From the Black Box to the Glass Box: The Effects of Explainability on Procedural Fairness and Trust in Algorithmic vs Human Decision-making

L. Jung

Student number: 14272059

M.Sc. Psychology track: Consultancy and Organizational Development

Department of Psychology

University of Amsterdam

Supervisor: Dr. Barbara Nevicka

Second Assessor: Drs. Yasmin Abbaszadeh

07.07.2023

Abstract

The present study investigated the effect of explainability on procedural fairness perceptions of algorithmic decision-making (ADM) compared to human decision-making (HDM) and whether this can affect trust in the decision maker. We used an online experiment (N = 336), with a 2 x 2 between-subjects design in which we manipulated the decision maker (human vs. algorithm) and explainability (no explanation vs. explanation). Surprisingly, the results showed that there was no significant difference between procedural fairness perceptions of ADM and HDM; however, we found that the algorithm was trusted more than the human decision maker. As expected, procedural fairness perception did have a significant and positive relationship with trust. Although explainability did not moderate procedural fairness perceptions of either ADM or HDM, adding an explanation did increase perceived accuracy of HDM. Other exploratory results showed that ADM was perceived as fairer than HDM when participants had more experience with ChatGPT and this resulted in more trust. Our findings can contribute to the implementation and development of AI, because we found that people do not seem to perceive AI as less fair than humans, and even trust AI more than humans, in an organizational decision-making context. This suggests that organizations may utilize the potential benefits of ADM over HDM. Since the explanation did not affect ADM, we encourage AI researchers and practitioners to explore other explanation styles and tools that might facilitate understanding, fairness perception, and trust in AI.

Keywords: algorithm, explainability, decision-making, trust, procedural fairness, online experiment

From the Black Box to the Glass Box: The Effects of Explainability on Procedural Fairness and Trust in Algorithmic vs Human Decision-making

Due to the rapid growth of digitalization, the use of Artificial Intelligence (AI), which is the simulation of human intelligence exhibited by machines, becomes more prominent and increasingly important in organizations (Helm et al., 2020). For example, in human resource (HR) departments, AI is used to make crucial decisions such as hiring, promotions, or bonus allocations (Malik et al., 2022). Deemed a subset of AI, algorithmic decision-making (ADM) refers to an automated process by an algorithm that carries out a large-scale collection and analysis of data to make a decision (Köchling & Wehner, 2020). Experts agree that the use of AI in organizations is inevitable to gain a competitive advantage because it is associated with significant individual and organizational benefits (Deloitte, 2020; Lindebaum et al., 2020). For instance, ADM allows organizations to review a larger quantity of data, and it can create a potentially more fair and objective process by reducing subjective biases (Lepri et al., 2018). Even though there are certain benefits to ADM, several disadvantages need to be disclosed, such as discriminatory outcomes for minority groups, or lower fairness perception of ADM compared to human decisions (Barocas & Selbst, 2016). Examining fairness perception is essential because, with fair decision-making, employees will trust the decision maker more, and will be more satisfied with their job (Lind et al., 2001). On the contrary, perceptions of unfair decision-making are associated with employee turnover (Wang et al., 2019).

Many studies assessed fairness in ADM scenarios by focusing on perceived procedural fairness, which is defined as the perception of fairness in which processes follow consistent standards for equal treatment in decisions (Starke et al., 2022). Procedural fairness is a main fairness criterion in HR decisions (Barrett-Howard & Tyler, 1986) since it is positively related to organizational outcomes such as organizational commitment, trust, and job satisfaction (Colquitt et al., 2001; Wang et al., 2019). However, findings about procedural fairness for ADM are inconsistent. Some researchers suggest that people perceive ADM as more

procedurally fair than human decision-making (HDM), others have shown that algorithms compared to humans in HR decisions are perceived as less procedurally fair (Langer & Landers, 2021; Newman et al., 2020). These inconsistencies indicate the need for a moderator, that can explain the differences in procedural fairness perceptions between HDM and ADM.

One explanation, as to why some people perceive algorithms as less fair is the black box effect, which assumes that people do not understand the mechanisms behind ADM due to its complexity (Yeomans et al., 2019). Hence, explainability may be important in establishing fairness perceptions (Shin, 2021). For AI, explainability is described as the capability to explain the mechanisms of an algorithm, and to understand how and why it produced certain results (Barredo Arrieta et al., 2020). In the field of AI, explainability is widely recognized as a key aspect of the successful application of AI (Miller, 2019). However, a consensus is missing on what constitutes a good explanation and which explanation tools and styles increase fairness perceptions. To shed light on the black box of algorithms, we will apply an explanation tool called LIME (Local Interpretable Model-agnostic Explanations), which creates quantitative visualizations of how AI predictions are generated and aim to examine whether it can increase procedural fairness perceptions of ADM (Ribeiro et al., 2016).

The foundation of decision-making is considered to be trust, but most people do not trust AI (Cho et al., 2015). Trust refers to the expectation that an agent will produce a reliable outcome in a situation categorized by uncertainty and vulnerability and trust is essential for successful cooperation and high performance (Edelenbos & Klijn, 2007; Lee & See, 2004). Prior work indicates that fairness is positively related to trust in AI (Shin, 2021) and the European Commission's High-Level Expert Group on AI (2019) includes fairness as a key requirement for the realization of "Trustworthy AI". Despite its significance for decisions, trust in AI received relatively little attention, especially in comparison with HDM and with procedural fairness as an antecedent (Knowles et al., 2022). This stresses the need to examine why ADM is linked to less trust than HDM, and how to improve this relationship.

The purpose of this research is to investigate the question of whether explainability can increase the procedural fairness perception of ADM compared to HDM and whether this affects trust. The current study will contribute to theory and practice, by advancing our understanding of fairness and trust in ADM. First, to help reconcile prior findings we will expand the theoretical knowledge of when ADM compared to HDM is perceived as fairer by introducing explainability as a moderator. Second, we will examine the usefulness of the explanation tool LIME and whether it can improve fairness perceptions and trust in ADM compared to HDM. Third, it will be assessed whether perceived procedural fairness is an antecedent of trust for both HDM and ADM. Finally, this study can contribute to AI development in practice and the insights from this study could also equip organizations with the right tools on how to turn the black box perception of algorithms into a more transparent and understandable glass box, ultimately affecting procedural fairness and trust.

Theoretical Development

Fairness in Decision-making

How people generate fairness perceptions in decisions can be explained by the uncertainty management theory of fairness (Lind & van den Bos, 2002). It is argued that a core function of fairness perceptions is, that it equips people with the means to resolve uncertainty by creating fairness expectations and estimations (van den Bos & Lind, 2002). Evolutionary speaking, uncertainty is an unpleasant and stressful experience and people generally intend to avoid uncertainty (Thau et al., 2009). A key tenet of the theory is that people apply cognitive shortcuts, called heuristics, to manage their uncertainty (Jordan et al., 2022). One such heuristic in a decision-making scenario is to base one's fairness judgment on previous fairness-related information. For example, in a longitudinal study, job seekers' prior fairness expectations, of attitudes such as supervisor's derogatory behavior, predicted their actual post-job-entry fairness perception of the organization (Jordan et al., 2022). Hence, when assessing an uncertain decision, people form fairness expectations based on prior

fairness-related information to reduce any uncertainty about the decision outcome or procedure. Ultimately, these fairness expectations will also influence one's actual fairness perceptions of the decision (Jordan et al., 2022). Since the mechanisms and procedures of how an outcome is achieved in ADM are mostly unknown to laypeople (Tešić & Hahn, 2022), and seeing how it is commonly associated with uncertainty (Liu, 2021), we expect them to have lower fairness expectations and in light thereof, we seek to find out how these laypeople will perceive procedural fairness of ADM compared to HDM.

Procedural fairness in decision-making derives from the four organizational justice dimensions, which are often studied in ADM (Starke et al., 2022). First, distributive fairness is focused on the outcome and the equal distribution of resources. Second, interpersonal fairness is concerned with the respectful and dignified interaction between the decision maker and the recipient. Third, informational fairness specifies that the information used to explain a decision is adequate and truthful. Fourth, procedural fairness refers to how the outcome is attained; here, the focus is on the processing of information and the mechanisms that relate to how an outcome is achieved. Although distributive and procedural fairness both impact organizational justice perceptions, procedural fairness is regarded as the more robust predictor (van den Bos et al., 2001). In a series of studies, it was demonstrated that participants' fairness perception of a decision procedure affects their reaction to the outcome, which differs from participants' distributive fairness perception of the outcome (Thibaut & Walker, 1975). In other words, when people believe the process of a decision is unfair then they are less likely to accept the outcome, even if that outcome is fair (Morse et al., 2022). Especially when people lack information about a decision maker's trustworthiness, which is frequently the case in ADM, then people rely on procedural fairness to shape their fairness perceptions (Glikson & Woolley, 2020; van den Bos et al., 1998).

A fair process is also associated with potential economic, ethical, and legal benefits for organizations (Gilliland, 1993; Hollander-Blumoff & Tyler, 2008; Nørskov et al., 2020).

Economically speaking, perceptions of fair procedures would make the use of algorithms more attractive and organizations could gain a financial advantage due to a faster processing of information and data. From an ethical perspective, a perception of fair procedures in ADM will also affect people's psychological well-being, such as job satisfaction or turnover intentions (Ötting & Maier, 2018; Wang et al., 2019). Lastly, from a legal standpoint, the perceived fairness of algorithmic procedures may lead to fewer discrimination cases. This highlights the importance of procedural fairness in decision-making and raises the question of how procedural fairness perceptions can be achieved for both ADM and HDM. Therefore, in the current study, we will focus on procedural fairness perceptions.

HDM vs ADM and Procedural Fairness

Algorithms have the potential to exceed a human decision maker with regard to fairness because algorithms are capable of processing a vast amount of representative data in a more standardized, consistent, and objective manner and can thereby reduce bias and error in decision-making (Lepri et al., 2018). Nonetheless, ADM also received a lot of criticism for its opacity, and its discriminatory practices and outcomes (Barocas & Selbst, 2016). It was shown that ADM procedures may provide unfavorable and discriminatory outcomes for minority groups, due to input that was based on human-biased data. When algorithms are fed with data that includes human prejudice and biases then this might amplify discriminatory tendencies or reflect general societal biases. This algorithmic bias, can also negatively affect people's procedural fairness perceptions (Kordzadeh & Ghasemaghaei, 2022). Nevertheless, algorithms, especially if they consist of neutral and representative input, have the potential to generate more procedurally fair processes and outcomes compared to human decision makers, who are more likely to suffer from information overload, negative emotions, and biases, such as confirmation bias, illusion of control, and overconfidence bias (Buchanan & Kock, 2001; Chira et al., 2008; Lepri et al., 2018; Lerner et al., 2015).

However, research has yet failed to create a consensus on whether people perceive ADM as more or less procedurally fair than HDM (Starke et al., 2022). A few studies showed no significant differences in procedural fairness between HDM and ADM (Langer et al., 2020; Suen et al., 2019). Other research found that algorithms compared to humans were perceived as procedurally fairer in selection procedures (Marcinkowski et al., 2020). However, most research points towards lower procedural fairness perceptions for algorithms compared to humans in various decision scenarios (Binns et al., 2018; Dineen et al., 2004; Newman et al., 2020; Zhang & Amos, 2023). Although ADM is potentially less biased than HDM, it is unlikely that the decision outcome will be accepted when the decision procedure itself is generally perceived as unfair (Morse et al., 2022).

These mixed findings call for the need of a moderator, that may explicate when procedural fairness perceptions differ between ADM and HDM, and how procedural fairness perceptions can be improved. We suspect that varying degrees of fairness perceptions in ADM may be due to varying degrees of explainability of the decision procedure. This is in line with the black box effect, which states that, due to a lack of transparency or explainability, people fail to understand the process behind an algorithm (Yeomans et al., 2019). With the increasing growth of technology, algorithms also become more complicated and people increasingly regard algorithms as black boxes that are impossible to comprehend, meaning that only people with expertise or specialized skills understand ADM and its processes (Castelvecchi, 2016). Therefore, with a lack of explainability people may have lower procedural fairness perceptions of algorithms compared to human decision makers, who are typically not perceived as a black box (Bonezzi et al., 2022). Conversely, with high explainability, ADM may be perceived as procedurally fairer than HDM because people might realize the algorithm's potential to outperform humans in decision-making scenarios.

Explainability of ADM

In a recent article, researchers developed the first experimentally validated theory of how individuals derive AI judgments from explanations (Yang et al., 2022). They postulate a psychological theory of explainability that is based on cognitive theories such as the theory of mind, belief formation, and generalization. Theory of mind is the ability to attribute mental states to others and thereby understand their internal processes (Buckner & Carroll, 2007). Belief formation occurs either directly through experience, inference, and deduction or indirectly via other people's experiences (Grayling, 2011). It is proposed that humans construct a mental model of AI, based on their own-, or others' experience. Finally, generalization refers to the propensity to react similarly to different but comparable situations (Shepard, 1987). Accordingly, when faced with an ADM scenario people access their mental model of AI and then generalize those prior beliefs to form perceptions about the new scenario. However, when people are provided with a source of explanation then they compare the explanation with their own prior beliefs and consequently update their mental model of AI. Thus, explainability might increase procedural fairness perceptions of ADM by changing people's mental models and by helping them realize its potential benefits.

As previously argued, people's mental model of AI is that of a black box since they lack knowledge about its mechanisms and generally there are conceptions about AI as being unfair or biased (Kordzadeh & Ghasemaghaei, 2022; Starke et al., 2022). Paradoxically, human decision makers are not perceived as black box entities because people are better at projecting their intuitive understanding of internal processes onto a human than onto an algorithm (Bonezzi et al., 2022). But concerning people's mental model of AI, multiple international large-scale surveys suggest that people are generally opposed to the application of ADM in various contexts, such as personnel selection or medical diagnosis because they are uncertain about how algorithms work (Fischer & Petersen, 2018; Grzymek & Puntschuh, 2019). Similarly, a general belief exists that algorithms fail to perform adequately when evaluating context-related aspects or subjective qualities, such as how well a person fits into

an organization (Lee, 2018; Newman et al., 2020). These negative mental models of AI may lead people to perceive algorithms as less procedurally fair than humans in decision-making.

