 2.329, p < 0.021$). Moreover, participants showed significantly higher user trust under example-based explanation ($t = 2.455, p < 0.015$) and feature importance-based explanation ($t = 4.645, p < 0.000$) than under the control condition (no explanation presented), respectively.

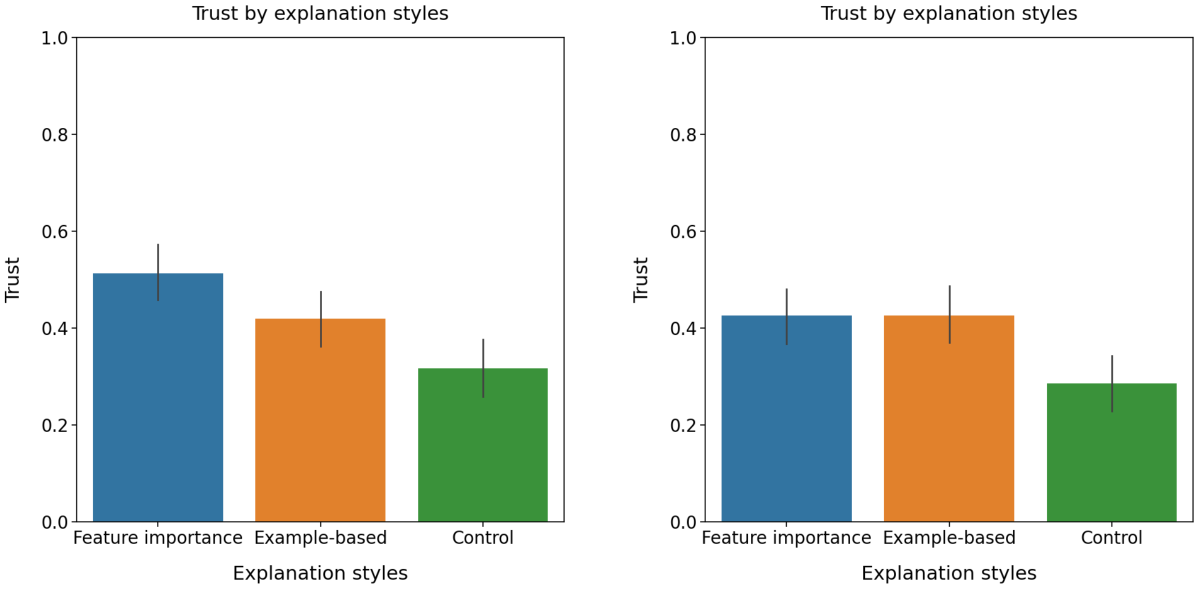
Figure 9b shows the normalised trust values per explanation types in the scenario of medical treatment decision-making. A one-way ANOVA test found that explanation types in connection with the medical treatment decision-making had a significant effect on user trust $F(2, 222) = 7.508, p < 0.001$. The further post-hoc t-tests found that participants showed significantly higher trust under example-based explanation ($t = 3.291, p < 0.001$) and feature importance-based explanation ($t = 3.384, p < 0.001$) than that under the control condition (no explanation presented), respectively. However, there were no significant differences in user trust found between both explanation types ($t = 0.000, p = 1$).

The results indicated that the effect of explanations on user trust was slightly different in two scenarios of health insurance decision-making and medical treatment decision-making despite explanations increasing user trust in both scenarios.

4.3.2. Introduced Fairness and Scenarios on Trust

To further investigate the effect of introduced fairness and scenarios on user trust, a two-way ANOVA test did not find any significant interactions in trust between introduced fairness levels and scenarios $F(2, 444) = 0.109, p = 0.896$.

Figure 10a shows the normalised trust values over introduced fairness levels in the scenario of health insurance decision-making. A one-way ANOVA test revealed that the introduced fairness condition in connection with the scenario of health insurance decision-making showed a significant influence on user trust $F(2, 222) = 8.446, p < 0.000$. The post-hoc t-tests found that participants had significantly higher user trust under the high level of introduced fairness than that under the low level of introduced fairness ($t = 4.185, p < 0.000$). Moreover, participants also had significantly higher user trust under the control condition (no fairness information presented) than that under the low-level introduced fairness ($t = 2.433, p < 0.016$).

