Ethics and Information Technology (2022) 24:2
https://doi.org/10.1007/s10676-022-09623-4

ORIGINAL PAPER

Fairness, explainability and in-between: understanding the impact of different explanation methods on non-expert users' perceptions of fairness toward an algorithmic system

Avital Shulner-Tal[1] · **Tsvi Kuflik**[1] · **Doron Kliger**[2]

Accepted: 5 January 2022

Abstract
In light of the widespread use of algorithmic (intelligent) systems across numerous domains, there is an increasing awareness about the need to explain their underlying decision-making process and resulting outcomes. Since oftentimes these systems are being considered as black boxes, adding explanations to their outcomes may contribute to the perception of their transparency and, as a result, increase users' trust and fairness perception towards the system, regardless of its actual fairness, which can be measured using various fairness tests and measurements. Different explanation styles may have a different impact on users' perception of fairness towards the system and on their understanding of the outcome of the system. Hence, there is a need to understand how various explanation styles may impact non-expert users' perceptions of fairness and understanding of the system's outcome. In this study we aimed at fulfilling this need. We performed a between-subject user study in order to examine the effect of various explanation styles on users' fairness perception and understanding of the outcome. In the experiment we examined four known styles of textual explanations (case-based, demographic-based, input influence-based and sensitivity-based) along with a new style (certification-based) that reflect the results of an auditing process of the system. The results suggest that providing some kind of explanation contributes to users' understanding of the outcome and that some explanation styles are more beneficial than others. Moreover, while explanations provided by the system are important and can indeed enhance users' perception of fairness, their perception mainly depends on the outcome of the system. The results may shed light on one of the main problems in explainability of algorithmic systems, which is choosing the best explanation to promote users' fairness perception towards a particular system, with respect to the outcome of the system. The contribution of this study is reflected in the new and realistic case study that was examined, in the creation and evaluation of a new explanation style that can be used as the link between the actual (computational) fairness of the system and users' fairness perception and in the need of analyzing and evaluating explanations while taking into account the outcome of the system.

Keywords Fairness · Explainability · Algorithmic systems · Decision support systems · Users perception

Introduction

Transparency of algorithmic systems [systems that apply artificial intelligence (AI) and/or machine learning (ML) in their reasoning process] is becoming fundamental, since trust-related problems and the fairness of such systems are becoming a pressing issue (Arrieta et al., 2020; Došilović et al., 2018; Rai, 2020; Wang & Benbasat, 2007). Algorithmic transparency refers to the ability of users to understand the decision-making process and outcome of a system (Lipton, 2016). The ability of a system to explain its' reasoning and results ("explainability") may make it appear more transparent and interpretable to its users (Abdollahi & Nasraoui, 2018). Explanations may be required in order

✉ Avital Shulner-Tal
avitalshulner@gmail.com

Tsvi Kuflik
tsvikak@is.haifa.ac.il

Doron Kliger
kliger@econ.haifa.ac.il

[1] Department of Information Systems, University of Haifa, Haifa, Israel

[2] Department of Economics, University of Haifa, Haifa, Israel

Published online: 24 January 2022

to allow observers to understand the reasons for a decision made by an algorithmic system. Furthermore, explanations may also be requested by regulators due to the user's legal 'right to explanation' (e.g., users can demand explanations of decisions that were made for them by an algorithmic system) (Goodman & Flaxman, 2017; Zhang et al., 2019). Explanation, in turn, may increase users' trust in the system and, therefore, encourage them to consider it as a fair system (Arrieta et al., 2020; Lipton, 2016). Hence, explainability and transparency of a system may promote trustworthiness and increase the fairness perception of the users regardless to the actual (computational) fairness of the system (Ribeiro et al., 2016; Singh et al., 2018; Theodorou et al., 2017; Wortham et al., 2016).

While computational definitions of fairness and transparency are quite popular research topics these days, there is an understanding regarding the need to look at algorithmic systems in a wider perspective that refers also to their social implications (e.g., conforming to social norms, moral judgments, users' perceptions) (Barocas et al., 2018; Green, 2018; Ribeiro et al., 2016). This research is conducted in order to understand the impact of the explanations that are provided by an algorithmic system on non-experts' understanding and fairness perception of it. Our goal was to investigate whether and how explanations affect the users' fairness perception with respect to the outcome of the system. We aimed to do this by examining the differences between various textual explanations in terms of users' fairness evaluation of a system, and their understanding of the systems' outcome. In order to do so we conducted a between-subject experiment using a recruitment decision support system (DSS) as a case study. The results of the experiment, with respect to the outcome of the system, revealed differences between the explanation styles and may help in selecting the most appropriate explanation for such systems.

Explainability of algorithmic systems

This section provides some background and related work about explainability of algorithmic system. We start with definitions, so to have a common understanding of the meaning of the terms used and provide an overview of the main types of explanation techniques. Then we focus on the content and structure of black-box explanations (that are the focus of the study), we continue with black-box explanations styles and finally, review techniques used for explanations evaluation.

Definitions and background

In general, explainability refers to the ability to explain the decisions made by algorithmic systems to their users. The following distinct, yet interrelated, terminologies of explainability are used in the AI and ML communities:

Understandability (intelligibility) The degree to which a human is able to understand a decision made by a model. For example, explaining how the model works in a comprehensible way to humans, without explaining its internal structure (Arrieta et al., 2020; Montavon et al., 2018).
Comprehensibility The ability of humans to understand the learned knowledge in the algorithm (Arrieta et al., 2020; Craven, 1996; Fernandez et al., 2019; Gleicher, 2016). This usually refers to model complexity evaluation (Guidotti et al., 2018).
Interpretability The ability to provide explanations regarding system's reasoning process and outcomes in terms that humans will understand (Abdollahi & Nasraoui, 2018; Arrieta et al., 2020; Lipton, 2016; Rai, 2020).
Explainability The ability to create explanations which will be used as an interface between humans and the decision makers. The explanations should be both accurate to the decision-making process and comprehensible to humans (Guidotti et al., 2018).
Transparency The degree to which the model is understandable by itself, mainly refers to the characteristics of the model (Abdollahi & Nasraoui, 2018; Arrieta et al., 2020; Lipton, 2016).
Causability The degree to which an explanation achieves a specified level of causal understanding with effectiveness, efficiency and satisfaction in a specified context of use (Holzinger et al., 2019).