On the one hand, the psychological theory of explainability assumes that due to the lack of knowledge about AI, humans project their own prior beliefs of AI being unfair onto an algorithm and then generalize those expectations to any future ADM scenario (Yang et al., 2022). On the other hand, when provided with an explanation about the algorithm, then people can update their mental model and they might correct any misconception about algorithms being unfair that was ingrained in their mental model. This may clarify why people perceive algorithms as less fair than human decision makers when the internal mechanics of a decision procedure are not transparent. Moreover, it implies that an explanation tool may help people recognize an algorithm's potential to make more objective and fair decisions than humans. Ultimately, people may perceive algorithms as fairer than human decision makers.

In response to the need for explanation tools for complex algorithms, a new research field emerged called XAI (eXplainable AI). The overall goal of XAI is to make AI output easier to understand and trusted by humans (Adadi & Berrada, 2018). For instance, researchers found that lectures and discussions about AI led students to prefer algorithms over human decision makers (Pierson, 2018). Additionally, the participants acknowledged the fact that humans have similar issues as algorithms but make more erroneous predictions. In line with Yang and colleagues' (2022) theory of explainability, this suggests that confronting people with information and explanations about AI can change their fairness perception of algorithms and can make them realize the potential benefits of algorithms and the potential flaws of human decision makers. Another study compared people's fairness perception of HDM and ADM when provided with explanations (Schoeffer et al., 2021) and found that ADM is perceived as fairer than HDM. They speculated that this may be due to the explanation that was provided for both conditions. Participants reported that the algorithm is more objective and less emotional than humans and that it may help to eliminate bias.

Although there is plenty of evidence that explanations do increase fairness, the question remains which explanation tool and which explanation style can provide the right kind and amount of explanatory information to facilitate understanding and fairness perception of ADM (Colquitt & Chertkoff, 2002; Langer et al., 2021; Schaubroeck et al., 1994; Shulner-Tal et al., 2022). A recent paper argued that different explanation styles also vary in effectiveness and some explanation styles are perceived as fairer and some as less fair (Shulner-Tal et al., 2023). Further research suggested that explanation style has a bigger impact on fairness perceptions when people are provided with more than one explanation style (Barredo Arrieta et al., 2020; Binns et al., 2018). This calls for an explanation tool that can incorporate multiple explanation styles.

Explanation Tool Features and Fairness in ADM

One novel explainability tool that uses several explanation styles is called LIME (Local Interpretable Model-Agnostic Explanations; Chowdhury et al., 2022). It is designed to fit any AI model and the goal is to extract the most relevant information from the model's prediction process. This tool usually consists of a combination of explanation styles such as visual explanations or feature relevance explanations. Visual explanations include simplified visualizations of the relations between the AI input and output. It is argued that visual explanations are perceived to convey more information and are easier to interpret than textual explanations (Szymanski et al., 2021). This can lead to a greater understanding of the underlying information. Research found that visual explanations of AI-assisted decisions in medical diagnostics influenced patient satisfaction and trust more than text explanations or no explanations (Alam & Mueller, 2021). Moreover, feature relevance explanations seek to illuminate the system's internal processes by generating a relevance score for all of its input variables (Barredo Arrieta et al., 2020). The relevance score is a quantified indicator of sensitivity that each input has on the output, which allows for comparisons between the variables and how much weight they have on the output. This can be useful to better

understand the underlying relationships between the input and output of ADM processes, thereby enhancing explainability (Auret & Aldrich, 2012). Empirical evidence suggests that people perceived AI-informed decisions as fairer when feature relevance explanations were added compared to having no explanations (Angerschmid et al., 2022).

Based on the psychological theory of explainability and research outlined above, it appears that a proficient explanation tool like LIME is crucial to shed light on the black box perception of algorithmic procedures because it can change people's mental models of algorithms (Ladbury et al., 2022). By establishing an understanding of the algorithm's procedures, people may recognize an algorithm's objectivity and potential to reduce bias and eventually perceive ADM as fairer than HDM (Schoeffer et al., 2021). Without explanations, people often fail to realize the potential of ADM and people may rely on negative prior beliefs that are influenced by a lack of understanding due to the black box effect. For human decision makers, people create the illusion that they can understand humans better than algorithms, while in reality, both can be seen as black boxes (Bonezzi et al., 2022). Therefore, we postulate that adding an explanation to a decision procedure might change one's fairness perception of both ADM and HDM and as a result we arrive to the following hypothesis:

Hypothesis 1: When explainability is high then ADM is perceived as higher in procedural fairness than HDM (H1a). However, when explainability is low then ADM is perceived as lower in procedural fairness than HDM (H1b).

ADM, Fairness, and Trust

Besides the effect of explainability on perceived procedural fairness, we are also interested in whether procedural fairness perceptions affect trust in ADM and HDM because trust is a fundamental aspect of decision-making (Shareef et al., 2021). Trust in the current study is characterized as an expectation that the other party will do a certain action (i.e.,

produce a reliable outcome) in a decision (Lee & See, 2004). This reflects that trust also carries some risk of the other party not fulfilling one's expectations, hence people often do not trust decision makers (Krishnan et al., 2006). Researchers from various fields agree that trust in the decision maker must be established to accept any decision outcome which highlights the importance of trust in both algorithms and humans (Carter & Bélanger, 2005; Pal et al., 2022; Schroeder & Fulton, 2017). Additionally, a lack of trust in AI systems may impede the development and adoption of AI in decision-making. Therefore, it is important to examine how one can trust the decision maker more. We propose that this may be achieved by establishing procedural fairness perceptions, or by increasing explainability of the decision process.

A theory that underlines people's expectations in trust formation is called the expectancy trust theory (Wierzbicki, 2010). The expectancy trust theory argues that trust is a subjective expectation that one party will do a certain action during an interaction. Those expectations are based on any relevant information about potential future outcomes of the interaction. One such piece of information was the perception of procedural fairness and it was found that procedural fairness can influence one's expectations about whether to trust a decision maker (Schroeder & Fulton, 2017; Viklund & Sjöberg, 2008). Scholars revealed that especially if decision scenarios are characterized by uncertainty and a lack of available information about the trustworthiness of the decision maker, then people rely on their expectations of procedural fairness to evaluate whether they can trust the decision maker (van den Bos et al., 1998). Hence, people's assessment of trust in the decision maker depends on their expectations of the decision's procedural fairness and whether they evaluate their procedural fairness perception as positive or negative. Indeed, multiple studies from the social sciences revealed a positive relationship between procedural fairness perceptions and trust in human decision-making (Alexander & Ruderman, 1987; Folger & Konovsky, 1989; Schroeder & Fulton, 2017). Taking into account the expectancy trust theory and related

empirical evidence, we expect to see this relationship in both ADM and HDM and thus replicate prior findings.

Hypothesis 2: Perceived procedural fairness is positively related to trust in the decision maker.

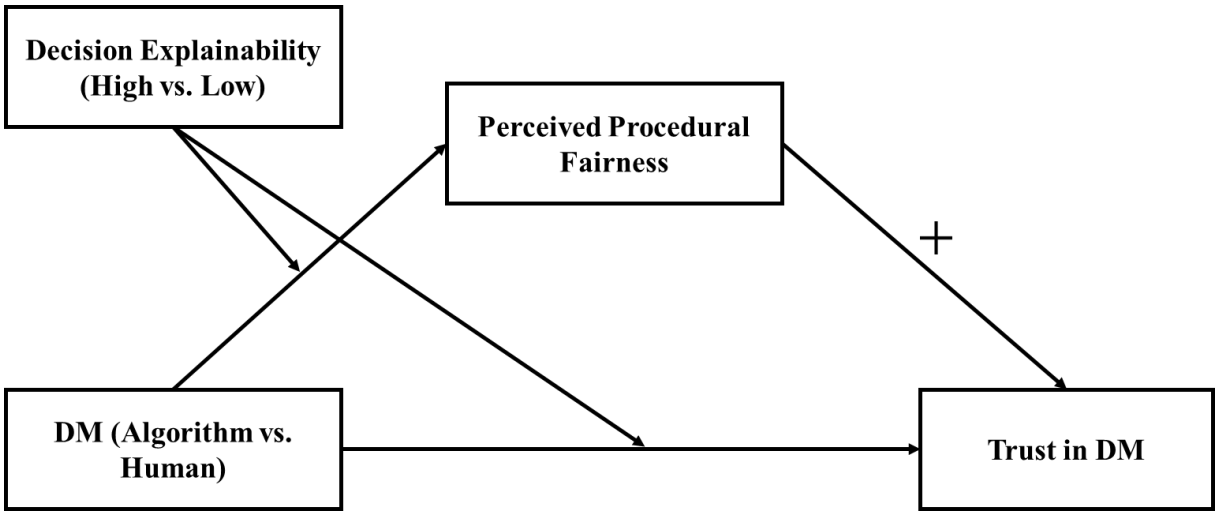
As hypothesized before, explainability may have the potential to increase procedural fairness perceptions and procedural fairness may positively impact trust which raises the question of whether explainability can also indirectly help to improve trust in the decision maker. Ribeiro et al. (2016) advocated in a recent paper that providing an explanation for a prediction can solve the problem of trusting a prediction and that providing multiple explanations for several predictions can solve the problem of trusting the model. The former describes the issue of distrusting the outcome of an algorithm and the latter refers to the problem of distrusting an algorithm before its application (Omrani et al., 2022). Specifically, they propose that LIME is capable of explaining a set of representative cases and can thereby tackle the problem of trusting the model. In their experiments, they demonstrate that LIME can increase trust in ADM and that it can help people recognize algorithms that should not be trusted. In other words, an explanation of an algorithm's internal process, can build confidence and trust in it (Ribeiro et al., 2016). Another study also found a positive main effect of explainability on fairness and a positive main effect of fairness on user trust in ADM (Shin, 2021). This indicates that explainability makes a system more interpretable and it facilitates procedural fairness perceptions. These procedural fairness perceptions are then used to assess trust in the decision maker. Building on the expectancy trust theory one could argue that prior expectations about AI being unfair are the reason why people mistrust AI (Lee, 2018). Therefore, we propose that ADM when paired with a suitable explanation tool like

LIME, will be perceived as more procedurally fair than HDM and will thereby lead to more trust in the algorithm (see Figure 1 for the whole model).

Hypothesis 3: The interaction between the decision maker (ADM vs. HDM) and explainability on trust is mediated by perceived procedural fairness such that, when explainability is high, then ADM is perceived as higher on procedural fairness than HDM and this is related to more trust in the decision maker (H3a). However, when explainability is low then ADM is perceived as lower on procedural fairness than HDM and this is related to less trust in the decision maker (H3b).

Figure 1

The Conceptual Research Model



Note. DM = decision maker.

Methods

Sample Characteristics

Based on a power analysis with the statistical analysis tool G*Power we established a minimum sample size of $N = 259$ but we aimed for 100 participants in each condition (Faul et

al., 2009). We conducted a priori power analysis for an ANOVA with a small to medium effect size of f = .175, a power of .80, alpha = .05, df = 1, and 4 groups.

The participants were recruited via convenience and snowball sampling from several private contacts (Emerson, 2015). The sample was contacted directly by the researchers via LinkedIn and WhatsApp with a standardized recruitment text, and indirectly from private contacts that redirected the link and the recruitment text to their contacts. We targeted employees from different organizations that worked at least part-time.

A total of 494 participants commenced the study, of which we excluded 89 because they did not sign the consent form (N = 45) or they did not enter any data (N = 44). Some participants did not finish the full study (N = 78), of which we decided to exclude only the participants that did not complete the variables of interest for our hypotheses (N = 69). While some participants had a suspiciously low duration of their response time, which could indicate inattentiveness, prior research shows that removing responses with low duration does not make a difference for substantive results (Greszki et al., 2015). We also tested whether the sample with and without these participants would make a difference for our findings and we did not find any differences, thus we decided to keep the participants with the fast and slow response times. After data cleaning, 336 participants remained for further analysis, which was above the required minimum sample size to achieve sufficient power.

The age range of the sample was 18 to 65 (M = 32.04, SD = 11.73) with 49.8% female participants (*male* = 44.1%, *Non-binary/third gender* = 2.7%, *prefer not to say* = 3.3%) and 38% with a bachelor degree as their highest level of completed education (*high school* = 13.4%, *post-secondary vocational education* = 12.8%, *graduate university (master)* = 26.1%, *PhD/doctorate* = 4.3%, *Other* = 5.5%). Most participants were Dutch (42.6%; see Table 1 for other nationalities) and had on average a more center political orientation (*range* = 1-5, M = 3.26 SD = 1.05; i.e., see demographics section for more details on the measure). We also asked participants about their experience with ChatGPT (*range* = 1-5, M = 2.43, SD = 1.22;

i.e., see demographics section for more details on the measure), with higher scores indicating more experience, their tenure in years (*range* = 0-41.92, *M* = 4.97, *SD* = 6.62) and how many hours per week they work (*range* = 5-100, *M* = 36.04, *SD* = 16.00). Lastly, participants indicated that they work on average 59.64% of their time (*range* = 0-100, *SD* = 29.78) at their work location (*home* = 29.46, *range* = 0-100, *SD* = 25.52; *elsewhere* = 10.90, *range* = 0-100, *SD* = 15.85) and on average their hierarchical position in their organization was 4.83 (*range* = 0-10, *SD* = 2.68; i.e., see demographics section for more details on the measure), with a higher score being indicative of a higher-level position.

Table 1

Nationality of Non-Dutch Participants

Country	%
Germany	49.6
UK	14.3
Switzerland	6.0
US	5.3
Austria	4.5
Ukraine	4.5
Turkey	2.3
Other	14

Design and Procedures

This study was part of a larger study and included other variables that were not relevant to our research. We conducted an online experiment with a 2(decision maker: human vs. algorithm) by 2(decision explainability: low vs. high) between-subjects design (see Table

2 for sample size per condition). The participants were provided with a link to an online questionnaire from the platform Qualtrics. The questionnaire started with an informed consent form that participants had to agree to and then participants were randomly allocated to one out of four conditions. First, participants were presented with a decision-making scenario involving either a human or an algorithm as the decision maker (see Decision Maker Manipulation). Next, they were either provided with an explanation or no explanation (see decision explainability manipulation). Finally, participants had to fill out a survey including their demographics, their perceived procedural fairness of the decision, their trust in the decision maker, a manipulation check, their AI literacy, and how typical it is for them to use ADM or HDM at their workplace. This study was available for a smartphone, laptop, or tablet and in English or German since a large part of the sample was expected to be native German-speaking, due to the majority of our private contacts being from German-speaking countries. In the end, only 26.5% of the participants filled it out in German and on average it took participants 77.18 minutes to finish the survey (*range* = 2.15-6020.43[1], *SD* = 484.79). As a reward for participation, we offered the chance to win a voucher or donate money to a charity.