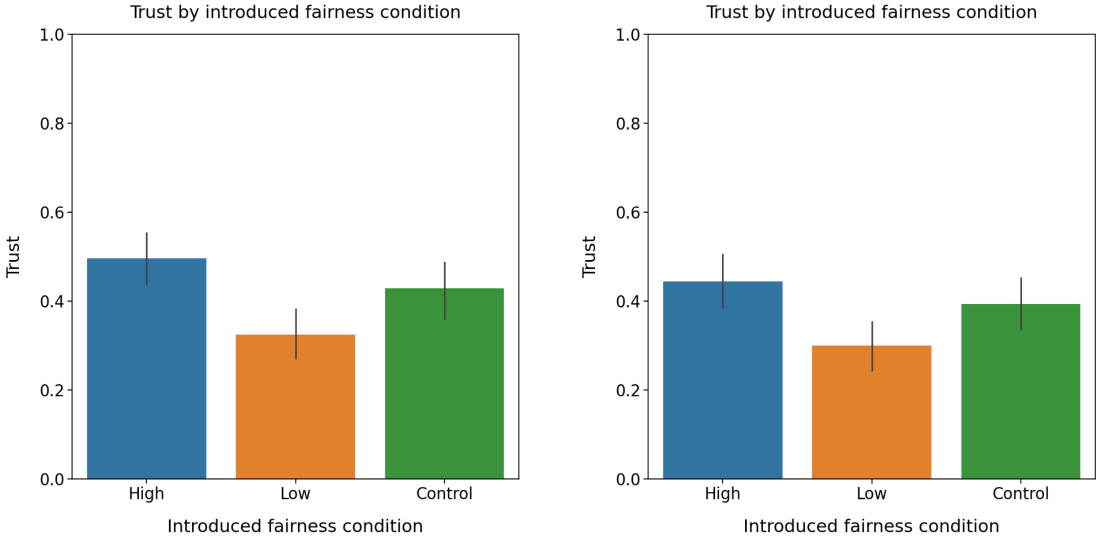


(**a**) (**b**)

Figure 9. User trust over explanation types in two scenarios. (**a**) Health insurance decision-making. (**b**) Medical treatment decision-making.

Figure 10b shows the normalised values of trust over introduced fairness conditions in the scenario of medical treatment decision-making. A one-way ANOVA test found that introduced fairness in connection with the scenario of medical treatment decision-making had a significant influence on user trust $F(2, 222) = 5.968$, $p < 0.003$. The post-hoc t-tests showed that user trust was significantly higher under the high level of introduced fairness than that under the low level of introduced fairness ($t = 3.494$, $p < 0.000$). However, there was neither a significant difference in user trust between the control condition and the high level of introduced fairness ($t = 1.164$, $p < 0.739$), nor between the control condition and the low level of introduced fairness ($t = 2.211$, $p < 0.086$).

The results showed that the effect of introduced fairness on user trust was similar in two scenarios, where the low level of introduced fairness decreased user trust while the high level of introduced fairness did not.



(**a**) (**b**)

Figure 10. User trust over introduced fairness in two scenarios. (**a**) Health insurance decision-making. (**b**) Medical treatment decision-making.

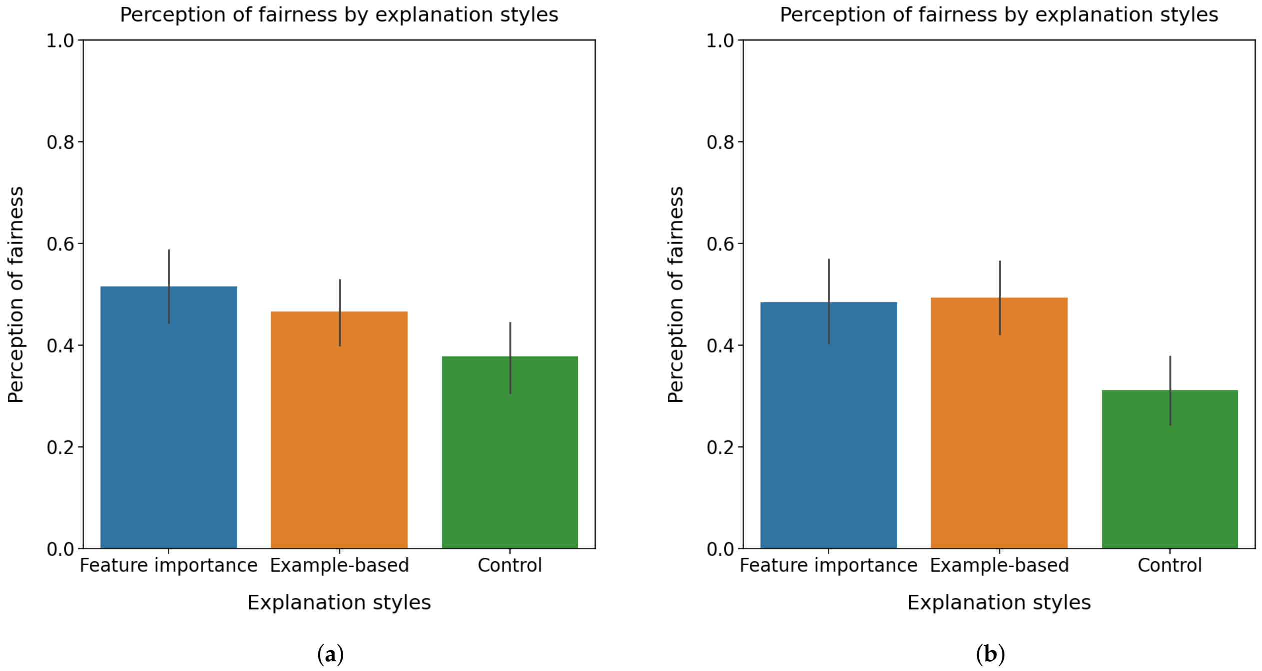
4.3.3. Explanation Types and Scenarios on Perception of Fairness