The above terminology is double-sided and includes the models' explanatory ability as well as the human's ability to understand the explanation (Arrieta et al., 2020). So, a suitable explanation can be created when there is a dialog between AI/ML experts and human experts (Gunning & Aha, 2019; Rudin, 2019). Therefore, users' perception should be taken into account when choosing the most appropriate explanation for a system (Gunning, 2017).

The purposes of explanations vary. One of the primary goals of explainability is to achieve transparency and as consequently establish the trustworthiness of the system (Abdollahi & Nasraoui, 2018; Arrieta et al., 2020; Felfernig & Gula, 2006; Kim et al., 2015; Ribeiro et al., 2016; Sinha & Swearingen, 2002). Both transparency and trustworthiness can improve users' perception of the system's, and may affect their willingness to use it (Abdollahi & Nasraoui, 2018; Griffin et al., 2017; Tintarev & Masthoff, 2011). Figure 1 presents the relations between explainability, transparency, trustworthiness and fairness perception (Fig. 1 was created as an extension to the diagram presented in (Abdollahi & Nasraoui, 2018), that presents the relation between explainability, transparency and fairness).

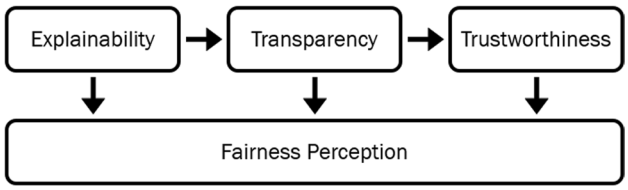

Fig. 1 Explainability, transparency, trustworthiness and fairness perception relations

A certain explanation may achieve many, though not necessarily all, goals. Moreover, for a given system, one style of explanation may be more efficient than another (Arya et al., 2019; Tintarev & Masthoff, 2011) and different purposes may be more important for different stakeholders (Arrieta et al., 2020).

We refer to two types of explainability of AI/ML systems: White-box explainability, and Black-box explainability. White-box explainability provides explanations for interpretable algorithms that reveal their structure: the algorithm, the reasoning process and the outcome can be easily explained by the model itself (Guidotti et al., 2018; Loyola-Gonzalez, 2019; Rudin, 2019; Tal et al., 2019). Black-box explainability, which is the focus of our study, provides explanations for intractable algorithms (deep learning, for instance) which are mathematically complicated and very hard to explain (Loyola-Gonzalez, 2019). Black-box explanations belong to the following two main categories (Guidotti et al., 2018): (i) model explanation—explaining the reasoning process of a black-box algorithm; and (ii) outcome explanation—explaining the relations between a particular input and its output, without explaining the reasoning process of the black-box model.

Black-box explanations refer to the ability to present a justified outcome to the user based on interpretable models that may be derived from the black-box model and, as a result, the decision may be more transparent to the user, and increase the fairness assessment of the system. Moreover, they are needed for verification of the system (experts' validation), improvement of the system (understanding the system's weaknesses and improving them), learning from the system (acquire new insights) and compliance with legislation (assignment of responsibilities in cases that the system is wrong) (Samek et al., 2017). In a way, black-box explanations fill an intention gap between system goals (accuracy) and users' needs and interests (understanding the model) (Tal et al., 2019).

Recent studies of black-box explainability aim at embedding explainability into the black-box model by integrating interpretability constraints into it or enrich the knowledge of the black box models with the knowledge of the transparent model (Adadi & Berrada, 2018; Arrieta et al., 2020; Gunning, 2017; Loyola-Gonzalez, 2019; Rai, 2020).

Content and structure of black-box explanations

The content of the explanations can be derived from the following four main groups (an explanation may contain information from more than one group) (Nunes & Jannach, 2017): (i) user preferences and inputs—provide information about the inputs that were provided to the system (e.g., decisive input values, users' preferences match, feature importance analysis, suitability estimate); (ii) decision inference process—provide information about the decision process of a specific problem or information about the logic of the system in general (e.g. Inference trace, inference and domain knowledge, decision method side-outcomes, self-reflective statistics); (iii) background and complementary information—provide additional background information regarding the system (e.g., knowledge about peers, knowledge about similar alternatives, relationship between knowledge objects, background data to the current problem instance, knowledge about the community); and (iv) alternatives and their features—provide analysis of the features of alternative outputs (e.g., decisive features, pros and cons, feature-based domination, irrelevant features). There are two perspectives regarding the scope of the explanation (Eiband et al., 2018a, 2018b): (i) a normative point of view which suggests that systems may only be used if their explanations are adequate and hence motivate to provide detailed and comprehensive explanations, and (ii) a pragmatic point of view which motivates less detailed explanations in order to enable a basic understanding of the system by integrating small explanations into the interface of the system.

Eiband et al. (2018a, 2018b, March) suggested a five-stage explanation model for integrating transparency into algorithmic systems. Their model deals with the content of the explanation (stages 1–3) and its presentation (stages 4–5). The first stage (Expert Mental Model) is used for gaining common understanding about data collection and processing methods; the second (User Mental Model) deals with the users' beliefs about the system logic and its transparency in order to create a list of differences between the user and expert mental models; the third (Target Mental Model) deals with the trade-off between transparency (providing more information) and the visual or cognitive load which can be formed; the fourth (Iterative Prototyping) aims to create several prototypes for integrating the explanations into an existing UI; and the fifth stage (Design Evaluation) deals with evaluating of the different prototypes with respect to design changed that can improve users' mental model.

Hence, human-in-the-loop psychological experiments are necessary and essential for evaluating the effectiveness of the explanations, since we need to consider users' satisfaction (users' rating regarding the clarity and utility of the explanation), users' mental model (their understanding regarding the reasoning process), users' task performance

(whether the explanation improves users' decisions) and users' future use and trust assessment (Gunning, 2017).

Styles of black-box explanation

Black-box explanations may be presented in various techniques such as textual explanations (using natural language to explain the reasoning process and/or outcome) and/or visual explanations (using graphs and pictures to explain the reasoning process and/or outcome) (Arrieta et al., 2020; Lipton, 2016; Ribeiro et al., 2016). When referring to textual explanations, there are four styles of black-box explanations that can be used for enhancing the fairness perception of the users (Binns et al., 2018; Dodge et al., 2019):

Case-based Explanation presents a case from the model's training data which is most similar to the users.
Demographic-based Explanation presents aggregate demographic statistics, such as age, gender, income level or occupation, regarding the structure of the training data and/or the distribution of the outcome.
Input Influence-based Explanation presents the influence of various input features (Decisive input values) on the decision by quantitative measures.
Sensitivity-based Explanation presents sensitivity analysis that shows how various changes in the input features values will modify the outcome.