Table 2

Sample Size per Condition after Data Cleaning

Decision maker	Low decision explainability	High decision explainability	Total
Human	75	81	156
Algorithm	88	92	180
Total	163	173	336

[1] The high upper limit of duration can be explained by the fact that participants were able to stop the survey and return to it later in time. We checked whether completion time had an effect on the manipulation check but it did not affect the manipulation check (for more details see Manipulation Check section).

Measures/Materials

Demographics

We measured several demographic variables such as age, gender (*male, female, non-binary/third gender, prefer not to say*), nationality, political orientation (1 = *right wing*, 2 = *center-right*, 3 = *center*, 4 = *center-left*, 5 = *left wing*; Chirumbolo, 2002), tenure in years, education (*high school, post-secondary vocational education, undergraduate university (bachelor), graduate university (master), PhD/Doctorate, Other*), work location in percentage (*at work, at home, elsewhere*), hierarchical position in the organization (0 = *bottom of the organization*, 10 = t*op of the organization*; Bell et al., 1990), and hours of work per week. In addition, we measured experience with ChatGPT (1 = *none at all*, 5 = *a great deal*), which is an AI chatbot that can respond to questions in a chat by generating human-like text-based messages (OpenAI, 2023).

Decision Maker Manipulation

We manipulated the type of decision maker in two different conditions. With a vignette from Newman et al. (2020), we described the same decision scenario, which was a bonus allocation decision with either an algorithm (AI decision maker) or a manager (human decision maker) as the decision maker (see Table A1 in Appendix A).

Decision Explainability Manipulation

Decision explainability consisted of two conditions, either with an explanation (high decision explainability) or without an explanation (low decision explainability). For the high decision explainability condition, participants were presented with a visual explanation of the algorithm's/manager's decision-making process (see Figure A1 in Appendix A). The visual explanation of the algorithm/manager was designed for this study, with the characteristics of the explainability tool LIME as a template. It depicted the input-output relationship of the variables and a relevance score for each variable (see Appendix A).

Several desired characteristics for LIME are proposed (Ribeiro et al., 2016). First, *interpretability* is emphasized in the sense that explanations of the relationship between input and output should be intuitive and easy to understand for laypeople. For our explanation this was achieved through a visual representation of the input variables and a relevance score for each input variable, indicating how much weight it has on the output. Second, the explanation entailed a complete and accurate description of a specific case. This is called *local fidelity* and it contributes towards a more global interpretation of the whole process by zooming into a specific case (Chowdhury et al., 2022). Also, it provides a user with a better contextual understanding of different examples. Third, LIME is *model agnostic* which means that it considers every AI model to be a black box and is suitable for any AI model as an explanation tool. To ensure that the visual explanation of the decision-making process was understandable and clear to participants we conducted a pilot study.

Pilot Study. The pilot study consisted of a 2(visual explanation: version A vs. version B) by 2(case example: present vs. absent) mixed design. The factor visual explanation was manipulated within-subjects and the diagrams across conditions differed only optically but not content-wise (see Figure B1 and B2 in Appendix B). For the between-subjects factor, participants were provided either with an additional textual example, that illustrated how the decision maker takes a decision for a fictitious case, or without the additional case example (see Figures B3 and B4 in Appendix B). To collect data, we used a convenience sampling method and Surveyswap, and in total we obtained a sample size of $N = 76$. The sample consisted of 56.6% female participants (*male* = 40.8%, *non-binary/third gender* = 2.6%) with an age range of 18 to 76 ($M = 28.03$, $SD = 12.10$), and the same participants were not included across the pilot study and the main study.

To test for the difference in perceived explainability we ran repeated measures mixed model ANOVAs with visual explanation as the within-subjects factor, case example as the between-subjects factor, and perceived explainability and explanation satisfaction as

dependent variables.[2] For perceived explainability the results showed a significant difference ($F(1,74) = 5.00$, $p = .028$ $\eta_p^2 = .06$) between version A ($M = 4.39$, $SD = 1.38$) and version B ($M = 4.66$, $SD = 1.29$) but no significant interaction with case example ($F(1,74) = 0.38$, $p = .537$, $\eta_p^2 = .01$). This indicates that version B was more understandable than version A but it did not make a difference whether a case example is present or absent. For explanation satisfaction, the results showed a marginally significant difference ($F(1,74) = 3.38$, $p = .070$, $\eta_p^2 = .04$) between version A ($M = 4.39$, $SD = 1.34$) and version B ($M = 4.60$, $SD = 1.32$) but again no significant interaction with case example ($F(1,74) = 1.18$, $p = .281$, $\eta_p^2 = .02$). In other words, people seemed to be more satisfied with version B as an explanation than with version A but it did not make a difference whether a case example is present or absent.

Finally, we used the open answer entries to code a new variable which informed whether participants reported an issue with the diagram (0 = *no issue*, 1 = *issue*). With a Chi-square test, we analyzed whether reporting an issue is dependent on the condition and we found a marginally significant difference between the case example present and the case example absent condition, $\chi^2(1, N = 76) = 3.15$, $p = .076$. From the case example absent condition 10.53% of the participants reported an issue and from the case example present condition 26.32% of the participants reported an issue. Hence, the odds of reporting an issue were 3 times higher in the case example present condition rather than in the case example absent condition.

To conclude, version B seemed to be the better explanation hence we chose version B for our main experiment. The case example did not seem to affect either perceived

[2] For details of the perceived explainability measure see the manipulation check section. We used the explanation satisfaction scale to measure participants' explanation satisfaction which refers to how understandable participants find the algorithm's process that is explained (Hoffman et al., 2018). The scale consists of eight items (1 = "I disagree strongly" to 7 = "I agree strongly"; see Table C5 in Appendix C) from which we replaced the item "This explanation lets me judge when I should trust and not trust the algorithm" with the item "The explanation lets me know how trustworthy the algorithm is" from the Items Explanation Goodness Checklist (Hoffman et al., 2018). Also, we left out two items (Item 5 and Item 6 in Table C5 in Appendix C) because they were not relevant to our decision scenario. The Cronbach's alpha of the explanation satisfaction scale was $\alpha = .86$ and it showed reasonably high content validity in prior research (Hoffman et al., 2018).

explainability or explanation satisfaction, and more people reported an issue when presented with a case example thus we did not include the case example for our main experiment. Furthermore, based on participant's comments about the explanations we decided to modify the instructions of the visual explanation and to add a sentence about the relevance score to the final version of the visual explanation.

Perceived Procedural Fairness

To assess perceived procedural fairness, we used a scale that was developed by Conlon et al. (2004) as an organizational justice measurement (see Table C1 in Appendix C). The scale consists of five items (1 = *strongly disagree* to 7 = *strongly agree*) and it showed good internal consistency with $\alpha = .82$. A sample item is "The process by which the algorithm/manager made this decision was fair". The total score of the scale was computed by averaging all responses, which is the same procedure that we used for the trust scale and the perceived explainability scale.

Trust in the Decision Maker

Trust in the decision maker was measured with an adapted version of the six-item self-report scale employed by Merritt (2011) to assess trust in algorithms. We adapted the scale to tailor the questions more toward the decision maker from our decision scenario. An example item from the scale is "I trust the manager/algorithm". We adapted the scale to match it with the experimental paradigm. The scale uses a 7-point Likert scale (1 = *strongly disagree* to 5 = *strongly agree*) and we could report a good internal consistency of $\alpha = .84$. See Table C2 in Appendix C for a full list of the adapted items as well as the original items.

Factor Analysis. To assure that perceived procedural fairness and trust in the decision maker are two distinct constructs, we ran a principal components analysis with an oblimin rotation and checked whether the items for each construct load on different factors (Abdi & Williams, 2010). Based on the Eigenvalues greater than 1 there were two factors extracted and these two factors explained together 58.27% of the variance. Factor 1 had an Eigenvalue

of 5.17 which explained 46.98% of the total variance, and Factor 2 had an Eigenvalue of 1.24 which explained 11.29% of the total variance. Looking at the pattern matrix it appears that the trust in the decision maker items load strongly together on factor 1 and the perceived procedural fairness items load strongly together on factor 2. This indicates that perceived procedural fairness and trust in the decision maker are two distinct factors because the factor loadings are sufficiently high and exceed the cut-off of 0.4 (Stevens, 2009; see Table 3).

Table 3

Results From a Factor Analysis of Perceived Procedural Fairness Items and Trust in the Decision Maker Items

Items	Factor loading	
	1	2
Factor 1: Perceived procedural fairness		
1. In my opinion, the outcome of the algorithm's/manager's decision was fair.	.08	**-.71**
2. The process by which the algorithm/manager made this decision was fair.	.07	**-.75**
3. I am satisfied with the way in which the algorithm/manager made the decision.	-.09	**-.90**
4. The algorithm/manager made this decision in an unbiased and neutral manner.	-.02	**-.71**
5. The algorithm/manager treated all employees with dignity and respect in making this decision.	.12	**-.66**
Factor 2: Trust in the decision maker		
1. I believe the manager/algorithm is a competent performer.	**.59**	-.19
2. I trust the manager/algorithm.	**.69**	-.23
3. I have confidence in the decision given by the manager/algorithm.	**.76**	-.12
4. I can depend on the manager/algorithm.	**.69**	-.12
5. I can rely on the manager/algorithm to behave in consistent ways.	**.77**	.18
6. I can rely on the manager/algorithm to do their/its best every time they/it makes a decision.	**.73**	.03

Note. N = 336. Extraction method: principal component analysis. Rotation method: oblique (oblimin with kaiser normalization) rotation. Factor loadings above .40 are in bold.

Manipulation Checks

To check whether our manipulation for the decision maker was successful, we added an instruction check question "The bonus allocation decision was made by…" with the answer options "A human" (1) and "An algorithm" (2). For decision explainability, we checked the manipulation with an adapted 3-item perceived explainability scale from Shin

(2021). We adapted the scale to match it with the experimental paradigm and we split Item 3 from the original scale into two separate items because the item seemed to be a double-barreled question, meaning it measured two different aspects of perceived explainability in one question (see Table C3 in Appendix C). The items were measured on a 7-point scale (1 = *strongly disagree* to 7 = *strongly agree*) and showed good reliability $\alpha = .80$ (see Table C3 in Appendix C for the item list). For example, an item from the scale is "I found the algorithm's/manager's decision process easily understandable".

Control Variables

Two factors that might have a potential influence on participants' fairness perception are typicality and AI literacy (Newman et al., 2020). *Typicality* refers to how typical it is for their work environment to use human or algorithmic decision-making. If ADM or HDM is typical in people's work environment, then they are more likely to perceive it as procedurally fair. This was assessed with the question "How typical is it in your company that a manager/algorithm makes these kind of decisions (e.g., bonus allocation)?" and participants used a slider to rate the typicality (0 = *Not typical at all* to 100 = *Completely typical*; Newman et al., 2020).

AI literacy describes one's familiarity and knowledge of algorithms and computers. If people possess a lot of knowledge about ADM, then they may be more inclined to perceive ADM as procedurally fair. AI literacy was measured using an 8-item scale (Wang et al., 2020; see Table C4 in Appendix C for a complete item list and anchors). A sample item from the scale is "I can make use of programming to solve a problem". For a total score, we normalized the three 4-point scale answers into a 7-point scale and then composited all the items to create a final AI literacy score.[3] In our study, the scale achieved a good reliability of

[3] To normalize the three 4-point scale items we used the following formula: (x-1)(6/3)+1.

α = .81. Based on previous research, we expect for both typicality and AI literacy a positive relationship with perceived procedural fairness (Newman et al., 2020; Wang et al., 2020).

Data analysis

The data analysis was carried out using IBM SPSS Statistics, Version 29 (IBM Corp., 2022). After testing for assumptions and the effect of control variables, we assessed the equivalence of the groups. We conducted a manipulation check by examining whether the instruction check question was answered correctly. For the decision explainability manipulation, we conducted a 2(decision explainability: low decision explainability vs high decision explainability) x 2(decision maker: human vs algorithm) between-subjects ANOVA to test for a difference in means of perceived explainability between the high decision explainability condition and the low decision explainability condition while controlling for the decision maker condition and testing for potential interaction effects. To analyze Hypothesis 1a & b, we used a 2(decision maker: human vs algorithm)x 2(decision explainability: low decision explainability vs high decision explainability) between-subjects ANOVA and PROCESS Model 1 to probe the interaction effect between decision explainability and decision maker on perceived procedural fairness (Hayes, 2022). For Hypothesis 2, we performed a hierarchical multiple regression with perceived procedural fairness as the predictor and trust in the decision maker as the dependent variable. To test for the conditional mediation hypothesis with stage one moderation (H3a & b), we employed PROCESS Model 8 with type of decision maker as the independent variable, decision explainability as the moderator, perceived procedural fairness as the mediator, and trust in the decision maker as the dependent variable. For each hypothesis, we used a significance level of α = .05.

Results

Descriptive Statistics

Table 4 displays the correlations and descriptives for all relevant variables. Consistent with our expectations and previous research (Newman et al., 2020; Wang et al., 2020), there

was a significant positive relationship between AI literacy and procedural fairness (r = .15, p = .007) as well as a significant positive relationship between typicality and procedural fairness (r = .23, p < .001). Similarly, AI literacy (r = .24, p < .001) and typicality (r = .22, p < .001) also showed a significant positive correlation with trust in decision maker. However, there were some missing values for typicality (N = 69) and AI literacy (N = 23), and due to random allocation of participants, it was found that each condition was equivalent in typicality and AI literacy which means that they cannot serve as a potential confound (see section Equivalence of Groups for additional details). Therefore, to preserve statistical power we decided to use neither AI literacy nor typicality as a control variable for all three hypotheses.