A two-way ANOVA test was performed to analyse the effect of explanation types and scenarios on the perception of fairness. There were no statistically significant interactions found in a perception of fairness between explanation types and scenarios $F(2,444) = 0.813$, $p < 0.444$.

Figure 11a shows the normalised values of perception of fairness under different explanation types in the scenario of health insurance decision-making. A one-way ANOVA test showed a statistically significant difference in perceived fairness, $F(2,222) = 3.863$, $p < 0.022$. The further post-hoc t-tests found that participants had significantly higher perception of fairness under feature importance-based explanations than that under the control condition (no explanation presented), $t = 2.688, p < 0.008$. However, there was neither a significant difference in perception of fairness between example-based explanation and control condition ($t = 1.751, p < 0.082$) nor between the two explanation types ($t = 1.003, p < 0.317$).

Figure 11b shows the normalised values of perception of fairness under different explanation types in the scenario of medical treatment decision-making. A one-way ANOVA test found a statistically significant difference in perceived fairness among explanation conditions, $F(2,222) = 7.190, p < 0.001$. The further post-hoc t-tests found that the perception of fairness was significantly higher under example-based explanation ($t = 3.450, p < 0.001$) and feature importance-based explanation ($t = 3.192, p < 0.002$), respectively, than that under the control condition (no explanation presented). However, there were no significant differences in perceived fairness found between the two explanation types ($t = 0.160, p < 0.873$).

The results revealed that participants showed slightly different behaviours of perception of fairness in two studied scenarios, where example-based explanations increased perception of fairness in the medical treatment decision-making but not in the health insurance decision-making despite feature importance-based explanations increased perception of fairness in both scenarios.



(**a**) (**b**)

Figure 11. Perception of fairness over explanation types in two scenarios. (**a**) Health insurance decision-making. (**b**) Medical treatment decision-making.

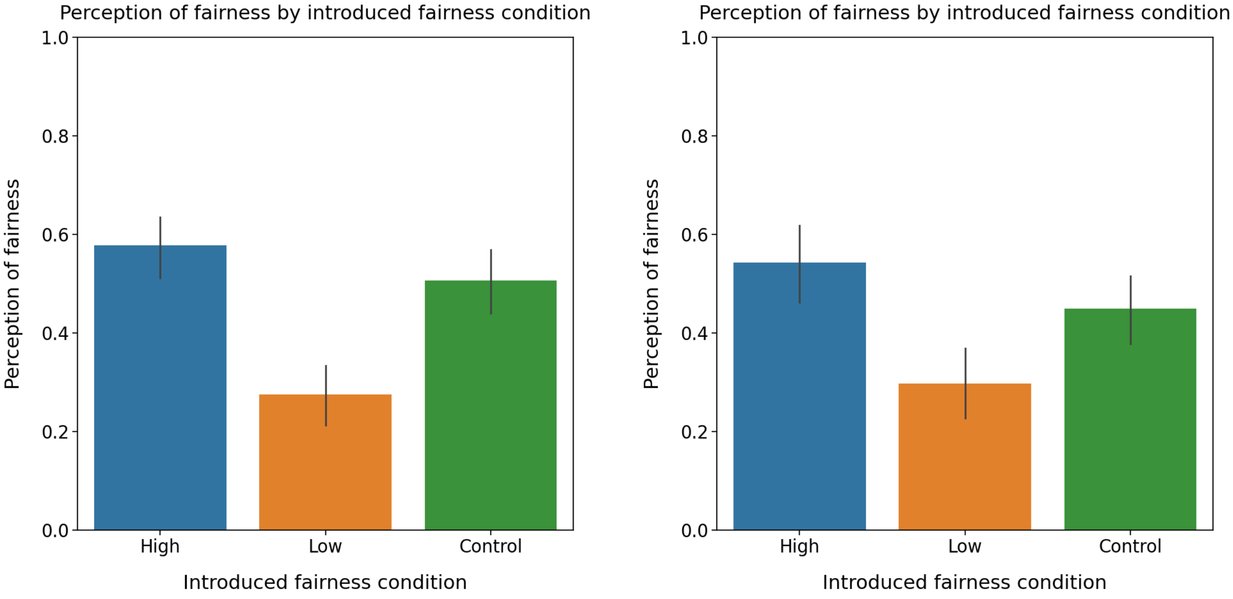
4.3.4. Introduced Fairness and Scenario on Perception of Fairness