According to (Dodge et al., 2019), sensitivity-based and case-based explanations are considered as local explanations of the system (outcome explanations) which try to justify why a particular outcome was received for a specific case, while input influence-based and demographic-based explanations are considered as global explanations (model explanations) which describe the model of the system and how it works. They found that global explanations enhance the perceived fairness of users who intend to trust algorithmic systems, while local explanations are more effective in revealing fairness perception differences among different cases and that users' prior trust level in such systems impact their reaction to the explanations.

Furthermore, Binns et al. (2018) suggested that users' fairness perception may be enhanced only when multiple explanation styles (combining global and local explanations) are presented to the users and that the scenario effects (e.g., system's domain, results, usage) outweigh the explanation effect in cases of one explanation style. Both (Binns et al., 2018) and (Dodge et al., 2019) claimed that in most cases the case-based explanation is perceived as less fair then other explanation styles and that there is a need to continue and examine the effect of explanation on fairness perceptions.

Design and evaluation of explanation

Designing an effective explanation requires a lot of efforts. The following four possible guidelines may help in designing or selecting the most suitable explanation for a system (Gedikli et al., 2014): (i) use domain specific content in order to increase the effectiveness of the explanation; (ii) use familiar explanation concepts in order to reduce users' cognitive effort; (iii) increase transparency through explanations in order to increase user satisfaction; and (iv) there is no need to adjust the explanation for efficiency.

After the explanation was designed, there is a need to evaluate its' effectiveness. Jesus et al. (2021) proposed XAI test, an evaluation methodology for measuring the utility of the explanations. They examined the following three outlines using three professional fraud analysts: (i) only transaction data; (ii) transaction data and ML model Score; and (iii) transaction data, ML model Score and explanation (explanation were created from LIME, SHAP and TreeInterpreter). They evaluated the user's perception of the explanation quality, with the following three statements:

1. The explanation covered all the relevant information to help me make a decision.
2. The explanation helped me decide faster.
3. The explanation was useful to help me make a decision.

They found that the relevance, usefulness, and diversity of the explanation are perceived differently by end users.

Another method for evaluating explanations was presented by Holzinger et al. (2020). They suggested System Causability Scale (SCS) which is based on System Usability Scale (SUS). They argued that this method, can help quickly determine to what extant an explanation is suitable for an intended purpose. They used the following ten statements for evaluating explanations in the medical domain:

1. Factors in data—I found that the data included all relevant known causal factors with sufficient precision and granularity.
2. Understood—I understood the explanations within the context of my work.
3. Change detail level—I could change the level of detail on demand.
4. Need teacher/support—I did not need support to understand the explanations.
5. Understanding causality—I found the explanations helped me to understand causality.
6. Use with knowledge—I was able to use the explanations with my knowledge base.
7. No inconsistencies—I did not find inconsistencies between explanations.

8. Learn to understand—I think that most people would learn to understand the explanations very quickly.
9. Needs references—I did not need more references in the explanations: e.g., medical guidelines, regulations.
10. Efficient—I received the explanations in a timely and efficient manner.

Methodology

As described above, explanations may be used to increase transparency, trustworthiness, and the perception of fairness of algorithmic systems (Abdollahi & Nasraoui, 2018; Felfernig & Gula, 2006; Griffin et al., 2017; Kim et al., 2015; Ribeiro et al., 2016; Sinha & Swearingen, 2002; Tintarev & Masthoff, 2011). Therefore, in this study we aimed to investigate whether and how explanations affect users' fairness perception regarding an algorithmic system by evaluating various textual explanation styles using a between-subject user study. The experiment was conducted for examining the following question:

How do different explanations impact non-expert users' perceptions of fairness toward the system?

This question can be divided into the following two sub questions:

(**Q1**) How do the various styles of explanations affect users' fairness evaluation of the system?
(**Q2**) How do the various styles of explanations affect users' understanding of the system's outcome?

We tested the following Null hypotheses, with respect to the research questions:

H1 The fairness evaluation of the users is similar among the various explanation styles

H2 The understanding of the outputs is similar among the various explanation styles.

We followed the methodology of (Binns et al., 2018; Dodge et al., 2019; Lundberg & Lee, 2017) and performed a between-subject user study, in order to examine the differences between the various styles of textual explanations in terms of their effect on users' fairness perceptions (captured by the users' fairness evaluations of a system), and the explanations' effect on users' understanding of system's outcome. We analyze the results with respect to the outcome of the system. The between-subject design was chosen in order to distinguish between the explanation styles and to avoid carryover effects. We used a realistic recruitment DSS that recommends whether to hire a candidate or not. The recruitment field was chosen because it is a familiar field and most people experienced a recruitment process during their lifetime.

We choose to examine the following four basic explanation styles, as suggested by (Binns et al., 2018; Dodge et al., 2019) (presented in "Styles of black-box explanation" section): Case-based, Demographic-based, Input Influence-based and Sensitivity-based. We also added a fifth explanation style: Certification-based which presents the results of an invited auditing process of the system [A regulator or certification authority that decides whether to certify an algorithmic system as fair, based on the auditing results (Kilbertus et al., 2018)]. The purpose of this style of explanation is to bridge the gap between the actual (computational) fairness that can be examined in the auditing process of the system and users' fairness perception, since we do not want to mislead users to think that a system is fair while it is not. In addition, this explanation style may increase the fairness perception of users even without revealing any details about the system decision making process and without hampering the accuracy of the system.

Participants

The experiment was conducted using Amazon's Mechanical Turk crowdsourcing platform. Although it is difficult to control the background of the participants when using crowdsourcing platforms, crowdsourcing enables to collect real and anonymous inputs (Van Berkel et al., 2019).

We ran our study on August 20, 2020. All the participants were native English speakers, above 18 years old and had HIT approval rate of at least 98% and at least 5000 HITs completed. A total of 600 participants took part in the experiment; 72 of them were excluded due to failing an attention check (a simple question in which the participants were asked to choose a specific answer), and 103 due to deviation from execution times ranging from 60 to 600 s (participants' time for conducting the experiment was around 300 s, the bottom 5% of the participants with the shortest times and the top 5% of the participants with the longest times were excluded). No participant participated more than once. The demographic characteristics (gender and age) distribution of the remaining 425 participants is presented in Table 1.