Other noteworthy significant relationships were the negative correlation between survey language and procedural fairness (r = -.23, p < .001) as well as survey language and trust in decision maker (r = -.19, p < .001). Although survey language shows a significant correlation with our variables of interest, survey language was equivalent across conditions due to random allocation and thus it was not added as a control variable (see section Equivalence of Groups). Furthermore, working hours showed a positive significant relationship with procedural fairness (r = .20, p < .001) and with trust in decision maker (r = .14, p = .008). This indicates that the more people work the more they perceive the decision scenario as procedurally fair and the more they trust the decision maker. Lastly, we found a significant positive relationship between ChatGPT experience and procedural fairness (r = .14, p = .013) as well as ChatGPT experience and trust in decision maker (r = .22, p < .001). These relationships suggest that the more experience participants have with ChatGPT the more they perceive the decision scenario as procedurally fair and the more they are likely to trust the decision maker. Both ChatGPT experience and working hours differed across the decision maker conditions but ChatGPT experience contained a few missing values (N = 7; see section Equivalence of Groups). Therefore, to preserve power we decided to use only working hours as a control variable for all three hypotheses.

Table 4

Descriptive Statistics and Correlations

	Variable	*M*	*SD*	1	2	3	4	5	6	7	8	9	10	11	12	13	14
1	Procedural fairness	4.37	1.08	-													
2	Trust in decision maker	4.35	1.02	.62**	-												
3	Algorithmic literacy	4.31	1.11	.15**	.24**	-											
4	Typicality	45.22	30.54	.23**	.22**	.19**	-										
5	Working hours	36.04	16.00	.20**	.14**	.25**	.15*	-									
6	Work from work location	59.64	29.78	-.01	.01	-.07	-.10	.11†	-								
7	Work from home	29.46	25.52	.04	-.01	.09	.05	-.08	-.85**	-							
8	Language[a]	1.26	0.44	-.23**	-.19**	-.37**	-.17**	-.08	.10†	-.09†	-						
9	Hierarchy	4.83	2.68	.10†	.09	.09	.12*	.42**	.07	-.04	.18**	-					
10	Decision maker[b]	1.54	0.50	-.04	-.01	.07	-.22**	.08	-.03	.03	.05	.00	-				
11	Decision explainability[c]	1.51	0.50	.02	.05	-.03	-.04	-.00	.05	-.02	-.01	-.07	-.01	-			
12	ChatGPT experience	2.43	1.22	.14*	.22**	.45**	.23**	.18**	-.03	-.00	-.16**	.06	.03	-.04	-		
13	Age	32.04	11.73	-.02	-.09†	-.16**	.04	.24**	.03	-.01	.41**	.49**	.03	.00	-.26**	-	
14	Tenure	4.97	6.62	-.05	-.13*	-.08	.06	.15**	.08	-.10†	.27**	.42**	.00	.02	-.18**	.65**	-
15	Nationality[d]	1.57	0.50	-.09†	-.15**	-.20**	-.25**	.01	.19**	-.08	.48**	.03	.10	.04	-.18**	.27**	.07

Note. N = 267-336[4]. [a]1 = *English*, 2 = *German*, [b]1 = *Human*, 2 = *Algorithm*. [c]1 = *Low decision explainability*, 2 = *High decision explainability*. [d]1 = *Dutch*, 2 = *Other*.

†$p < .10$. *$p < .05$. **$p < .01$.

[4] The wide range is mainly due to missing values of the control variables typicality and AI literacy which were missing a forced choice response format in the study.

Non-Response Bias

Since a lot of participants dropped out of the study, we checked for a non-response bias and compared whether there were demographic differences between participants who did not finish the full study (N = 78) and participants who completed everything (N = 327). The results of an independent t-test suggested that participants who did not finish (M = 68.17, SD = 35.96) compared to the participants who finished (M = 59.17, SD = 29.56) reported significantly more percentage of their work time spent at their work location, $t(103.18) = 2.05$, $p = .043$, $d = .29$.[5]

Furthermore, the results of a chi-square test indicated that participants who selected German as their preferred language for the survey (27.90%) were less likely to finish the survey than participants who selected English as the language (72.10%), $\chi^2(1, N = 405) = 5.36$, $p = .021$. The odds of finishing the survey were 1.83 times higher for participants who filled out the survey in English rather than for the participants that filled out the survey in German.

Lastly, we checked with a three-way loglinear analysis (decision maker vs decision explainability vs finished) whether the conditions can predict higher rates of dropping out and found a higher order effect between decision explainability and finished, $\chi^2(1, N = 405) = 5.34$, $p = .021$. For the low decision explainability condition, 13.2% out of the decision explainability total did not finish the study compared to 6.1% of the participants in the high decision explainability condition out of the decision explainability total did not finish. This means that the odds of finishing the survey were 2.33 times higher for the participants in the high decision explainability condition than for the participants in the low decision explainability condition.

[5] The Levene's test for equality of variances was significant, thus we reported the statistics of the t-test for equal variances not assumed.

To summarize, several factors influenced whether people completed the study and these factors (e.g., working from the work location, survey language, and the decision explainability condition) might have caused a non-response bias. Especially, survey language poses a potential problem since it is significantly related to our variables of interest, perceived procedural fairness, and trust in decision maker (see Table 4). However, our analysis of equivalence of groups revealed that each condition, due to the random allocation of participants, was equivalent in survey language. Thus, it is unlikely that survey language affected our results, even if it caused a non-response bias.

Manipulation Check

For the *decision maker manipulation*, we analyzed the instruction check. First, we conducted a Chi-square test to assess the effect of the decision maker condition on the manipulation check. We found that the results of the Chi-square test were significant and the manipulation was successful, $\chi^2(1, N = 336) = 5.58$, $p = .018$, $\varphi = .13$. Within the algorithm condition 57.8% successfully indicated the algorithm as the decision maker and within the human condition, 55.1% successfully indicated the human as the decision maker.

Surprisingly, in total 43.5% of the participants failed to answer the instruction check correctly. Therefore, we tested whether there was a difference in proportions that failed the instruction check for each condition. The results showed that the proportion of participants who failed the instruction check did not differ between the algorithm and human decision maker condition, $\chi^2(1, N = 336) = 0.24$, $p = .625$. Within the algorithm condition, 42.2% failed the instruction check and within the human decision maker condition, 44.9% failed the instruction check. Since this error occurred equally in both conditions, we can assume that it did not occur because of a difference in conditions. Second, the results of a three-way loglinear analysis (decision maker vs decision explainability vs instruction check) indicated that there was also no significant three-way interaction between the decision maker and decision explainability on the instruction check, $\chi^2(1, N = 336) = 0.64$, $p = .425$.

Since the sample was from various nationalities that were not all native English or German-speaking countries, we examined whether the likelihood to fail the instruction check is different for participants who took the survey in English compared to participants who took the survey in German. With a Chi-square test, we found that there was a marginally significant association between language and instruction check, $\chi^2(1, N = 336) = 2.77$, $p = .096$, $\varphi = .09$. The odds of passing the instruction check were 1.52 times higher for the participants who took the survey in German than for the participants who took the survey in English. This implies that non-native English speakers might have had more problems comprehending the decision scenario or the items related to the decision scenario.[6]

Lastly, we explored whether potential inattentiveness due to time spent on the survey (i.e. duration) affected failing the instruction check. It is possible that participants who took a break in between and returned to the survey later could not recall the exact decision maker or that participants who were too fast may have not paid enough attention to the scenario. The results of a logistic regression analysis show that duration did not significantly affect the likelihood of answering the instruction check incorrectly, $\chi^2(1, N = 336) = 0.02$, $p = .889$. Since we could not find a valid reason as to why such a large proportion failed the instruction check, and thus could not control for this statistically, we concluded that those participants should be removed from the data set and we continued further analyses with the sample that answered the instruction check correctly ($N = 190$; see Table 5 for sample size per condition).[7]

[6] We reran the analyses for the hypotheses with language as a control variable and we could not find any differences in terms of results.

[7] For robustness purposes we reran the analyses with the total sample ($N = 336$) including the participants that failed the instruction check (see Footnotes 10-12).

Table 5

Sample Size per Condition after Removal of Participants who Failed the Instruction Check

Decision maker	Low decision explainability	High decision explainability	Total
Human	44	42	86
Algorithm	50	54	104
Total	94	96	190

To examine whether the *decision explainability manipulation* was successful, we conducted a 2(decision maker: human vs algorithm) x 2(decision explainability: decision explainability low vs decision explainability high) ANOVA. There was a significant main effect of decision explainability on explainability perceptions ($F(1,186) = 5.44$, $p = .021$ $\eta_p^2 = .03$) and no significant effect of decision maker ($F(1,186) = 0.03$, $p = .858$ $\eta_p^2 < .01$). The participants in the high decision explainability condition ($M = 4.59$, $SD = 1.05$) reported more perceived explainability than the participants in the low decision explainability condition ($M = 4.18$, $SD = 1.23$). There was also no significant interaction found between decision maker and decision explainability, $F(1,186) = 1.10$, $p = .295$ $\eta_p^2 = .01$. Thus, the decision explainability manipulation was successful.

Equivalence of Groups

Before we analyzed our hypotheses, we tested for equivalence of groups between the conditions. To assure an equivalence of groups in both the decision explainability conditions and the decision maker conditions, we carried out a 2x2 ANOVA with the continuous demographical variables as the dependent variables and a chi-square test with the categorical demographical variables. We found a significant difference in working hours between the human ($M = 31.80$, $SD = 14.15$) and the algorithm ($M = 38.89$, $SD = 14.67$) condition,

$F(1,186) = 11.13$, $p = .001$, $\eta_p^2 = .06$. Also, participants in the human decision maker condition ($M = 2.13$, $SD = 1.13$) reported significantly less experience with ChatGPT than participants in the algorithm condition ($M = 2.62$, $SD = 1.30$), $F(1,182) = 7.53$, $p = .007$, $\eta_p^2 = .04$. However, ChatGPT experience contained four missing values and with regard to the equivalence of groups, working hours showed a larger effect size with a medium effect size compared to the small to medium effect size of ChatGPT experience. Therefore, we decided to include only working hours as a statistical control variable in subsequent analyses, while simultaneously conserving the statistical power of the study. We, additionally, reran all the analyses to check whether the results changed if ChatGPT experience was included as a statistical control, which was not the case.[8]

Lastly, the results of the chi-square test revealed that there was a significant association between gender and the decision maker condition, $\chi 2$ (3, $N = 186$) = 8.99, $p = .029$, $\varphi_c = .22$. There was a significantly higher proportion of males in the human decision maker condition (52.9%) than in the algorithm condition (36.6%) and there was also a significantly higher proportion of non-binary/third gender participants in the algorithm condition (5%) than in the human decision maker condition (0%). Since only 5 participants in total indicated non-binary/third gender as their gender, we decided not to add gender as a control variable.[9]

Hypothesis Testing

Assumption Testing

Prior to hypothesis testing, we assessed whether the data met the assumptions of normality, homogeneity, and linearity, and whether the data contains any outliers.

[8] There was no difference in the results of the hypotheses when adding ChatGPT experience as a control variable or having ChatGPT experience as the only control variable compared to the results with working hours as the control variable.

[9] We also ran the analysis of the hypotheses with gender as an additional control variable and we found no difference in the results of the analyses with working hours as a control variable only.

For the assumption of normality, we checked the skewness and kurtosis statistics and visually inspected the histograms and normal Q-Q plots. Some of the histograms looked slightly skewed and the points in the Q-Q plots deviated slightly from the reference line at the tails. However, the skewness (*Min* = -0.80, *Max* = 0.24) and kurtosis statistics (*Min* = -0.77, *Max* = 1.09) were still within an acceptable range of -2 and +2 (Hair et al., 2010). Thus, it was deemed that data transformation was not necessary.

Furthermore, the Levene's tests of homogeneity were not significant (*ps* > .05) and the scatterplot of standardized residuals and standardized predicted values did not show any indication of heteroscedasticity. Therefore, the assumption of equal variance was also met.

Although a few outliers were found with a boxplot and a scatterplot, none exceeded the cook's distance of 0.5 (*Max* = 0.26) (Cook, 1977). After careful inspection of each case, we could not find a valid reason to remove the data points from the data set.

Moderation Hypothesis

Hypothesis 1 predicted that when decision explainability is high then ADM is perceived as higher on procedural fairness than HDM (H1a) but, when decision explainability is low then ADM is perceived as lower on procedural fairness than HDM (H1b). For the analysis of H1a and H1b, we conducted a 2 by 2 factorial ANOVA with working hours as the control variable (covariate), decision maker (algorithm vs human) as the predictor, decision explainability (high decision explainability vs low decision explainability) as the moderator, and perceived procedural fairness as the dependent variable. There was no significant main effect of decision maker on procedural fairness ($F(1,185) = 1.09$, $p = .298$) as well as no significant main effect of decision explainability on procedural fairness ($F(1,185) = 0.68$, $p = .411$). The results of the same ANOVA showed there was no significant interaction effect between the decision maker and decision explainability on procedural fairness, $F(1,185) = 1.40$, $p = .238$. This means that when adding an explanation about the procedure of a decision then there is no difference in procedural fairness perception of an algorithmic decision

compared to a human decision (see Table 6). Without an explanation of the procedure of a decision, there is also no difference in procedural fairness perception of an algorithmic decision compared to a human decision (see Table 6). Therefore, H1a and H1b were not supported.[10]

Table 6

Means and Standard Deviations of Procedural Fairness for HDM and ADM in the Low and High Decision Explainability Conditions

Decision maker	Decision explainability	*M*	*SD*
Human	Low	4.31	1.19
	High	4.25	1.12
	Total	4.28	1.15
Algorithm	Low	4.34	1.17
	High	4.69	0.85
	Total	4.52	1.03
Total	Low	4.33	1.17
	High	4.50	1.00
	Total	4.41	1.09

Main Effect Hypothesis

[10] To check the robustness of the results we ran the analysis again with the total sample (N = 336), including the participants that failed the instruction check. Similarly, we found no support for H1a and H1b. The results suggested that there was no significant interaction effect between decision maker and decision explainability on perceived procedural fairness, $F(1,331) = 1.47$, $p = .226$.