A two-way ANOVA test was performed to investigate the effect of explanation types and scenarios on the perception of fairness. There were no statistically significant interactions in the perception of fairness between introduced fairness conditions and scenarios, $F(2, 444) = 0.680, p = 0.507$.

Figure 12a shows the normalised perception of fairness under introduced fairness conditions in the scenario of health insurance decision-making. A one-way ANOVA test revealed a statistically significant difference in perception of fairness among introduced fairness conditions, $F(2, 222) = 23.078, p = 0.000$. The further post-hoc t-tests found that the perception of fairness was significantly higher under the high level of introduced fairness ($t = 6.642, p < 0.000$) and the control condition (no fairness information presented)($t = 4.983, p < 0.000$), respectively, than that under the low level of introduced fairness. However, there were no significant differences in perceived fairness found between high-level introduced fairness and the control condition ($t = 1.493, p < 0.138$).

Figure 12b shows the normalised values of perception of fairness under introduced fairness conditions in the scenario of medical treatment decision-making. A one-way ANOVA test found a statistically significant difference in perceived fairness among introduced fairness conditions, $F(2, 222) = 10.671, p < 0.000$. The further post-hoc t-tests found that the perception of fairness was significantly higher under the high level of introduced fairness ($t = 4.542, p < 0.000$) and the control condition ($t = 2.895, p < 0.004$), respectively, than that under the low level of introduced fairness. However, there were no significant differences in perceived fairness found between high level introduced fairness and the control condition ($t = 1.723, p < 0.087$).

The results indicated that participants showed similar behaviours of perception of fairness in two studied scenarios, where the low level of introduced fairness decreased perceived fairness while the high level of introduced fairness did not.



(**a**) (**b**)

Figure 12. Perception of fairness over introduced fairness in two scenarios. (**a**) Health insurance decision-making. (**b**) Medical treatment decision-making.

5. Discussion

As discussed in earlier sections, explanation and fairness are two indispensable components in AI-informed decision-making for trustworthy AI. AI-informed decision-making and automated aids have been becoming much more popular with the advent of new AI-based intelligent applications. Therefore, we opted to study the effects of both AI explanations and fairness on human-AI trust and human perception of fairness, respectively, in specialised AI-informed decision-making scenarios.

From the trust's perspective, this 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. Similar conclusions for user trust were obtained under the feature importance-based explanation condition. Furthermore, from the introduced fairness' perspective, it revealed that, under the low level of introduced fairness, the feature importance-based explanation significantly increased user trust in decision-making, but the example-based explanation did not. The high level of introduced fairness showed similar effects on user trust as the low level of introduced fairness. It also implies that the introduced fairness levels did not affect user trust too much.

From the perceived fairness' perspective, this study found that the fairness statement in the scenario did affect the user's perception of fairness, where the low level of introduced fairness decreased the user's perception of fairness, while the high level of introduced fairness increased the user's perception of fairness. Moreover, the addition of explanations benefited the perception of fairness, and different explanation types did not show differences in affecting the perception of fairness. We also drilled down into the effects on perceived fairness under specific conditions. From the explanation's perspective, it was found that, under example-based explanations, the low level of introduced fairness decreased the perception of fairness, while the high level of introduced fairness did not. Similar conclusions were obtained for the perception of fairness under feature importance-based explanations. From the introduced fairness' perspective, the results showed that, under the low level of introduced fairness, feature importance-based explanations significantly increased participants' perception of fairness, while example-based explanations did not. However, different explanation types did not benefit a human's perception of fairness under the high level of introduced fairness. It shows that the effects of fairness levels and explanation types are different from their effects on user trust.