Method

In the experiment, each participant received a case study with a description of a recruitment DSS, followed by the description of the candidate that the system had to determine whether to recommend for hiring. The participants were

Table 1 Demographic distribution of the participants

Gender	# Participants	Age	# Participants
Female	186	18–25	62
Male	238	25–35	156
Unknown	1	35–45	103
		45 +	102
		Unknown	2

informed that there are only a few open job positions and a lot of candidates. Then the participant was presented with the output of the DSS which included the recommendation for the candidate (positive recommendation—the candidate is recommended by the system, or negative recommendation—the candidate is not recommended by the system) and one of five studied explanations. We did not examine the option in which there is no explanation at all because we wanted to address only cases where some explanation is provided. Furthermore, comparing cases with an explanation versus a case where there is no explanation at all may add additional noise since it is related to the effect of the existence of the explanation and not necessarily to the style of explanation itself.

We chose to use an average candidate (average graduate student, some relevant professional experience, good communication skills, good recommendation letters) description in our experiment since it is unlikely to present an outcome of the system that contradicts what may be considered a logical outcome (e.g. successful candidate can be linked to an acceptance for a job and an unsuccessful candidate can be linked to a rejection for a job). The characteristics of the candidate were slightly positive, hence we assumed that both system's decisions, to accept or reject the candidate will be considered reasonable and allow us to focus better on the impact of the explanations.

The system's description, selected candidate's description and explanation styles that were used in the experiment are presented in Appendix 1.

Explanation quality evaluation

The explanations that were used in the experiment were created in accordance with the explanations presented in (Binns et al., 2018; Dodge et al., 2019) and according to the guidelines in (Gedikli et al., 2014). Their quality was evaluated using the System Causability Scale (SCS) (Holzinger et al., 2020). In this evaluation, participants were presented with the recruitment DSS system description, the candidate description and one explanation style and were requested to rank the SCS statement according to a 5-point Likert scale (strongly agree to strongly disagree). The evaluation was conducted using Amazon Mechanical Turk as well. 205 participants took part in the evaluation, 9 of them were excluded from the sample due to failing an attention check (one statement in which the participant were requested to choose a specific score). Than an average ranking score was calculated for each statement with respect to the explanation style.

The results of this evaluation and the distribution of the explanation across the participant are presented in Table 2. The results suggest that the quality of the explanations is above average (score of 0.68 or higher), except from the certification-based explanation which is slightly below average. This is reasonable since the certification-based explanation doesn't reveal any information regarding the reasoning process of the system.

Table 2 Quality evaluation results

System Causability Scale (SCS) statements	Case-based	Certification-based	Demographic-based	Input influence-based	Sensitivity-based
# Participants	42	38	33	40	43
Factors in data	3.476	3.263	3.727	3.475	3.512
Understood	3.762	3.605	3.667	3.500	3.605
Change detail level	3.595	3.447	3.727	3.325	3.455
Need teacher/support	3.548	3.421	3.515	3.375	3.532
Understanding causality	3.667	3.184	3.606	3.650	3.721
Use with knowledge	3.786	3.474	3.515	3.625	3.644
No inconsistencies	3.476	3.079	3.394	3.250	3.605
Learn to understand	3.595	3.421	3.636	3.800	3.698
Needs references	3.476	3.079	3.424	3.325	3.512
Efficient	3.786	3.342	3.606	3.700	3.744
SCS Score	0.723	0.666	0.716	0.701	0.721

Between subject experiment

Taking into account the above research questions and their respective hypotheses, we noted that differences between two scenarios—when the system provides a "negative" recommendation and when the system provides a "positive" recommendation. Hence, we distinguish between the two system's outcomes:

(1) *Negative recommendation* The candidate is not recommended by the system
(2) *Positive recommendation* The candidate is recommended by the system.

The effectiveness of the explanation styles, with respect to the recommendation output, was examined in two aspects:

(1) The *fairness* evaluation of the users (in order to examine **H1**)
(2) The level of *understanding* of the outcome (in order to examine **H2**).

In order to examine the aspect of users' fairness evaluation, the users were requested to answer the following question: "What is your view about the fairness of the system?" Answers were given on a 6-point Likert scale, ranging from "extremely fair", represented as (3), to "extremely unfair", represented as (− 3). We excluded the option of "neither fair or unfair" (represented as 0) from the scale. In order to examine the participants understanding level, they were requested to answer the following question: "To what extent the explanation helps you to understand the output of the system?" Answers were given on a 5-point Likert scale from "much better", represented as (2), to "much worse", represented as (− 2). The joint distribution of the recommendations and the explanations among the participants in the filtered sample is presented in Table 3.

We used Non-Parametric Kruskal–Wallis H test (analogue to one-way ANOVA), followed by post-hoc pairwise comparison with Bonferroni correction for multiple comparisons in order to examine the above Null hypotheses.

Table 3 Recommendation and Explanation Distribution

Explanation style	Negative recommendation	Positive recommendation	Total
Case-based	43	44	87
Demographic-based	41	45	86
Input influence-based	44	44	88
Sensitivity-based	37	42	79
Certification-based	39	46	85
Total	204	221	425

Results

We start by showing the results for the cases of *negative recommendation* (where the candidate was not recommended by the system). The descriptive statistics results, presented in Table 4, show that the sensitivity-based explanation was judged by the participants as the best in both *fairness* and *understanding* aspects, with respective average scores of − 0.027 and 0.865. The certification-based explanation scored lowest for *fairness*, with an average score of − 0.641. The case-based and certification-based explanations scored lowest for *understanding*, with an average score of 0.023 and 0.025 respectively. Furthermore, the average scores for users' *fairness* evaluation of the system was negative in all explanation styles and the *understanding* level of the output was positive in all explanation styles.

The results of the Kruskal–Wallis test, with post-hoc pairwise comparisons, in cases of negative recommendation, indicated the following (presented in the top part of Table 6):

- There is no significant difference in fairness evaluation among the different explanation styles [H (9.49) = 2.426, P value > 0.05], therefore we cannot reject **H1**.
- There is a significant difference in output's understanding among the different explanations [H (9.49) = 17.759, P-value < 0.05], therefore we reject **H2**.

The post-hoc pairwise comparisons with Bonferroni correction results shows that for understanding, there is a significant difference between certification-based explanation and sensitivity-based explanation and between case-based explanation and sensitivity-based explanation (Adj. P-value < 0.05).