Hypothesis 2 stated that perceived procedural fairness is positively related to trust in the decision maker. A hierarchical multiple regression was performed with working hours (control variable) entered in step one, perceived procedural fairness entered in step two, and trust in decision maker as the dependent variable. It was established that the model with procedural fairness as the predictor explained significantly more variance in trust in decision maker ($F(2,187) = 52.45$, $p < .001$, $R^2 = .36$, $\Delta R^2 = .35$, $\Delta F(1,187) = 101.49$, $p < .001$) compared to the model with the control variable only ($F(1,188) = 2.22$, $p = .138$, $R^2 = .01$). The results show that there is a positive significant relationship between procedural fairness and trust in decision maker (see Table 7). Accordingly, the higher participants' fairness perception of a decision's procedure, the more likely they are to trust the decision maker. As expected, H2 was supported.[11]

[11] We did a robustness check for H2 with the total sample (N = 336), including the participants that failed the instruction check. Again, the results showed support for H2 with a significant positive relationship between perceived procedural fairness and trust in decision maker, $\beta = 0.61$, $t(333) = 13.86$, $p < .001$, 95% CI [0.50, 0.66].

Table 7

Regression Analyses Testing the Effect of Procedural Fairness on Trust in Decision Maker

Variables	β	*t*	*df*	*p*	*Lower 95% CI*	*Upper 95% CI*
Step 1						
Working hours	.11	1.49	188	.138	-0.002	0.02
Step 2						
Working hours	.01	0.18	187	.855	-0.01	0.01
Procedural fairness	.60	10.07	187	<.001	0.46	0.68

Note. Step 1 = only control variable; Step 2 = control variable and predictor (procedural fairness).

Conditional Mediation Hypothesis with Stage One Moderation

With H3a we hypothesized that the interaction between the decision maker (HDM vs. ADM) and decision explainability on trust in the decision maker is mediated by perceived procedural fairness such that when decision explainability is high then ADM is perceived as higher on procedural fairness than HDM and this is related to more trust in the decision maker. However, when decision explainability is low then ADM is perceived as lower on procedural fairness than HDM and this is related to less trust in the decision maker (H3b). To analyze H3a and H3b, we used Model 8 in the PROCESS macro for SPSS with 5000 bootstrapped samples and working hours as the control variable, decision maker as the independent variable, decision explainability as the moderator, procedural fairness as the mediator, and trust in decision maker as the dependent variable.

For the total effect, we ran a separate analysis with PROCESS Model 1 and the results indicated that the interaction the between decision maker and decision explainability on trust in decision maker was not significant, B = -0.26, SE = 0.30, t(185) = -0.86, p = .394, 95% CI [-0.84, 0.33]. Moreover, the interaction effect between decision maker and decision explainability on procedural fairness was not statistically significant, B = 0.37, SE = 0.31, t(185) = 1.18, p = .238, 95% CI [-0.25, 0.99]. When procedural fairness was entered into the model with the interaction term of decision maker and decision explainability, then the interaction effect became marginally significant, B = -0.47, SE = 0.24, t(184) = -1.93, p = .055, 95% CI [-0.94, 0.01]. The simple effects analysis suggested that the effect of the decision maker on trust in decision maker is only significant at the level of the low decision explainability condition, (B = 0.46, SE = 0.17, t(184) = 2.66, p = .009, *95% CI* [0.12, 0.79]) but not significant at the level of the high decision explainability condition, B = -0.01, SE = 0.17, t(184) = -0.06, p = .953, *95% CI* [-0.35, 0.33]. This means that without an explanation and when controlled for procedural fairness, trust is higher in the algorithm than in the human decision maker but with an explanation there is no difference in trust between the algorithm and the human. Furthermore, there was a positive significant relationship between procedural fairness and trust in decision maker, B = 0.57, SE = 0.06, t(184) = 10.08, p < .001, 95% CI [0.46, 0.68]. Lastly, the indirect effect of the interaction between decision maker and decision explainability on trust in decision maker via procedural fairness was also not significant since the 95% confidence interval did include a zero, $B_{indirect}$ = 0.21, SE = 0.18, 95% CI [-0.15, 0.56]. These results suggest that there is a direct effect of the interaction between the decision maker and decision explainability on trust in decision maker but no indirect effect of the interaction on trust in decision maker via procedural fairness. Thus, we could not find support for H3a and H3b.[12]

[12] For H3a and H3b we also conducted a robustness check and performed the same analysis with the total sample (N = 336) that also included the participants that failed the instruction check. The only difference in the results was that the direct effect of the interaction between decision maker and decision explainability, with

Exploratory Analyses

Main Effect of Decision Maker on Trust in Decision Maker

The results of the main analysis suggested that there is no significant difference between ADM and HDM in procedural fairness perceptions. Therefore, we were interested in whether there is a difference in trust between ADM and HDM. With a two-way ANOVA (decision maker vs decision explainability) we found a significant main effect of the type of decision maker on trust in the decision maker ($F(1,186) = 5.60$, $p = .019$, $\eta_p^2 = .03$.), with higher ratings of trust in the algorithm ($M = 4.54$, $SD = 0.98$) than in the human decision maker ($M = 4.18$, $SD = 1.07$).

Understandability of the Explanation

Since adding an explanation compared to no explanation did neither affect procedural fairness perceptions nor trust in the decision maker, this raises the question of whether the explanation was too difficult for laypeople to comprehend. To test this assumption, we were interested in whether AI literacy or ChatGPT experience can moderate the relationship between decision explainability and perceived explainability as measured by the manipulation check. It is possible that when people are more experienced and familiar with AI then they might find the explanation more understandable than people who are less experienced with AI (Ehsan et al., 2021). For the main effects, we only found a positive significant relationship of ChatGPT experience on perceived explainability, $B = 0.43$, $SE = 0.21$, $t(182) = 2.06$, $p = .041$, *95% CI* [0.02, 0.84]. The results of PROCESS Model 1 suggest that there was no significant interaction between ChatGPT experience and decision explainability on perceived explainability ($B = -0.21$, $SE = 0.14$, $t(182) = -1.51$, $p = .132$, *95% CI* [-0.47, 0.06]) as well as no significant interaction of AI literacy and decision explainability on perceived explainability ($B = 0.05$, $SE = 0.15$, $t(175) = 0.31$, $p = .756$, *95% CI* [-0.25, 0.34]). Thus, the more

procedural fairness in the model, on trust in decision maker was not significant, $B = -0.28$, $SE = 0.18$, $t(330) = -1.55$, $p = .122$, 95% CI [-0.62, 0.07]. Otherwise, the results stayed the same and we could find no support for H3a and H3b.

experience people have with ChatGPT the more they understand the decision process. However, being more literate in AI or more familiar with ChatGPT does not change one's perception of explainability of the explanation.

Experience with AI

Furthermore, with the growing popularity of AI people also become more familiar with it (Lund et al., 2023). People who are more acquainted with AI might react differently to fairness perceptions of ADM compared to HDM and might even show different levels of trust in the decision maker. Therefore, we explored whether AI literacy or ChatGPT experience can moderate the relationship between the decision maker and procedural fairness and whether AI literacy and ChatGPT experience can moderate the relationship between the decision maker and trust in decision maker.

With PROCESS Model 1 we found no significant interaction between AI literacy and decision maker on procedural fairness (B = -0.01, SE = 0.14, t(175) = -0.05, p = .957, *95% CI* [-0.29, 0.28]) and no significant interaction effect between AI literacy and decision maker on trust in decision maker (B = 0.15, SE = 0.14, t(175) = 1.09, p = .278, *95% CI* [-0.12, 0.42]). However, there was a negative significant main effect of ChatGPT experience on procedural fairness (B = -0.50, SE = 0.22, t(182) = -2.26, p = .025, *95% CI* [-0.93, -0.06]) and a significant interaction effect of decision maker and ChatGPT experience on both procedural fairness (B = 0.33, SE = 0.13, t(182) = 2.51, p = .013, *95% CI* [0.07, 0.59]; see Figure 2) and trust in decision maker (B = 0.25, SE = 0.13, t(182) = 2.00, p = .047, *95% CI* [0.004, 0.50]; see Figure 3). Simple slopes analysis indicated that the effect of the decision maker on procedural fairness was significant and positive at higher levels of ChatGPT experience, B = 0.71, SE = 0.24, t(182) = 3.03, p = .003, *95% CI* [0.25, 1.18], meaning that participants that were more experienced with ChatGPT perceived ADM as procedurally fairer than HDM. Similarly, the effect of the decision maker on trust in decision maker was significant and positive at higher levels of ChatGPT experience, B = 0.68, SE = 0.23, t(182) = 3.00, p = .003,

95% CI [0.23, 1.13], meaning that participants that were more experienced with ChatGPT also trusted algorithms more than human decision makers. The effect of decision maker was not significant at lower levels of ChatGPT experience for either procedural fairness (*B* = -0.11, *SE* = 0.22, *t*(182) = -0.49, *p* = .626, *95% CI* [-0.54, 0.33]), or trust (*B* = 0.05, *SE* = 0.21, *t*(182) = 0.24, *p* = .813, *95% CI* [-0.37, 0.47]). In other words, the human decision maker is trusted less and perceived as less procedurally fair than the algorithm when participants have more ChatGPT experience. However, the algorithm is rated the same on procedural fairness and trust regardless of one's experience with ChatGPT (see Figures 2 & 3).

Figure 2

The Moderating Effect of ChatGPT Experience on the Relationship between Decision Maker and Perceived Procedural Fairness

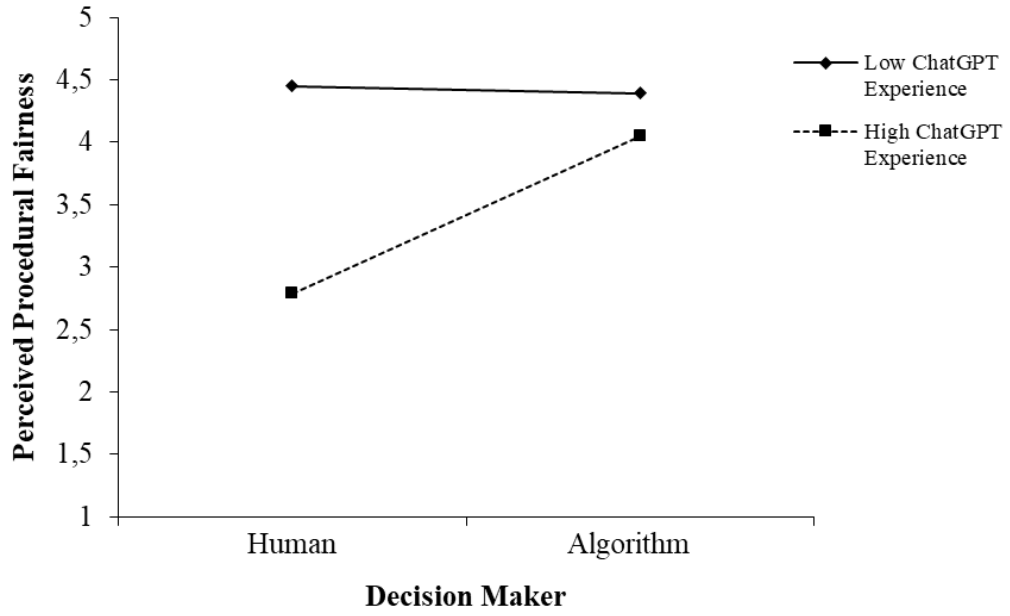
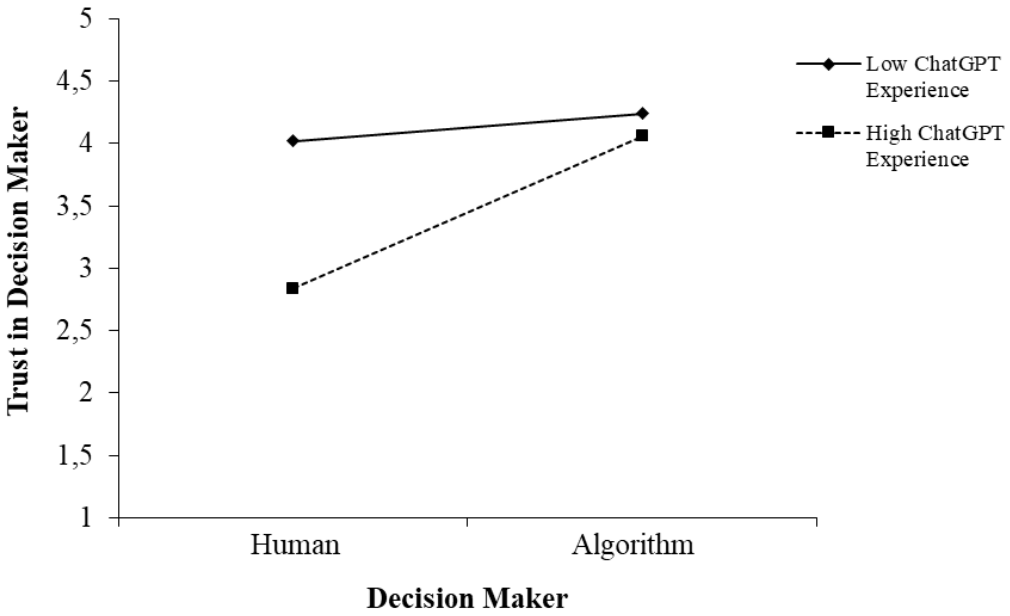


Figure 3

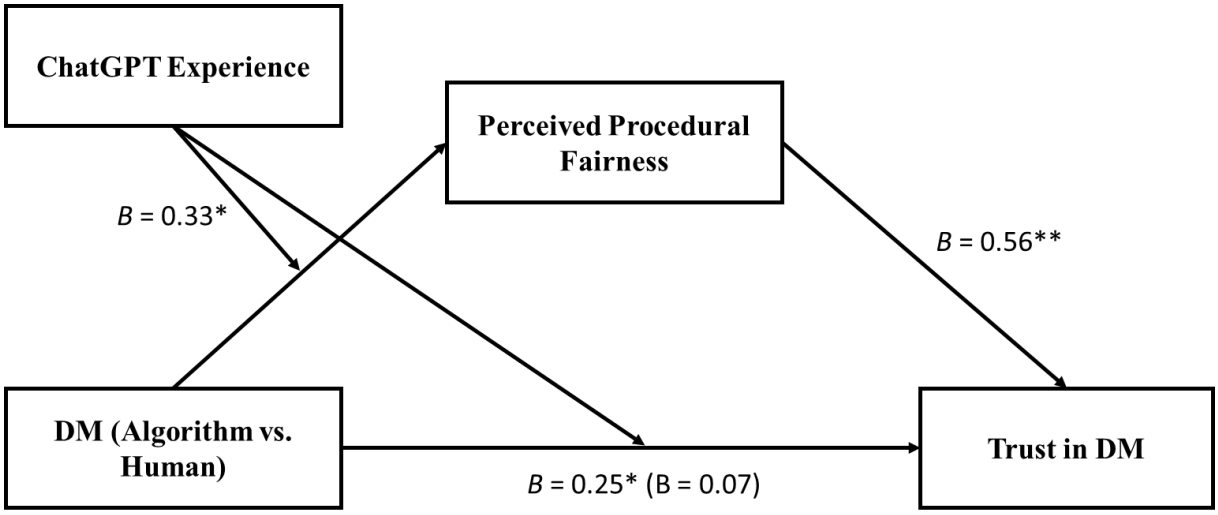
The Moderating Effect of ChatGPT Experience on the Relationship between Decision Maker and Trust in Decision Maker