We also compared participants' responses of trust and perception of fairness under different application scenarios of health insurance decision-making and medical treatment decision-making. It was found that the effect of explanations on user trust was slightly different in the two scenarios despite explanations increasing user trust in both scenarios. The effect of introduced fairness on user trust was similar in two scenarios, where the low level of introduced fairness decreased user trust while the high level of introduced fairness did not. From the perceived fairness' perspectives, participants showed slightly different responses to the perception of fairness under different explanation types in two studied scenarios, where example-based explanations increased user trust in the medical treatment decision-making but not in the health insurance decision-making despite feature importance-based explanations increased user trust in both scenarios. While participants showed similar behaviours of perception of fairness in two studied scenarios, where the low level of introduced fairness decreased perceived fairness and the high level of introduced fairness did not.

Compared with previous studies [68] which only focus on the analysis of explanations on human's perception of fairness and trust, this study investigated the effects of both AI explanations and introduced fairness on perceived fairness and trust in AI-informed decision-making. Such analysis helps to have a more comprehensive understanding of interrelations between explanations and fairness in AI-informed decision-making.

These findings suggest that the deployment of AI explanation and fairness statements in real-world applications is complex: we need to not only consider explanation types and levels of introduced fairness but also consider the scenarios that AI-informed decision-making is used for. In order to maximise user trust and perception of fairness 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. Conclusions

This paper investigated the effects of introduced fairness and explanation on the perception of fairness and 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. From the perceived fairness' perspective, the introduced fairness affected the user's perception of fairness, and the low level and high level of introduced fairness affected the user's perception of fairness differently. In addition, the study found that the AI explanations increased users' perceptions of fairness. The effects of application scenarios on trust and perception of fairness were also investigated in this study. These findings demonstrated important implications in the deployment of AI applications with explanations and fairness. Besides trust and fairness, other AI ethical principles such as accountability also play important roles in trustworthy AI [69]. Our future work will focus on the effects of explanation and introduced fairness on the accountability of AI, as well as the effects of other factors such as application scenarios on the accountability of AI. We will also investigate the gender discrimination in AI-informed decision-making regarding the perception of fairness and trust.

Author Contributions: Conceptualization, J.Z., F.C. and A.H.; Data curation, K.T.; Formal analysis, A.A.; Funding acquisition, J.Z., F.C. and A.H.; Methodology, A.A., J.Z., K.T. and A.H.; Writing—original draft, A.A., J.Z. and A.H.; Writing—review and editing, J.Z., F.C. and A.H. All authors have read and agreed to the published version of the manuscript.

Funding: 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.

Institutional Review Board Statement: This work does not raise any ethical issues. No information that can be used to identify a participant was collected. Moreover, the data collection was anonymous. Therefore, this work was exempted from the ethics review according to the University ethics guidelines.

Informed Consent Statement: Informed consent was obtained from all subjects involved in the study.

Data Availability Statement: Not applicable.

Conflicts of Interest: The authors declare that there are no conflict of interests.

References

1. White Paper on Artificial Intelligence—A European Approach to Excellence and Trust. Available online: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52020DC0065 (accessed on 31 May 2022).
2. Zhou, J.; Arshad, S.Z.; Luo, S.; Chen, F. Effects of Uncertainty and Cognitive Load on User Trust in Predictive Decision Making. In *Human-Computer Interaction—INTERACT 2017*; Bernhaupt, R., Dalvi, G., Joshi, A.K., Balkrishan, D., O'Neill, J., Winckler, M., Eds.; Springer: Cham, Switzerland, 2017; pp. 23–39.
3. 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), Doha, Qatar, 29–31 October 2021; pp. 1–5.
4. Castelvecchi, D. Can we open the black box of AI? *Nat. News* **2016**, *538*, 20. [CrossRef] [PubMed]
5. 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. *Int. J. Comput. Sci. Eng.* **2016**, *13*, 378–389. [CrossRef]