We now turn to the analysis of the cases of *positive recommendation*. The descriptive statistics results, presented in Table 5, show that certification-based explanation scored highest in *fairness*, with average scores of 1.565, and

Table 4 Fairness and Understanding scores in cases of *Negative Recommendation* (the candidate was not recommended by the system)

Explanation style	Fairness score AVG (STD)	Understanding score AVG (STD)
Case-based	− 0.465 (1.856)	0.023 (1.113)
Demographic-based	− 0.268 (2.122)	0.463 (1.051)
Input influence-based	− 0.409 (1.884)	0.227 (1.138)
Sensitivity-based	**− 0.027 (1.803)**	**0.865 (0.918)**
Certification-based	− 0.641 (1.940)	0.025 (1.000)

The sign (±) represent the positive/ negative affect of the explanation. The higher the result, the more positive the effect. Best result for each aspect are bolded

Table 5 Fairness and Understanding scores in cases of *Positive Recommendation* (the candidate was recommended by the system)

Explanation style	Fairness score AVG (STD)	Understanding score AVG (STD)
Case-based	1.159 (1.697)	0.795 (0.734)
Demographic-based	− 0.356 (1.885)	0.133 (1.217)
Input influence-based	1.205 (1.636)	0.909 (0.802)
Sensitivity-based	1.167 (1.464)	**0.976 (0.924)**
Certification-based	**1.565 (1.470)**	0.804 (0.885)

The sign (±) represent the positive/ negative affect of the explanation. The higher the result, the more positive the effect. Best result for each aspect are bolded

sensitivity-based explanation scored highest in *understanding* with an average score of 0.976. The demographic-based explanation scored lowest, with average scores of − 0.356 and 0.133 for *fairness* and *understanding*. Furthermore, the *understanding* of the output was positive in all explanation styles while the *fairness* evaluation of the system was positive in all explanation styles, except for demographic-based explanation, in which the *fairness* evaluation was negative.

The results of the Kruskal–Wallis tests with post-hoc pairwise comparisons, in cases of positive recommendation, indicated the following (presented in the bottom part of Table 6):

- There is a significant difference in fairness evaluation among the different explanation styles [H (9.49) = 29.214, P-value < 0.05], therefore we reject H1. The post-hoc pairwise comparisons with Bonferroni correction results show that there is a significant difference between demographic-based explanation and all the other explanation styles (Adj. P-value < 0.05).
- There is a significant difference in output's understanding among the different explanations [H (9.49) = 17.464, P-value < 0.05], therefore we reject H2. The post-hoc pairwise comparisons with Bonferroni correction results show that there is a significant difference between demographic-based explanation and all the other explanation styles except of case-based explanation (Adj. P-value < 0.05).

According to the results, we can argue that in cases of *negative recommendation* (the candidate is not recommended by the system) the system was considered as unfair no matter what was the explanation, and still the explanations helped in understanding the output of the system. Furthermore, it is interesting to note that our findings show that it doesn't really matter which explanation to present since no significant differences were found in terms of *fairness* evaluation, but the sensitivity-based explanation achieved higher scores than the case based-explanation and the certification-based explanation in terms of *understanding* the output of the system.

On the other hand, in cases of *positive recommendation* (the candidate is recommended by the system) the results were different. The system was considered as fair, except when the demographic-based explanation was provided and all explanations helped in understanding the output of the system. Moreover, our findings show that any style of explanation is better (achieved significantly higher scores) than the demographic-based explanation in both *fairness* and *understanding* aspects. It is also interesting to note that even though their difference is not statistically significant, the certification-based explanation scored considerably higher in fairness and understanding.

Figure 2 presents the *fairness* and *understanding* scores of the various explanation styles for both cases of negative recommendation and *positive recommendation*.

Table 6 Significant differences according to Kruskal–Wallis H tests with post-hoc pairwise comparisons

Outcome	Parameter	Pairwise comparisons (Bonferroni correction)	Adj. Sig
Negative recommendation	Understanding	Certification—Sensitivity	0.004
		Case—Sensitivity	0.004
Positive recommendation	Fairness	Demographic—Sensitivity	0.007
		Demographic—Input	0.001
		Demographic—Case	0.001
		Demographic—Certification	0.000
	Understanding	Demographic—Sensitivity	0.001
		Demographic—Input	0.013
		Demographic—Certification	0.045

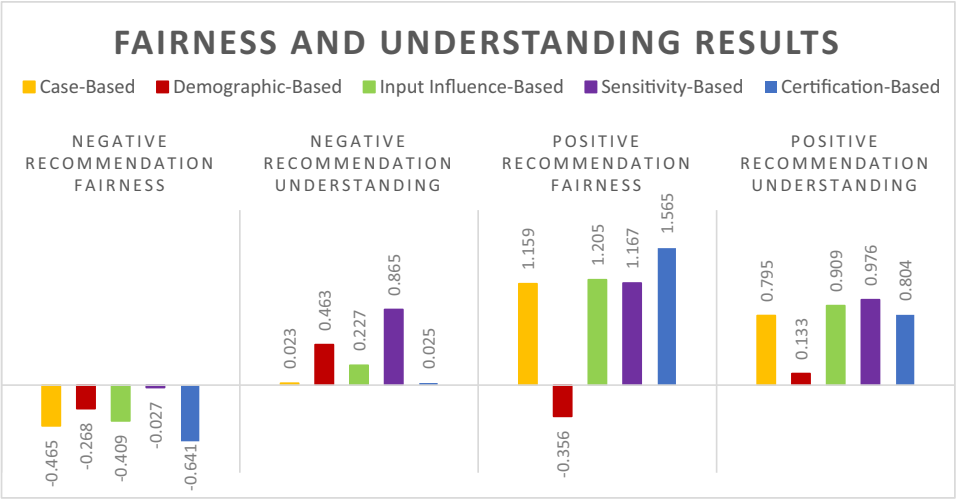

Fig. 2 Fairness and Understanding scores. Fairness and understanding scores in cases of negative *recommendation* are presented in the left part of the graph. Fairness and understanding scores in cases of *positive recommendation* are presented in the right part of the graph. The colors represent the different explanation styles

Discussion

This study aimed to explore the effect of various explanation styles on users' fairness perception regarding a system and their understanding of the outcome. Our main contributions lie in the examination of a realistic DSS case study, in the creation and evaluation of a new explanation style and, in the examination of the various explanation styles with respect to the outcome of the system. In this study, we devised a recruitment DSS. We have chosen to deal with recruitment decisions because most people have experienced recruitment processes during their lifetime. Moreover, alongside to four known explanation styles, we created and evaluated a new explanation style (certification-based explanation), aimed at increasing the fairness perception of users without revealing any details about the system decision-making process and without hampering the accuracy of the system. This new type of explanation requires an execution of an auditing process and can be used as the connecting thread between the actual (computational) fairness of the system and the perceived fairness of the system. The results of our experiment suggest that although the explanation of the system is important, the perceived fairness is mainly affected by the outcome of the system.