Given that the interaction between the decision maker and procedural fairness is strong on the mediator procedural fairness, from our conceptual research model, we also tested for a conditional meditation with stage one moderation, with ChatGPT experience as the moderator instead of decision explainability, decision maker as the independent variable, and trust as the dependent variable. Above we already found a significant total effect, between the decision maker and ChatGPT experience on trust, and a significant effect of the interaction between the decision maker and ChatGPT experience on procedural fairness. To check for the rest of the conditional mediation model with stage one moderation, we used PROCESS Model 8 and found that when procedural fairness was entered into the model, then the interaction between the decision maker and ChatGPT experience became non-significant (B = 0.07, SE = 0.10, t(181) = 0.64, p = .521, *95% CI* [-0.14, 0.27]), but there was a significant positive relationship between procedural fairness and trust, B = 0.56, SE = 0.06, t(181) = 9.73, p < .001, *95% CI* [0.45, 0.68]. Finally, the indirect effect of the interaction between the decision maker and ChatGPT experience on trust via procedural fairness was significant, $B_{indirect}$ = 0.19, SE = 0.08, 95% CI [0.03, 0.36]. The indirect effect of decision maker on trust via procedural fairness was only significant at higher levels of ChatGPT experience ($B_{indirect}$ = 0.40, SE = 0.15, 95% CI [0.13, 0.72]) but not at lower levels of ChatGPT experience ($B_{indirect}$ = -0.06, SE = 0.13, 95% CI [-0.33, 0.20]. These results present support for the conditional mediation model with stage one moderation (see Figure 4 for the full model). It occurs that with more experience in ChatGPT, people perceive algorithms as procedurally fairer than human decision makers and this leads indirectly to more trust.

Figure 4

Statistical Pathways of the Conditional Mediation Model with Stage one Moderation



Note. DM = decision maker. Unstandardized regression coefficients for each pathway are presented. The regression coefficient for the direct effect between DM and ChatGPT experience on trust, while controlling for procedural fairness, is in parentheses.

* p < .05, ** p< .001.

Perceived Accuracy

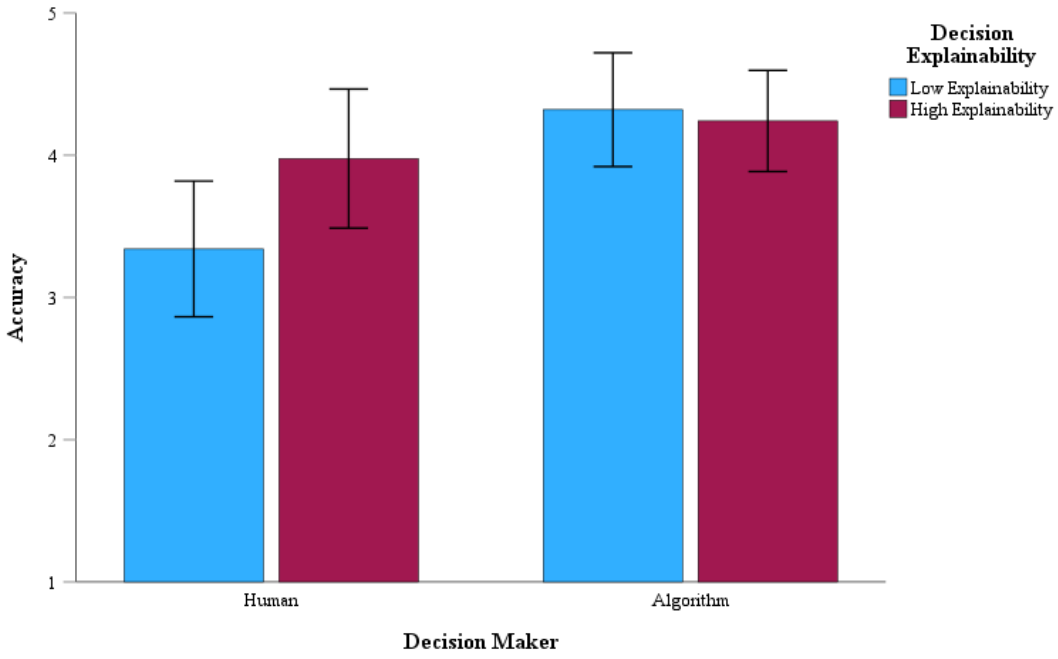
Although we did not find any differences in fairness perceptions between ADM and HDM, we did explore whether participants would perceive differences in accuracy[13] of the decision maker. With a 2(decision maker: human vs algorithm) x 2(decision explainability: low decision explainability vs high decision explainability) between-subjects ANOVA, we investigated whether the type of decision maker and decision explainability affect perceived accuracy. We found that there was a significant main effect of decision maker on perceived accuracy ($F(1,186) = 8.59$, $p = .004$, $\eta_p^2 = .04$), with the algorithm ($M = 4.28$, $SD = 1.35$) being perceived as more accurate than the human decision maker ($M = 3.65$, $SD = 1.59$).

[13] Perceived accuracy was measured with a single item on a 7-point Likert scale (1 = *very inaccurate* to 7 = *very accurate*). The item question was "How accurate do you think the algorithm/manager is in determining whether an employee should receive a bonus or not?". The accuracy item was developed in a paper that has not been published yet but a similar item was used by Kocielnik et al., (2019).

Also, there was a marginally significant interaction effect between the decision maker and decision explainability on perceived accuracy, $F(1,186) = 2.84$, $p = .094$, $\eta_p^2 = .02$. The simple effects analysis with PROCESS Model 1 revealed that the effect of the decision maker is significant only for low decision explainability, ($B = 0.98$, $SE = 0.30$, $t(186) = 3.26$, $p = .001$, *95% CI* [0.39, 1.57]), but not significant for high decision explainability, $B = 0.26$, $SE = 0.30$, $t(186) = 0.88$, $p = .378$, *95% CI* [-0.33, 0.86] (see Figure 5). It seems that when decision explainability is low then accuracy of HDM ($M = 3.34$, $SD = 1.57$) is perceived as lower than accuracy of ADM ($M = 4.32$, $SD = 1.41$) but when decision explainability is high then there is no difference in perceived accuracy between the human decision maker ($M = 3.98$, $SD = 1.57$) and the algorithm ($M = 4.24$, $SD = 1.30$).

Figure 5

The Moderating Effect of Decision Explainability on the Relationship between Decision Maker and Perceived Accuracy



Note. Error bars show 95% confidence intervals.

Discussion

The current study sought to investigate whether explainability can increase one's procedural fairness perception of algorithmic decision-making (ADM) compared to human decision-making (HDM) and whether this affects trust in the decision maker. Research is inconsistent in answering the question of whether ADM is perceived as more or less procedurally fair than HDM (Starke et al., 2022). Thus, the purpose of this research was to add to the fairness debate of ADM versus HDM by introducing the potential moderator explainability. In doing so, we intend to contribute to the XAI (eXplainable AI) research field, by investigating the usefulness of a LIME-like explanation tool for laypeople, in an organizational decision-making context. This might provide insight into what constitutes a good explanation and whether such an explanation can increase procedural fairness perceptions. Finally, we studied trust in ADM and HDM because trust is a key determinant for a successful cooperation in decision-making, and without trust, it is unlikely that a decision outcome will be accepted (Edelenbos & Klijn, 2007; Pal et al., 2022). Therefore, the focus of the study was procedural fairness perception as a determinant of trust and whether explainability can affect procedural fairness perceptions of ADM compared to HDM.

In summary, our findings suggest that there was no difference in procedural fairness perceptions between ADM and HDM but we found that the algorithm was trusted more than the human decision maker. Although explainability could not influence the relationship between the decision maker and procedural fairness perceptions, explainability did influence the relationship between the decision maker and perceived accuracy. We found that an explanation can increase accuracy perceptions of the human decision maker but not the algorithm. Furthermore, procedural fairness perception did indeed predict trust in the decision maker but procedural fairness perception did not mediate the relationship between the interaction of the decision maker and explainability on trust. Nonetheless, the exploratory

analyses showed that the algorithm was perceived as fairer than the human decision maker when the participants had more experience with ChatGPT and this resulted in more trust.

Contrary to our expectations, there did not seem to be a significant difference in procedural fairness perceptions between the algorithm and the human decision maker but the algorithm was trusted more than the human. The nonsignificant difference in fairness perceptions between the algorithm and the human decision maker may be explained by the uncertainty management theory (Lind & van den Bos, 2002). It is argued that fairness perceptions are based on prior fairness-related experience, to resolve uncertainty in an uncertain context such as the decision-making scenario from our experiment. Although we proposed, based on previous research, that ADM will be associated with more uncertainty than HDM due to the algorithm's black box appearance (Liu, 2021), it remains unknown whether this was in fact the case since we did not measure uncertainty. Based on the results, one could reason that in our decision scenario, independent of the decision maker, both conditions might have been affected by uncertainty about the outcome and the procedure. This might have led to both the HDM condition and the ADM condition being equally associated with uncertainty and may have resulted in an equal fairness perception of the decision procedure. However, trust was found to be higher in the algorithm than in the human decision maker and it had a positive relationship with fairness. This indicates that there could be differences in other fairness dimensions such as distributive fairness or interpersonal fairness and one could argue that algorithms might be associated with less uncertainty and higher fairness perceptions due to their accuracy and potential to be more objective (Starke et al., 2022). We urge future research, to measure uncertainty directly and to assess other fairness dimensions when studying ADM versus HDM.

Another theory, that can support the nonsignificant difference in procedural fairness perceptions between ADM and HDM, is called computers-are-social-actors theory (CASA). It is proposed that individuals apply the same social norms and expectations to algorithms as

they do to humans (Nass & Moon, 2000). Individuals display social behaviors towards algorithms such as politeness and reciprocity and to some extent even portray algorithms as having a personality. This may be due to an algorithm's anthropomorphism which is the perception of an algorithm having human-like characteristics (Złotowski et al., 2015). Thus, according to CASA perceptions of procedural fairness should be equal between algorithmic decision agents and human agents. In future research this could be investigated by comparing algorithms with varying degrees of anthropomorphism and human decision makers, and whether this results in different effects on procedural fairness perceptions.

Regarding explainability we failed to find support for the assumption that procedural fairness perceptions of ADM can be improved by adding an explanation to the decision procedure but we did find evidence that adding an explanation can increase people's accuracy perception of HDM. Our theoretical reasoning for this assumption was based on the psychological theory of explainability, which states that people form beliefs about a decision maker based on their own-, or others' experiences, thereby creating mental models that are used to shape perceptions of fairness or accuracy (Yang et al., 2022). Going further, these mental models are generalized to assess perceptions of similar but comparable situations, such as the decision scenario from our experiment. According to the theory, explanations function as a source of information to revise one's mental model and to replace one's prior beliefs or expectations with facts about the decision maker. Although our research did not provide any evidence that people's mental models about fairness perceptions of AI or the human changed when receiving an explanation, our exploratory analysis did indicate that people change their accuracy perception of HDM when provided with an explanation. Without an explanation, participants tend to perceive the human to be less accurate in making a decision than the algorithm; however, when an explanation was added, this seemed to increase the accuracy perception of HDM, meaning that there was no difference in perceived accuracy between HDM and ADM. In other words, the explanations can in fact change people's view of a

decision agent and update people's mental model but this was only true in our experiment for the human decision maker and not for the algorithm. Since we did not assess mental models directly, we recommend for future studies to use tests of comprehension that are used to assess mental models directly, for instance with diagramming tasks in which participants have to create diagrams mapping their understanding of the relations within HDM or ADM (Hoffman et al., 2018).

It remains unclear why we did not observe a change in procedural fairness perception when participants were provided with an explanation and why the explanation did not change people's accuracy perceptions of ADM. One possible reason might be a lack of understanding of the explanation or the fact that the explanation did not sufficiently illuminate the complex mechanisms of ADM procedures. Although we found a higher mean level of perceived explainability for the group with an explanation (M = 4.59) compared to the group without an explanation (M = 4.18), the overall difference between the groups was not excessively large considering that for both groups, the average response aligned between the response anchors of "neither agree nor disagree" to "somewhat agree".

The explanation tool LIME that we used in our study was designed by AI researchers and may be more appropriate for experts (Ladbury et al., 2022). A common pitfall of the XAI field is that AI researchers design explanation tools for themselves rather than for the end users who often possess little knowledge about AI (Miller et al., 2017). This can harm the user's comprehension of the system's processes and outcomes (Nourani et al., 2019). Therefore, it is important in XAI research to select appropriate explanations to make AI systems more understandable for laypeople (Goebel et al., 2018). Although we did test the explanation beforehand in our pilot study, four participants reported that the explanation was too complicated. Research has found that the meaningfulness of explanations and their alignment with human logic is essential to make AI systems more intelligible for laypeople (Nourani et al., 2019). In addition, researchers demonstrated that laypeople prefer simplicity

over detailed probabilistic explanations (Roy et al., 2021). With respect to the LIME explanations, which are rather probabilistic, it is possible that these explanations lack meaningfulness or do not follow human logic (Hagras, 2018). This implies that LIME explanations may not be appropriate explanations for laypeople to grasp the complexity of decision-making processes, especially for algorithms. It could explain why in our experiment, explainability did not affect procedural fairness perceptions for both ADM and HDM and why it did not affect accuracy perceptions of ADM. We urge future research to compare different explanation tools and to consider the meaningfulness and complexity of explanations when studying laypeople. One such approach might be the fuzzy logic systems, which are explanations that are based on simple human thinking and entail if-then rules that represent linguistic human concepts such as low/high instead of numerical probabilities (Hagras, 2018). For instance, the LIME explanation could be designed in the same fashion but the numbers that represent the weights would be replaced by anchors ranging from low to high. Thus, the input-output relationship of the algorithm is described in linguistic terms instead of probabilities and might be easier for laypeople to understand.