6. Zhou, J.; Gandomi, A.H.; Chen, F.; Holzinger, A. Evaluating the Quality of Machine Learning Explanations: A Survey on Methods and Metrics. *Electronics* **2021**, *10*, 593. [CrossRef]
7. 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*; Human–Computer Interaction Series; Springer International Publishing: Cham, Switzerland, 2018; pp. 3–19.
8. Zhou, J.; Chen, F. (Eds.) *Human and Machine Learning: Visible, Explainable, Trustworthy and Transparent*; Springer: Cham, Switzerland, 2018.
9. Holzinger, K.; Mak, K.; Kieseberg, P.; Holzinger, A. Can we Trust Machine Learning Results? Artificial Intelligence in Safety-Critical Decision Support. *ERCIM News* **2018**, *112*, 42–43.
10. Stoeger, K.; Schneeberger, D.; Kieseberg, P.; Holzinger, A. Legal aspects of data cleansing in medical AI. *Comput. Law Secur. Rev.* **2021**, *42*, 105587. [CrossRef]
11. Stoeger, K.; Schneeberger, D.; Holzinger, A. Medical Artificial Intelligence: The European Legal Perspective. *Commun. ACM* **2021**, *64*, 34–36. [CrossRef]
12. Pieters, W. Explanation and trust: What to tell the user in security and AI? *Ethics Inf. Technol.* **2011**, *13*, 53–64. [CrossRef]
13. 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*; Springer: Cham, Switzerland, 2019; pp. 94–113.
14. Alam, L.; Mueller, S. Examining the effect of explanation on satisfaction and trust in AI diagnostic systems. *BMC Med Inform. Decis. Mak.* **2021**, *21*, 178. [CrossRef]
15. Zhou, J.; Chen, F. Making machine learning useable. *Int. J. Intell. Syst. Technol. Appl.* **2015**, *14*, 91–109. [CrossRef]
16. Dwivedi, Y.K.; Hughes, L.; Ismagilova, E.; Aarts, G.; Coombs, C.; Crick, T.; Duan, Y.; Dwivedi, R.; Edwards, J.; Eirug, A.; et al. Artificial Intelligence (AI): Multidisciplinary perspectives on emerging challenges, opportunities, and agenda for research, practice and policy. *Int. J. Inf. Manag.* **2021**, *57*, 101994. [CrossRef]
17. Berk, R.; Heidari, H.; Jabbari, S.; Kearns, M.; Roth, A. Fairness in criminal justice risk assessments: The state of the art. *Sociol. Methods Res.* **2018**, *50*, 0049124118782533. [CrossRef]
18. Feldman, M.; Friedler, S.A.; Moeller, J.; Scheidegger, C.; Venkatasubramanian, S. Certifying and removing disparate impact. In Proceedings of the KDD2015, Sydney, NSW, Australia, 10–13 August 2015; pp. 259–268.
19. Starke, C.; Baleis, J.; Keller, B.; Marcinkowski, F. Fairness Perceptions of Algorithmic Decision-Making: A Systematic Review of the Empirical Literature. *arXiv* **2021**, arXiv:2103.12016.
20. Duan, Y.; Edwards, J.S.; Dwivedi, Y.K. Artificial intelligence for decision-making in the era of Big Data—Evolution, challenges and research agenda. *Int. J. Inf. Manag.* **2019**, *48*, 63–71. [CrossRef]
21. Kuzior, A.; Kwilinski, A. Cognitive Technologies and Artificial Intelligence in Social Perception. *Manag. Syst. Prod. Eng.* **2022**, *30*, 109–115. [CrossRef]
22. Komodromos, M. Employees' Perceptions of Trust, Fairness, and the Management of Change in Three Private Universities in Cyprus. *J. Hum. Resour. Manag. Labor Stud.* **2014**, *2*, 35–54.
23. Roy, S.K.; Devlin, J.F.; Sekhon, H. The impact of fairness on trustworthiness and trust in banking. *J. Mark. Manag.* **2015**, *31*, 996–1017. [CrossRef]
24. 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 (IUI'19), Marina del Ray, CA, USA, 17–20 March 2019; pp. 275–285.
25. Kilbertus, N.; Carulla, M.R.; Parascandolo, G.; Hardt, M.; Janzing, D.; Schölkopf, B. Avoiding discrimination through causal reasoning. In Proceedings of the Advances in Neural Information Processing Systems, Long Beach, CA, USA, 4–9 December 2017; pp. 656–666.
26. Bellamy, R.K.E.; Dey, K.; Hind, M.; Hoffman, S.C.; Houde, S.; Kannan, K.; Lohia, P.; Martino, J.; Mehta, S.; Mojsilovic, A.; et al. AI Fairness 360: An Extensible Toolkit for Detecting, Understanding, and Mitigating Unwanted Algorithmic Bias. *arXiv* **2018**, arXiv:1810.01943.
27. Shin, D. User Perceptions of Algorithmic Decisions in the Personalized AI System:Perceptual Evaluation of Fairness, Accountability, Transparency, and Explainability. *J. Broadcast. Electron. Media* **2020**, *64*, 541–565. [CrossRef]
28. Corbett-Davies, S.; Goel, S. The measure and mismeasure of fairness: A critical review of fair machine learning. *arXiv* **2018**, arXiv:1808.00023.
29. Nabi, R.; Shpitser, I. Fair inference on outcomes. In Proceedings of the Thirty-Second AAAI Conference on Artificial Intelligence. AAAI Conference on Artificial Intelligence, New Orleans, LA, USA, 2–7 February 2018; Volume 2018, p. 1931.
30. Glymour, B.; Herington, J. Measuring the biases that matter: The ethical and casual foundations for measures of fairness in algorithms. In Proceedings of the Conference on Fairness, Accountability, and Transparency, Atlanta, GA, USA, 29–31 January 2019; pp. 269–278.
31. Lee, M.K.; Baykal, S. 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, Portland, OR, USA, 25 February–1 March 2017; pp. 1035–1048.
32. Lee, M.K.; Jain, A.; Cha, H.J.; Ojha, S.; Kusbit, D. Procedural Justice in Algorithmic Fairness: Leveraging Transparency and Outcome Control for Fair Algorithmic Mediation. *Proc. ACM Hum.-Comput. Interact.* **2019**, *3*, 1–26. [CrossRef]