Accordingly, we argue that: (i) when the system produces a "negative" output, it may be considered unfair, no matter what explanation is provided (**H1** not rejected for "negative" output); (ii) when the system produces a "positive" output, it may be considered fair and every explanation style (except for demographic-based) works, with a slight advantage to our new certification-based explanation (**H1** was rejected for "positive" output). Therefore, we assert that users' *fairness* perception regarding the system is affected mainly by the output of the system and that the explanation may not be relevant when the output is negative; lastly, (iii) the explanations help in *understanding* the output of the system in both "negative" and "positive" outputs (**H2** was rejected) and sensitivity-based explanations are probably the best explanation style for increasing the understanding level.

The results confirm the claim made in (Binns et al., 2018) that suggested that the scenario effects (e.g. system's results) outweigh the explanation effect in cases when one explanation style is presented. For example, it is possible that the output of the system ("positive" or "negative") outweighed the effect of the explanation in cases where the users assumed that the candidate may fit the position and expected that the candidate will be recommended and when the candidate was not recommended, they felt that there is no correlation between the input (candidate description) and the output of the system. However, according to our results, selecting the appropriate explanation style to present is useful especially in cases of positive output. Furthermore, our results are partially contradicting the results of (Binns et al., 2018) and (Dodge et al., 2019) that claimed that in most cases the case-based explanation is perceived as less fair then other explanation styles, since we found that users' fairness perception mainly depends on the outcome of the system and that case-based explanation was perceived as less fair then sensitivity-based explanation only in cases when the system provide "negative" outcome, while in cases when the system provide "positive" outcome the explanation style that is perceived as less fair then others is the demographic-based explanation.

Another interesting issue is that the certification-based explanation scored highest average score in *fairness* aspect in cases of "positive" recommendation. The uniqueness of this style is that it satisfies the users' need for an explanation regarding the system on the one hand, and on the other hand, it is a generic style that does not provide an explanation of the reasoning process or the outcome of the system. This of course may help companies that are not interested in disclosing the decision-making process of the system, but it

will require them to carry out an auditing process (e.g., like "VeriSign" certificate of online payments) in order to obtain the required fairness certification.

As any study, this study has limitations. The scenario could have an impact on the fairness perception of the users. First of all, the explanation styles were examined for a specific input (slightly positive candidate description), and it is possible that different results will be obtained if different candidate description will be presented (e.g., a really successful candidate that may be rejected or a really unsuccessful candidate that may be accepted). Such cases may change users' expectation towards the output of the system and then the users' may consider the system as fairer or not. Second, the experiment was performed in a specific domain (job recommendation) and other results may be obtained when the explanation styles will be examined for systems with different levels of impact on our life (e.g., medical DSS or legal DSS). Another limitation that needs to be considered is that the participants' fairness and understanding assessments are based on self-report and there aren't any personal stakes on their behalf, so, it is possible that the results do not reflect their true assessments in case they will encounter such system in real-life. Additionally, it is difficult to control the background of the participants when using crowdsourcing platforms. Hence, the results may contain noise due to the use of Amazon Mechanical Turk.

Conclusions and future work

In the current study, we conducted a human-in-the-loop experiment in order to examine the relationship between users' fairness perception and various explanation styles. Three main conclusions may be drawn from this study: (i) explanations effect the fairness perception of a DSS, though it is mainly determined based on the output of the system (positive/negative); (ii) it is not important which explanation will be presented when the user receives a "negative" output, however, this requires further investigation, given the limitation of the study; and (iii) explanations increase users' level of understanding of the output. The results of this study highlight the need for performing human-in-the-loop experiments for evaluating and improving the explainability, transparency and fairness of algorithmic system.

Although selecting the most appropriate explanation for a system is a difficult task, the results of this study may help in assessing the best explanation that can be presented with respect to the outcome of the system. Of course, there is a need for further examination of this issues (e.g., examine users' fairness perception when a combination of textual explanation styles is presented and examine visual explanations). Additionally, we suggested and examined a new explanation style (certification-based explanation), which was found to be as effective as the other explanation styles in all the parameters that were examined in this study. It will be interesting to further examine the certification-based explanation, since, this explanation style may fulfill users' need for explanation without revealing the reasoning process of the system. The certification-based explanation describes the actual fairness of the system in a general and understandable way and may be used only if the system was verified as a fair system. Hence, the use of this type of explanation will not mislead users into thinking that the system is fair when it is actually not.

Finding which explanation is better to explain the outcome of an AI model is a challenging task, and further studies in this field will be beneficial. Noteworthy, additional research may shed further light on the topic. In future studies we aim to extend the current experiment and explore more inputs (candidate descriptions), more outputs ("strongly recommend" and "strongly not recommend"). Additional directions may be examining other explanation styles, both textual and visual, that are presented in real-life systems, as well as evaluating the effectiveness of the explanation styles with domain experts and comparing their fairness perception with non-experts' fairness perception. It would also be interesting to observe the different results in fairness and understanding if each participant had access to all of the explanation styles provided, as well as to examine the relationship between explanation styles and users' fairness perception on different domain specific DSS (e.g. legal or medical DSSs) and with various input–output correlations.

Appendix 1

System's description

"CANRA.Inc" is an intelligent DSS that uses AI and ML techniques to predict the likelihood of a candidate succeeding in a new job. The system recommends to recruiters and others in the HR system whether or not to recruit a candidate.

The system receives the candidate's CV, rating of the university, class rank at the university (student's performance compared to other students in her/his graduating class), relevant experience, personality test results, recommendation letters from former employers and a brief summary of an internal interview with the company's interviewer.

The system then produces a recommendation score (Strongly not recommend/Not recommend/Neutral/Recommend/Strongly recommend) for hiring the candidate as well as an explanation letter explaining the output of the system.