In our exploratory analyses, we tested whether experience with ChatGPT and AI literacy would increase perceived explainability for the explanation group however both ChatGPT experience and AI literacy did not change explainability perceptions for the explanation group nor the group without an explanation. Therefore, being more familiar with AI did not seem to matter for the understandability of the explanation. But we want to emphasize that scoring higher on AI literacy or ChatGPT experience does not necessarily make you an expert in ADM and one could still be considered a lay person whilst being more familiar with AI or ChatGPT than an average person (Ehsan et al., 2021). One should keep in mind that these results do not rule out the possibility that LIME explanations are too complex for laypeople to comprehend.

Taking into account the recent trends in AI systems such as ChatGPT or DALL-E it is possible that people currently are more acquainted with algorithms and might associate ADM as less of a black box than in previous years (Lund et al., 2023). Hence, due to AI's saliency in the media and frequent exposure to AI tools such as ChatGPT, people nowadays are more aware of the potential advantages and disadvantages of both ADM and HDM (Ehsan et al., 2021; Ouchchy et al., 2020). In our exploratory analyses, we found evidence that can strengthen that claim; we found that people who have more experience with ChatGPT tend to perceive HDM as less procedurally fair than ADM and ultimately trusted the human less. Conversely, for participants with less ChatGPT experience, there was no difference in procedural fairness perceptions between ADM and HDM. It seems that experience with AI changed the fairness perception of the human decision maker but not the algorithm. A possible explanation is that with increased familiarity or experience with AI, people become more aware of the potential benefits as well as the limitations associated with ADM, such as algorithmic bias (Horowitz et al., 2023; Kordzadeh & Ghasemaghaei, 2022). At the same time, people might recognize that algorithmic bias is mostly due to human-biased data and that ADM can still be advantageous over HDM. This may explain why in our experiment we observed that with more ChatGPT experience, participants lowered their procedural fairness perception of the human decision maker, but fairness perceptions for the algorithm neither increased nor decreased. Although we did not measure awareness of algorithmic bias, these results still showed that there is a difference in fairness perception and trust in the decision maker between people who are more acquainted with AI and people who are less acquainted with AI. Future research needs to replicate these findings with confirmatory studies and should investigate the role of algorithmic bias awareness in ADM and HDM.

In line with previous research (Alexander & Ruderman, 1987; Folger & Konovsky, 1989; Schroeder & Fulton, 2017), we detected a positive relationship between perceived procedural fairness and trust. The results provide confidence in the notion that trust in the

decision maker can be established by creating fair procedures and that procedural fairness perception is a determinant of trust in the decision maker (van den Bos et al., 1998). Taking into account the expectancy trust theory, it can be argued that procedural fairness perceptions of the decision influence one's subjective expectation that the other party will do a certain action and deliver potential future outcomes. If people think that the decision procedure is fair then this will signal that the decision maker will hold up to their expectations and people are more likely to trust the decision maker (Viklund & Sjöberg, 2008).

Finally, this study did not find support for the assumption that the interaction between the decision maker and explainability on trust will be mediated by perceived procedural fairness. Rather, the analysis showed that, the interaction between decision maker and explainability did not affect trust indirectly via procedural fairness. In an article by Ribeiro et al. (2016) it is proposed that providing multiple explanations for several predictions can resolve the problem of distrusting the model (i.e., the decision maker). However, we only provided participants with an explanation for a single prediction instead of providing them with an explanation for several cases. Besides, we failed to find the proposed trust issue with the algorithm as the decision maker, instead, we observed more trust in the algorithm than in the human decision maker. Hence, this may explain why we did not find the desired effect.

Likewise, it is also possible that trust in AI might be conceptually different from trust in a human (Bedué & Fritzsche, 2022). Measuring trust often entails human-like constructs such as ability, integrity, and benevolence, whilst ignoring additional AI-relevant factors like data security or alignment with social norms and values (Banavar, 2016; Mayer et al., 1995). In a recent paper, it is argued that trust in AI is a multifaceted and complex phenomenon that is difficult to capture in a single dimension (Vereschak et al., 2021). This poses the issue of identifying viable measures for trust. Unfortunately, this goes beyond the scope of our research but other research suggests that trust in AI can also be captured using indirect metrics such as perceived accuracy (Nourani et al., 2019). In fact, in our additional analysis, we also

explored whether there is an interaction between the decision maker and explainability on perceived accuracy and we only found an increase in accuracy for the human and not the algorithm when adding an explanation. This is consistent with the notion that trust might be conceptually different for AI and humans. We call for future research to investigate whether trust dimensions differ between algorithms and humans and we highlight the importance of considering different trust dimensions when assessing trust as a construct.

Strengths, Limitations, and Suggestions for Future Research

This study contains several methodological strengths. First of all, it is worth mentioning that we conducted a pilot study to assess the perceived explainability and explanation satisfaction for two different versions of our explanation manipulation and the inclusion of a case example. Based on qualitative data from the pilot study we were able to make minor improvements to the explanation and the results of the quantitative analyses informed us about which design works better in terms of perceived explainability and satisfaction of the explanation.

Second, the experimental design of the study allowed for causal inference, manipulation of the variables, and the minimization of confounding factors. By randomly allocating participants to different conditions, we could control for any potential confounding variables that may impact the results, and by implementing a manipulation check we were able to assess the effectiveness of the experimental manipulation. In doing so, we ensured that the intended manipulations had the desired effect on the dependent variables, hence strengthening the study's internal validity.

Third, in terms of the robustness of the study, we employed only existing and validated measures which also showed high internal consistencies and we conducted a factor analysis to confirm that fairness and trust are indeed two distinct constructs. Besides, we tested the robustness of the results by examining whether they hold under different conditions. We first ran the analyses for our hypotheses on the sample that passed the manipulation check

for the decision maker and then we ran the analyses on the total sample, including the participants that failed the manipulation check. Since the results were consistent for our hypotheses with both samples, it increases the confidence in the robustness of our results. In addition, the sample contained participants from various nationalities and from across different companies which adds to the generalizability of the results.

Apart from its strengths, there are also a few limitations that warrant attention. Surprisingly, while the manipulation of the decision maker was successful, a sizeable proportion of the sample failed the manipulation check which meant that we had to exclude a lot of participants for the hypothesis testing. We ran multiple tests to inspect why this may have been the case and the only indication we found was that participants who filled out the survey in English were more likely to fail the manipulation check than participants who filled out the survey in German. Given the various nationalities and the large proportion of German participants, one could speculate that the participants who selected English as the survey language were mostly non-native English speakers; however, the participants who selected German as the survey language were mostly native German speakers. Therefore, it is possible that the participants had language issues with the English survey due to a lack of English proficiency. However, this does not account for the still large proportion of participants with the German survey that failed the manipulation check.

Based on informal feedback from the participants, we learned that some participants thought the instruction check was too easy and assumed that it was testing some hidden information which is why they chose the opposite answer. For future research, this suggests that the relation between the scenario and the instruction check should be made clear which can be done for instance by incorporating the instruction check immediately after the decision scenario instead of placing it after the scales, as was done in our study. Although the decision scenario was a validated scenario from previous research, it did not sufficiently serve the purpose of this study (Newman et al., 2020).

After consultation with one of the researchers from the study that developed the decision scenario, we learned that they found almost 100% correct responses on the instruction check. The only explanation they had for the large proportion of our participants failing the instruction check, was the characteristics of our online sample. They emphasized that online samples are often prone to inattentiveness and although we failed to find that duration affects the likelihood of failing the instruction check, these results alone do not rule out a problem with attention. Also, in line with our findings of a potential language issue, they reasoned that language could have been a problem with many non-native speakers in the sample. Therefore, we recommend future researchers to take these issues into account and to conduct research in the lab with surveys that are in the participants' mother tongue.

Furthermore, we need to point out that we observed a non-response bias in our study. It was found that the more time people spent in the office, the less likely they were to finish the survey. People who work from the office may have less time to participate in research compared to people who work from home or elsewhere. Also, the participants who did the survey in German were less likely to finish the study compared to the participants who filled it out in English. A non-response bias could have affected the conclusiveness of our results since it makes the sample less representative in terms of work location and language. However, it is unlikely that work location and language have caused a difference in results because the survey questions were the same in each language and they primarily focused on perceptions about a hypothetical scenario that was unrelated to their work environment. Also, work location did not correlate with our variables of interest and both language and work location were equivalent for each condition due to random allocation of participants. The potential issue of the non-response bias can be resolved in future studies by rewarding participants financially, by sending out reminders, or with longer data collection periods.

Another limitation of our study is that the content of the explanation was fictitious and based on a hypothetical decision scenario. We copied the format and design of a LIME-

generated explanation to develop an explanation that fits our decision scenario. Usually, these explanations are generated by an AI system and are based on real data (Zhang et al., 2019). This raises the question of whether the results would be different with real data in a real-life scenario. Research has demonstrated that algorithms are perceived as more fair when the outcome of a decision is in favor of the person affected by the decision (Wang et al., 2020). Thus, we urge future research to investigate the effect of outcome favorability, for instance with different scenarios either allocating a bonus or no bonus. Further research could use decision scenarios in which people are either directly affected by ADM or need to imagine that they receive an outcome of a decision that is either positive or negative. We expect that outcome favorability can increase fairness perceptions for both ADM and HDM but with unfavorable outcomes, we expect ADM to be perceived as less fair because people might fear that the algorithm makes systematically unfair decisions (Noble et al., 2021).

Finally, due to the removal of participants, this study was slightly underpowered, thereby reducing the chances of discovering true effects in the sample. Based on a priori power analysis we needed 259 participants to achieve the desired power level of .80 but the actual sample size that was used for the hypothesis testing was only 190. With a post-hoc power analysis, using the same parameters, we determined a power of .67 which is generally considered too low (Abraham & Russell, 2008). Low power limits the ability to detect complex relationships, such as moderation or mediation, and low power hampers the generalizability of the findings. The lack of power could also explain why we did not find the expected results for our conditional mediation model. Therefore, for future research, we emphasize the importance of a large sample size to attain a desired power level of at least .80.

Other directions for future research could be longitudinal studies about AI and how opinions, attitudes, or perceptions of AI may change over time. For instance, future researchers could explore the effect of more intensive explanation styles such as lectures or discussions and to test whether those affect procedural fairness perceptions or trust. Based on

previous research, we assume that intensive explanation styles can have an even bigger effect on people's mental model of AI than shorter explanations and may be an effective means to increase fairness perceptions and trust in AI (Pierson, 2018).

Theoretical Implications

The results of this study offer several theoretical implications. First, this study can contribute to the fairness literature. In particular, we discovered that there seems to be no difference in fairness perceptions between ADM and HDM. However, with higher levels of ChatGPT experience, procedural fairness perceptions were higher for ADM than for HDM. Although many studies point towards lower procedural fairness perceptions of algorithms compared to humans, our results suggest that this is not the case (Binns et al., 2018; Dineen et al., 2004; Newman et al., 2020). Yet, still various other studies are consistent with our findings and also show that either there is no significant difference in procedural fairness between ADM and HDM (Langer et al., 2020; Suen et al., 2019) or that people with more knowledge of or experience with AI perceive ADM as fairer (Schoeffer et al., 2021).

Second, this study can contribute to the XAI literature. Our results emphasize that explainability does not always improve fairness perceptions of the decision procedure or trust in the decision maker but it did improve perceived accuracy. Although the explanation tool LIME might not be a useful tool to enhance laypeople's accuracy perceptions of ADM, it appeared to be an effective tool to increase accuracy perceptions of HDM. This shows that it is essential to further study which explanation tool and style is compatible with laypeople to establish a basic understanding of the processes and outcomes in ADM and HDM. Despite an expanding interest and effort of the XAI literature to develop and apply explainable AI systems, only a few studies evaluated these tools and considered laypeople's assessment of them (Adadi & Berrada, 2018). We highlight that users of-, and subjects to AI decisions may benefit from appropriate explanations because it enhances understanding of the output and promotes people to create more adequate mental models of AI (Kulesza et al., 2013).

Finally, by applying an interdisciplinary approach and by attempting to combine and integrate theories from data science, psychology, and social sciences, we advanced the theoretical knowledge of laypeople's perception of AI in decision-making (Boykin et al., 2021). We utilized the reasoning of the uncertainty management theory to explain that procedural fairness perceptions may arise due to uncertainty about the decision procedure which is resolved by relying on prior fairness-related experiences (Lind & van den Bos, 2002). Taking into account the psychological theory of explainability, we put forward that people change their mental models about AI when presented with an explanation (Yang et al., 2022). Additionally, with the expectancy trust theory, we illustrated that people's trust in the decision maker is determined by their perception of whether they think the decision's procedure is fair (Wierzbicki, 2010). By integrating already existing and validated theoretical frameworks, we were able to extend their applicability to the field of AI, and by introducing a psychological perspective to AI research, we tried to broaden the scope of the field to a more interdisciplinary approach, thereby opening up new avenues for future research.

Practical Implications

Next to the above-mentioned theoretical implications, the current study also contains several practical implications. We found that LIME explanations for laypeople are rather ineffective in increasing procedural fairness and trust for both ADM and HDM. That being said, the explanation did increase perceived accuracy of HDM. This advances our understanding of what constitutes a good explanation in AI and it can be used by explainable AI designers as a guideline to build more comprehensible explanation tools to also inform non-experts about the internal processes of an algorithm. When designing explanation tools researchers should take into account that the logic of an explanation should be meaningful and should follow simple human rationale (Nourani et al., 2019).