33. Helberger, N.; Araujo, T.; de Vreese, C.H. Who is the fairest of them all? Public attitudes and expectations regarding automated decision-making. *Comput. Law Secur. Rev.* **2020**, *39*, 105456. [CrossRef]
34. Harrison, G.; Hanson, J.; Jacinto, C.; Ramirez, J.; Ur, B. An Empirical Study on the Perceived Fairness of Realistic, Imperfect Machine Learning Models. In Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, FAT* '20, Barcelona, Spain, 27–30 January 2020; pp. 392–402. [CrossRef]
35. Shin, D.; Park, Y.J. Role of fairness, accountability, and transparency in algorithmic affordance. *Comput. Hum. Behav.* **2019**, *98*, 277–284. [CrossRef]
36. Shin, D.; Zhong, B.; Biocca, F.A. Beyond user experience: What constitutes algorithmic experiences? *Int. J. Inf. Manag.* **2020**, *52*, 102061. [CrossRef]
37. Shin, D. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *Int. J. Hum.-Comput. Stud.* **2021**, *146*, 102551. [CrossRef]
38. Binns, R.; Van Kleek, M.; Veale, M.; Lyngs, U.; Zhao, J.; Shadbolt, N. '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, Association for Computing Machinery, Montreal, QC, Canada, 21–26 April 2018; pp. 1–14. [CrossRef]
39. 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, Association for Computing Machinery, CHI EA '15, Seoul, Korea, 18–23 April 2015; pp. 923–928. [CrossRef]
40. 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 Trans. Comput.-Hum. Interact.* **2015**, *21*, 1–23. [CrossRef]
41. Kizilcec, R.F. How Much Information? Effects of Transparency on Trust in an Algorithmic Interface. Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, Association for Computing Machinery, CHI '16, San Jose, CA, USA, 7–12 May 2016; pp. 2390–2395. [CrossRef]
42. 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, FAT* '20, Barcelona, Spain, 27–30 January 2020; pp. 295–305.
43. Yin, M.; Vaughan, J.W.; Wallach, H. Does Stated Accuracy Affect Trust in Machine Learning Algorithms? In Proceedings of the ICML2018 Workshop on Human Interpretability in Machine Learning (WHI 2018), Stockholm, Sweden, 14 July 2018; pp. 1–2.
44. Earle, T.C.; Siegrist, M. On the Relation Between Trust and Fairness in Environmental Risk Management. *Risk Anal.* **2008**, *28*, 1395–1414. [CrossRef]
45. Nikbin, D.; Ismail, I.; Marimuthu, M.; Abu-Jarad, I. The effects of perceived service fairness on satisfaction, trust, and behavioural intentions. *Singap. Manag. Rev.* **2011**, *33*, 58–73.
46. 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, FAccT '21, Virtual Event, 3–10 March 2021; pp. 690–700. [CrossRef]
47. Holzinger, A.; Langs, G.; Denk, H.; Zatloukal, K.; Müller, H. Causability and Explainability of Artificial Intelligence in Medicine. *Wiley Interdiscip. Rev. Data Min. Knowl. Discov.* **2019**, *9*, 1–13. [CrossRef]
48. Holzinger, A.; Malle, B.; Saranti, A.; Pfeifer, B. Towards Multi-Modal Causability with Graph Neural Networks enabling Information Fusion for explainable AI. *Inf. Fusion* **2021**, *71*, 28–37. [CrossRef]
49. 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. *Knowl. Based Syst.* **2021**, *220*, 106916. [CrossRef]
50. Holzinger, A.; Carrington, A.; Mueller, H. Measuring the Quality of Explanations: The System Causability Scale (SCS). Comparing Human and Machine Explanations. *KI -Kuenstliche Intell.* **2020**, *34*, 193–198. [CrossRef]
51. Koh, P.W.; Liang, P. Understanding Black-box Predictions via Influence Functions. *Proc. ICML* **2017**, *70*, 1885–1894.
52. Papenmeier, A.; Englebienne, G.; Seifert, C. How model accuracy and explanation fidelity influence user trust. *arXiv* **2019**, arXiv:1907.12652.
53. 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, Cagliari, Italy, 17 March 2020; pp. 1–6.
54. Wang, X.; Yin, M. Are Explanations Helpful? A Comparative Study of the Effects of Explanations in AI-Assisted Decision-Making. In Proceedings of the 26th International Conference on Intelligent User Interfaces, College Station, TX, USA, 14–17 April 2021; pp. 318–328.
55. Kelley, K.H.; Fontanetta, L.M.; Heintzman, M.; Pereira, N. Artificial Intelligence: Implications for Social Inflation and Insurance. *Risk Manag. Insur. Rev.* **2018**, *21*, 373–387. [CrossRef]
56. Article 29 Working Party. Guidelines on Automated Individual Decision-Making and Profiling for the Purposes of Regulation 2016/679. Available online: https://ec.europa.eu/newsroom/article29/items/612053/en (accessed on 19 January 2022).
57. 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). 2016. Available online: https://eur-lex.europa.eu/legal-content/EN/TXT/PDF/?uri=CELEX:02016R0679-20160504 (accessed on 19 January 2022).