Candidate's description

The data of the following candidate was inserted to the system:

The candidate is an average graduate student (ranked 48th out of 103 students in the class). The candidate worked and did voluntary service while studying. The candidate was appreciated by co-workers in both places.

The internal interviewer's impression:

– The candidate has relevant experience for the position.
– My impression from the candidate's recommendation letters from former employers is that the candidate fulfills the job responsibilities as required.
– My impression from the internal interview is that the candidate has good communication skills.

Interviewer's recommendation: We may consider proceeding with this candidate.

Explanations descriptions

The following explanations were used in the experiment:

Case-based explanation

A similar case (which received the same outcome) is the following candidate: "The candidate was an average performing student with some relevant experience for the job, S/he was positively recommended by her/his co-workers and fulfills her/his job responsibilities as required. The candidate had a similar CV to yours and the personality test results were also similar."

Certification-based explanation

The system was tested and verified by authorized experts and regulators for fairness towards different population segments guarding against biases and discrimination. It was found to satisfy the required fairness constraints.

Demographic-based explanation

The outputs are distributed in a normal distribution. Furthermore, it is known that:

- 17% of candidates who are ranked in the top 10% in their graduating class are positively recommended by the system.
- 36% of candidates with 10 years of relevant experience are negatively recommended by the system.
- 28% of candidates with good communication skills in the internal interview are negatively recommended by the system.
- 41% of candidates who were appreciated by former employers are negatively recommended by the system.

Input influence-based explanation

Our predictive model assessed the candidate's information in order to predict his/her chances of progressing in the recruitment process. The more + signs or—signs, the more positively or negatively that factor impacted the probability of being recommended. Unimportant factors are not indicated. The following features and their impact on the outcome for this particular candidate are:

- Rating of the university (++).
- Candidate's ranking in the university (+).
- Candidate's CV (+).
- Candidate's personality test results (−).
- Candidate's experience (+++).
- Candidate's recommendation letters (−−).
- Internal interviewer's recommendation (++).

Sensitivity-based explanation

Our predictive model The following changes in the input features will change the outcome of the system:

- If this candidate were to be ranked in the top 10 percent of her/his graduating class—the likelihood of positive recommendation by the system would be increased by 23%.
- If this candidate had another year of relevant experience to this job—the likelihood of positive recommendation by the system would be increased by 34%.
- If this candidate had better communication skills in the internal interview—the likelihood of positive recommendation by the system would be increased by 15%.
- 12% of candidates who were recommended by the internal interviewer are positively recommended by the system.

Acknowledgements Partial financial support was received from the Cyprus Center for Algorithmic Transparency, which has received funding from the European Union's Horizon 2020 Research and Innovation Program under Grant Agreement No. 810105 (CyCAT—Call: H2020-WIDESPREAD-05-2017-Twinning), by a scholarship program for doctoral students in High-Tech professions at the University of Haifa, Israel and by Data Science Research Center (DSRC) at the University of Haifa, Israel.

Author contributions All authors contributed equally to the study conception and design. Material preparation, data collection and analysis were performed by AS-T. The first draft of the manuscript was written by AS-T and all authors commented on previous versions of the manuscript. All authors read, reviewed and commented on interim versions of the paper until the final manuscript was submitted.

Funding Partial financial support was received from the Cyprus Center for Algorithmic Transparency, which has received funding from the European Union's Horizon 2020 Research and Innovation Program under Grant Agreement No. 810105 (CyCAT—Call: H2020-WIDESPREAD-05-2017-Twinning), by a scholarship program for doctoral students in High-Tech professions at the University of Haifa, Israel and by Data Science Research Center (DSRC) at the University of Haifa, Israel.

Data availability The datasets generated during and/or analyzed during the current study are available from the corresponding author on request.

Code availability Not applicable.

Declarations

Conflict of interest The authors have no conflict of interest to declare that are relevant to the content of this article.

Ethical approval The experiments conducted in this study were approved by the Committee for Ethical Research and the Protection of Human Participants, University of Haifa, Israel (Approval 350/19).

Consent to participate The following consent to take part in academic research was presented to the participants: "This research is conducted by researchers from the Departments of Information Systems and Economics at the University of Haifa, which deals with the transparency and fairness of algorithmic systems. We request your participation in this online study. It should be emphasized that the answers to the questionnaires will be kept confidential and used only for research purposes. No personal or identifying information is requested or kept. Your participation in this study is voluntary. If you decide at any time that you do not wish to participate, you may do so without penalty. This research is approved by Committee for Ethical Research and the Protection of Human Participants, University of Haifa: (350/19) Thank you in advance for your cooperation."

Consent for publication This work has not been published before; it is not under consideration for publication anywhere else; its publication has been approved by all co-authors.

References

Abdollahi, B., & Nasraoui, O. (2018). Transparency in fair machine learning: The case of explainable recommender systems. *Human and Machine Learning* (pp. 21–35). Springer.

Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). *IEEE Access, 6*, 52138–52160.

Arrieta, A. B., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., ... Chatila, R. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. Information Fusion, 58, 82-115.

Arya, V., Bellamy, R. K., Chen, P. Y., Dhurandhar, A., Hind, M., Hoffman, S. C., ... Mourad, S. (2019). One explanation does not fit all: A toolkit and taxonomy of AI explainability techniques. arXiv preprint arXiv:1909.03012

Barocas, S., Hardt, M., & Narayanan, A. (2018). Fairness and Machine Learning. fairmlbook. org. Retrieved from http://www.fairmlbook.org

Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., & Shadbolt, N. (2018, April). 'It's reducing a human being to a percentage' perceptions of justice in algorithmic decisions. In Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems (pp. 1–14).

Craven, M. W. (1996). *Extracting comprehensible models from trained neural networks*. University of Wisconsin-Madison Department of Computer Sciences.

Dodge, J., Liao, Q. V., Zhang, Y., Bellamy, R. K., & Dugan, C. (2019, March). Explaining models: an empirical study of how explanations impact fairness judgment. In Proceedings of the 24th International Conference on Intelligent User Interfaces (pp. 275–285).

Došilović, F. K., Brčić, M., & Hlupić, N. (2018). Explainable artificial intelligence: A survey. *2018 41st International convention on information and communication technology, electronics and microelectronics (MIPRO)* (pp. 0210–0215). IEEE.

Eiband, M., Schneider, H., & Buschek, D. (2018). Normative vs. Pragmatic: Two perspectives on the design of explanations in intelligent systems. In: IUI workshops on explainable smart systems (EXSS)

Eiband, M., Schneider, H., Bilandzic, M., Fazekas-Con, J., Haug, M., & Hussmann, H. (2018). Bringing transparency design into practice. *23rd International conference on intelligent user interfaces* (pp. 211–223). ACM.

Felfernig, A., Gula, B. (2006). Consumer behavior in the interaction with knowledge-based recommender applications. In: ECAI 2006 workshop on recommender systems, pp. 37–41

Fernandez, A., Herrera, F., Cordon, O., del Jesus, M. J., & Marcelloni, F. (2019). Evolutionary fuzzy systems for explainable artificial intelligence: Why, when, what for, and where to? *IEEE Computational Intelligence Magazine, 14*(1), 69–81.

Gedikli, F., Jannach, D., & Ge, M. (2014). How should I explain? A comparison of different explanation types for recommender systems. *International Journal of Human-Computer Studies, 72*(4), 367–382.

Gleicher, M. (2016). A framework for considering comprehensibility in modeling. *Big Data, 4*(2), 75–88.

Goodman, B., & Flaxman, S. (2017). European Union regulations on algorithmic decision-making and a "right to explanation." *AI Magazine, 38*(3), 50–57.

Green, B. (2018). "Fair" risk assessments: A precarious approach for criminal justice reform. In: 5th Workshop on fairness, accountability, and transparency in machine learning.

Griffin, R. W., Phillips, J., & Gully, S. M. (2017). Organizational behavior: Managing people and organizations.

Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2018). A survey of methods for explaining black box models. *ACM Computing Surveys (CSUR), 51*(5), 1–42.

Gunning, D. (2017). Explainable artificial intelligence (xai). Defense Advanced Research Projects Agency (DARPA), nd Web, 2, 2.

Gunning, D., & Aha, D. W. (2019). DARPA's explainable artificial intelligence program. *AI Magazine, 40*(2), 44–58.

Holzinger, A., Carrington, A., & Müller, H. (2020). *Measuring the quality of explanations: The system causability scale (SCS)* (pp. 1–6). KI-Künstliche Intelligenz.

Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Müller, H. (2019). Causability and explainability of artificial intelligence in medicine. *Wiley Interdisciplinary Reviews: Data Mining and Knowledge Discovery, 9*(4), e1312.

Jesus, S., Belém, C., Balayan, V., Bento, J., Saleiro, P., Bizarro, P., & Gama, J. (2021). How can I choose an explainer? An Application-grounded Evaluation of Post-hoc Explanations. arXiv preprint arXiv:2101.08758

Kilbertus, N., Gascón, A., Kusner, M. J., Veale, M., Gummadi, K. P., & Weller, A. (2018). Blind justice: Fairness with encrypted sensitive attributes. arXiv preprint arXiv:1806.03281

Kim, B., Glassman, E., Johnson, B., & Shah, J. (2015). iBCM: Interactive Bayesian case model empowering humans via intuitive interaction.

Lipton, Z. C. (2016). The mythos of model interpretability. arXiv preprint arXiv:1606.03490

Loyola-Gonzalez, O. (2019). Black-box vs. white-box: Understanding their advantages and weaknesses from a practical point of view. *IEEE Access, 7*, 154096–154113.

Lundberg, S. M., & Lee, S. I. (2017). A unified approach to interpreting model predictions. In: Advances in neural information processing systems (pp. 4765–4774).

Montavon, G., Samek, W., & Müller, K. R. (2018). Methods for interpreting and understanding deep neural networks. *Digital Signal Processing, 73*, 1–15.

Nunes, I., & Jannach, D. (2017). A systematic review and taxonomy of explanations in decision support and recommender systems. *User Modeling and User-Adapted Interaction, 27*(3–5), 393–444.

Rai, A. (2020). Explainable AI: From black box to glass box. *Journal of the Academy of Marketing Science, 48*(1), 137–141.

Ribeiro, M. T., Singh, S., & Guestrin, C. (2016). Why should I trust you?: Explaining the predictions of any classifier. *Proceedings of the 22nd ACM SIGKDD international conference on knowledge discovery and data mining* (pp. 1135–1144). ACM.

Rudin, C. (2019). Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. *Nature Machine Intelligence, 1*(5), 206–215.

Samek, W., Wiegand, T., & Müller, K. R. (2017). Explainable artificial intelligence: Understanding, visualizing and interpreting deep learning models. arXiv preprint arXiv:1708.08296

Singh, C., Murdoch, W. J., & Yu, B. (2018). Hierarchical interpretations for neural network predictions. arXiv preprint arXiv:1806.05337

Sinha, R., Swearingen, K. (2002). The role of transparency in recommender systems. In: Conference on Human Factors in Computing Systems, pp. 830–831

Tal, A. S., Batsuren, K., Bogina, V., Giunchiglia, F., Hartman, A., Loizou, S. K., Kuflik, T. & Otterbacher, J. (2019) "End to End" towards a framework for reducing biases and promoting transparency of algorithmic systems. In: 2019 14th International Workshop on Semantic and Social Media Adaptation and Personalization (SMAP), Larnaca, Cyprus, , pp. 1-6. https://doi.org/10.1109/SMAP.2019.8864914

Theodorou, A., Wortham, R. H., & Bryson, J. J. (2017). Designing and implementing transparency for real time inspection of autonomous robots. *Connection Science, 29*(3), 230–241.

Tintarev, N., & Masthoff, J. (2011). Designing and evaluating explanations for recommender systems. *Recommender systems handbook* (pp. 479–510). Springer.

Van Berkel, N., Goncalves, J., Hettiachchi, D., Wijenayake, S., Kelly, R. M., & Kostakos, V. (2019). Crowdsourcing perceptions of fair predictors for machine learning: A recidivism case study. *Proceedings of the ACM on Human-Computer Interaction, 3*(CSCW), 1–21.

Wang, W., & Benbasat, I. (2007). Recommendation agents for electronic commerce: Effects of explanation facilities on trusting beliefs. *Journal of Management Information Systems, 23*(4), 217–246.

Wortham, R. H., Theodorou, A., & Bryson, J. J. (2016, June). What does the robot think? Transparency as a fundamental design requirement for intelligent systems. In: Ijcai-2016 ethics for artificial intelligence workshop.

Zhang, J. M., Harman, M., Ma, L., & Liu, Y. (2019). Machine learning testing: Survey, landscapes and horizons. arXiv preprint arXiv:1906.10742