58. 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). Available online: https://eur-lex.europa.eu/legal-content/EN/TXT/?uri=CELEX:52020IP0275 (accessed on 19 January 2022).
59. High-Level Export Group on Artificial Intelligence. Ethics Guidelines for Trustworthy AI. Available online: https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai (accessed on 19 January 2022).
60. Charles, C.; Gafni, A.; Whelan, T. Decision-making in the physician–patient encounter: Revisiting the shared treatment decision-making model. *Soc. Sci. Med.* **1999**, *49*, 651–661. [CrossRef]
61. Makary, M.A.; Daniel, M. Medical error—The third leading cause of death in the US. *BMJ* **2016**, *353*, i2139. [CrossRef]
62. Obermeyer, Z.; Powers, B.; Vogeli, C.; Mullainathan, S. Dissecting racial bias in an algorithm used to manage the health of populations. *Science* **2019**, *366*, 447–453. [CrossRef]
63. Pourhomayoun, M.; Shakibi, M. Predicting mortality risk in patients with COVID-19 using artificial intelligence to help medical decision-making. *MedRxiv* **2020**. [CrossRef]
64. Renkl, A.; Hilbert, T.; Schworm, S. Example-Based Learning in Heuristic Domains: A Cognitive Load Theory Account. *Educ. Psychol. Rev.* **2009**, *21*, 67–78. [CrossRef]
65. 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 (IUI '19), Marina del Ray, CA, USA, 17–20 March 2019; pp. 258–262. [CrossRef]
66. 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. *Hum. Factors* **2013**, *55*, 520–534. [CrossRef]
67. Colquitt, J.A.; Rodell, J.B. Measuring Justice and Fairness. In *The Oxford Handbook of Justice in the Workplace*; Cropanzano, R.S., Ambrose, M.L., Eds.; Oxford University Press: Oxford, UK, 2015; pp. 187–202.
68. Schoeffer, J.; Machowski, Y.; Kuehl, N. Perceptions of Fairness and Trustworthiness Based on Explanations in Human vs. Automated Decision-Making. *arXiv* **2021**, arXiv:2109.05792
69. Zhou, J.; Chen, F.; Berry, A.; Reed, M.; Zhang, S.; Savage, S. A Survey on Ethical Principles of AI and Implementations. In Proceedings of the 2020 IEEE Symposium Series on Computational Intelligence (SSCI), Canberra, ACT, Australia, 1–4 December 2020; pp. 3010–3017.