Next, our findings suggest that procedural fairness perceptions can predict trust in the decision maker. In an organizational context, establishing trust is essential to achieve high

performance in teams and to foster cooperation (Edelenbos & Klijn, 2007). Employees will be less inclined to accept the outcome of any decision if they lack trust in the decision maker (Schroeder & Fulton, 2017). Therefore, our results present valuable knowledge for organizations to help improve trust in the decision maker. This can be done by using fair procedural characteristics in the decision-making process, such as voice, neutrality, consistency, accuracy, reversibility, and transparency (Dolan et al., 2007). Giving employees *voice* involves providing them with an opportunity to contribute to the decision-making process, for instance by expressing their opinion (Bies & Shapiro, 1988). *Neutrality* refers to the decision maker who must be able to ignore one's interest in order to be unbiased, which may be assured with a declaration of interest stating the nature and extent of the decision maker's interest (Magner et al., 2000). For a decision process to be *consistent*, it is required that the process remains the same under varying circumstances (Leventhal, 1980). This may be done with a standardized set of procedures or a protocol that is used in every context. Concerning *accuracy*, decisions must be based on accurate information and evidence that can lead to valid and reliable decisions (Dolan et al., 2007). To achieve valid and reliable decisions, practitioners should use a data-driven approach in decision-making and rely on evidence-based practice instead of personal preference and intuition. In case the outcome may be extremely unfavorable for a certain group then there should be an opportunity for *reversibility* of the decision, for example by providing the opportunity to appeal an unfavorable decision which could then be reviewed and, if necessary, reversed (Tsuchiya et al., 2005). Lastly, *transparency* of a decision can be accomplished by explaining to employees how the decision was made (Chowdhury et al., 2022). Although the explanation used in our experiment did not influence procedural fairness perceptions, we would still assert that other explanations can increase procedural fairness perceptions of a decision. Hence, in practice it is wise to select the type of explanation carefully and with sufficient empirical evidence.

Moreover, while this study found no differences in procedural fairness perceptions between ADM and HDM, it showed that laypeople do trust algorithms more than human decision makers and that they are perceived as more accurate than humans. For people with more experience with AI tools like ChatGPT, we found that algorithms are indeed perceived as procedurally fairer and are trusted more than human decision makers. This implies that people may gradually accept AI in decision-making due to its growing popularity and due to greater degrees of familiarization. When perceived accuracy and trust in algorithms are higher than in humans and without a difference in fairness perception, then it could be beneficial for organizations to employ algorithms as decision makers. Algorithms, when programmed with neutral and representative input, are capable of producing more standardized, consistent, and objective output compared to humans whose decision-making is often influenced by emotions, information overload, stereotypes, and biases (Buchanan & Kock, 2001; Lepri et al., 2018; Lerner et al., 2015; Woods et al., 2020). On top of that, organizations can gain a financial advantage because algorithms can handle larger quantities of data than humans in a more efficient and timely manner (Davenport, 2018).

Conclusion

Since our society is becoming increasingly digitized, it appears inevitable that AI will someday be part of our daily life. Especially at work, AI is already applied to make significant decisions, which highlights the importance of trust in AI and positive fairness perceptions to promote the successful implementation and development of AI. Our findings suggest that procedural fairness perceptions and trust in AI are shifting towards a more positive picture of AI in decision-making, especially for people who are more experienced with AI tools such as ChatGPT. Although we failed to find evidence for the effectiveness of a visual explanation tool for AI, we still found support for the usefulness of the explanation in increasing accuracy perceptions of human decision makers.

Beyond the previously described practical and theoretical strengths of our study, we hope to inspire future research to explore other ways of how to improve fairness perceptions of a decision procedure and how to increase trust in the decision maker. Finally, due to the increasing popularity of the use of AI in practice, we encourage researchers and practitioners to further develop explainable AI systems to make them more accessible and understandable for everyone and to eventually turn the black box perception of AI into a glass box.

References

Abdi, H., & Williams, L. J. (2010). Principal component analysis: Principal component analysis. *Wiley Interdisciplinary Reviews: Computational Statistics*, *2*(4), 433–459. https://doi.org/10.1002/wics.101

Abraham, W. T., & Russell, D. W. (2008). Statistical Power Analysis in Psychological Research. *Social and Personality Psychology Compass*, *2*(1), 283–301. https://doi.org/10.1111/j.1751-9004.2007.00052.x

Adadi, A., & Berrada, M. (2018). Peeking Inside the Black-Box: A Survey on Explainable Artificial Intelligence (XAI). *IEEE Access*, *6*, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052

Alam, L., & Mueller, S. (2021). Examining the effect of explanation on satisfaction and trust in AI diagnostic systems. *BMC Medical Informatics and Decision Making*, *21*(1), Article 178. https://doi.org/10.1186/s12911-021-01542-6

Alexander, S., & Ruderman, M. (1987). The role of procedural and distributive justice in organizational behavior. *Social Justice Research*, *1*(2), 177–198. https://doi.org/10.1007/BF01048015

Angerschmid, A., Zhou, J., Theuermann, K., Chen, F., & Holzinger, A. (2022). Fairness and Explanation in AI-Informed Decision Making. *Machine Learning and Knowledge Extraction*, *4*(2), 556–579. https://doi.org/10.3390/make4020026

Auret, L., & Aldrich, C. (2012). Interpretation of nonlinear relationships between process variables by use of random forests. *Minerals Engineering*, *35*, 27–42. https://doi.org/10.1016/j.mineng.2012.05.008

Banavar, Guruduth. (2016). *Learning to trust artificial intelligence systems Accountability, compliance and ethics in the age of smart machines*. IBM. https://lexing.eu/wp-content/uploads/2017/06/34348524.pdf

Barocas, S., & Selbst, A. D. (2016). Big Data's Disparate Impact. *SSRN Electronic Journal*. https://doi.org/10.2139/ssrn.2477899

Barredo Arrieta, A., Díaz-Rodríguez, N., Del Ser, J., Bennetot, A., Tabik, S., Barbado, A., Garcia, S., Gil-Lopez, S., Molina, D., Benjamins, R., Chatila, R., & Herrera, F. (2020). Explainable Artificial Intelligence (XAI): Concepts, taxonomies, opportunities and challenges toward responsible AI. *Information Fusion*, *58*, 82–115. https://doi.org/10.1016/j.inffus.2019.12.012

Barrett-Howard, E., & Tyler, T. R. (1986). Procedural justice as a criterion in allocation decisions. *Journal of Personality and Social Psychology*, *50*(2), 296–304. https://doi.org/10.1037/0022-3514.50.2.296

Bedué, P., & Fritzsche, A. (2022). Can we trust AI? An empirical investigation of trust requirements and guide to successful AI adoption. *Journal of Enterprise Information Management*, *35*(2), 530–549. https://doi.org/10.1108/JEIM-06-2020-0233

Bell, R. A., Roloff, M. E., Van Camp, K., & Karol, S. H. (1990). Is It Lonely at the Top?: Career Success and Personal Relationships. *Journal of Communication*, *40*(1), 9–23. https://doi.org/10.1111/j.1460-2466.1990.tb02247.x

Bies, R. J., & Shapiro, D. L. (1988). Voice and Justification: Their Influence on Procedural Fairness Judgements. *Academy of Management Journal*, *31*(3), 676–685. https://doi.org/10.2307/256465

Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., & Shadbolt, N. (2018). "It's Reducing a Human Being to a Percentage": Perceptions of Justice in Algorithmic Decisions. *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, 1–14. https://doi.org/10.1145/3173574.3173951

Bonezzi, A., Ostinelli, M., & Melzner, J. (2022). The human black-box: The illusion of understanding human better than algorithmic decision-making. *Journal of*

Experimental Psychology: General, *151*(9), 2250–2258. https://doi.org/10.1037/xge0001181

Boykin, C. M., Dasch, S. T., Rice Jr., V., Lakshminarayanan, V. R., Togun, T. A., & Brown, S. M. (2021). Opportunities for a More Interdisciplinary Approach to Measuring Perceptions of Fairness in Machine Learning. *Equity and Access in Algorithms, Mechanisms, and Optimization*, 1–9. https://doi.org/10.1145/3465416.3483302

Buchanan, J., & Kock, N. (2001). Information Overload: A Decision Making Perspective. In M. Köksalan & S. Zionts (Eds.), *Multiple Criteria Decision Making in the New Millennium* (Vol. 507, pp. 49–58). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-56680-6_4

Buckner, R. L., & Carroll, D. C. (2007). Self-projection and the brain. *Trends in Cognitive Sciences*, *11*(2), 49–57. https://doi.org/10.1016/j.tics.2006.11.004

Carter, L., & Bélanger, F. (2005). The utilization of e-government services: Citizen trust, innovation and acceptance factors. *Information Systems Journal*, *15*(1), 5–25. https://doi.org/10.1111/j.1365-2575.2005.00183.x

Castelvecchi, D. (2016). Can we open the black box of AI? *Nature*, *538*(7623), 20–23. https://doi.org/10.1038/538020a

Chira, I., Adams, M., & Thornton, B. (2008). Behavioral Bias within the Decision Making Process. *Journal of Business & Economics Research (JBER)*, *6*(8). https://doi.org/10.19030/jber.v6i8.2456

Chirumbolo, A. (2002). The relationship between need for cognitive closure and political orientation: The mediating role of authoritarianism. *Personality and Individual Differences*, *32*(4), 603–610. https://doi.org/10.1016/S0191-8869(01)00062-9

Cho, J.-H., Chan, K., & Adali, S. (2015). A Survey on Trust Modeling. *ACM Computing Surveys*, *48*(2), 1–40. https://doi.org/10.1145/2815595

Chowdhury, S., Joel-Edgar, S., Dey, P. K., Bhattacharya, S., & Kharlamov, A. (2022). Embedding transparency in artificial intelligence machine learning models: Managerial implications on predicting and explaining employee turnover. *The International Journal of Human Resource Management*, 1–32. https://doi.org/10.1080/09585192.2022.2066981

Colquitt, J. A., & Chertkoff, J. M. (2002). Explaining Injustice: The Interactive Effect of Explanation and Outcome on Fairness Perceptions and Task Motivation. *Journal of Management*, *28*(5), 591–610. https://doi.org/10.1177/014920630202800502

Colquitt, J. A., Conlon, D. E., Wesson, M. J., Porter, C. O. L. H., & Ng, K. Y. (2001). Justice at the millennium: A meta-analytic review of 25 years of organizational justice research. *Journal of Applied Psychology*, *86*(3), 425–445. https://doi.org/10.1037/0021-9010.86.3.425

Conlon, D. E., Porter, C. O. L. H., & Parks, J. M. (2004). The Fairness of Decision Rules. *Journal of Management*, *30*(3), 329–349. https://doi.org/10.1016/j.jm.2003.04.001

Cook, R. D. (1977). Detection of Influential Observation in Linear Regression. *Technometrics*, *19*(1), 15–18. https://doi.org/10.2307/1268249

Davenport, T. H. (2018). *The AI Advantage: How to Put the Artificial Intelligence Revolution to Work*. The MIT Press. https://doi.org/10.7551/mitpress/11781.001.0001

Deloitte. (2020). *State of AI in the Enterprise – 3rd Edition*. https://www2.deloitte.com/content/dam/Deloitte/de/Documents/technology-media-telecommunications/DELO-6418_State%20of%20AI%202020_KS4.pdf

Dineen, B. R., Noe, R. A., & Wang, C. (2004). Perceived fairness of web-based applicant screening procedures: Weighing the rules of justice and the role of individual differences. *Human Resource Management*, *43*(2–3), 127–145. https://doi.org/10.1002/hrm.20011

Dolan, P., Edlin, R., Tsuchiya, A., & Wailoo, A. (2007). It ain't what you do, it's the way that you do it: Characteristics of procedural justice and their importance in social decision-making. *Journal of Economic Behavior & Organization*, *64*(1), 157–170. https://doi.org/10.1016/j.jebo.2006.07.004

Edelenbos, J., & Klijn, E.-H. (2007). Trust in Complex Decision-Making Networks: A Theoretical and Empirical Exploration. *Administration & Society*, *39*(1), 25–50. https://doi.org/10.1177/0095399706294460

Ehsan, U., Passi, S., Liao, Q. V., Chan, L., Lee, I.-H., Muller, M., & Riedl, M. O. (2021). *The Who in Explainable AI: How AI Background Shapes Perceptions of AI Explanations*. ArXiv. https://doi.org/10.48550/ARXIV.2107.13509

Emerson, R. W. (2015). Convenience Sampling, Random Sampling, and Snowball Sampling: How Does Sampling Affect the Validity of Research? *Journal of Visual Impairment & Blindness*, *109*(2), 164–168. https://doi.org/10.1177/0145482X1510900215

Faul, F., Erdfelder, E., Buchner, A., & Lang, A.-G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods*, *41*(4), 1149–1160. https://doi.org/10.3758/BRM.41.4.1149

Fischer, S., & Petersen, T. (2018). Was Deutschland über Algorithmen weiß und denkt: Ergebnisse einer repräsentativen Bevölkerungsumfrage. *Impuls Algorithmenethik*. https://doi.org/10.11586/2018022

Folger, R., & Konovsky, M. A. (1989). Effects of Procedural and Distributive Justice on Reactions to Pay Raise Decisions. *Academy of Management Journal*, *32*(1), 115–130. https://doi.org/10.2307/256422

Gilliland, S. W. (1993). The Perceived Fairness of Selection Systems: An Organizational Justice Perspective. *The Academy of Management Review*, *18*(4), 694–734. https://doi.org/10.2307/258595

Glikson, E., & Woolley, A. W. (2020). Human Trust in Artificial Intelligence: Review of Empirical Research. *Academy of Management Annals*, *14*(2), 627–660. https://doi.org/10.5465/annals.2018.0057

Goebel, R., Chander, A., Holzinger, K., Lecue, F., Akata, Z., Stumpf, S., Kieseberg, P., & Holzinger, A. (2018). Explainable AI: The New 42? In A. Holzinger, P. Kieseberg, A. M. Tjoa, & E. Weippl (Eds.), *Machine Learning and Knowledge Extraction* (Vol. 11015, pp. 295–303). Springer International Publishing. https://doi.org/10.1007/978-3-319-99740-7_21

Grayling, A. C. (2011). Psychology: How we form beliefs. *Nature*, *474*(7352), 446–447. https://doi.org/10.1038/474446a

Greszki, R., Meyer, M., & Schoen, H. (2015). Exploring the Effects of Removing "Too Fast" Responses and Respondents from Web Surveys. *Public Opinion Quarterly*, *79*(2), 471–503. https://doi.org/10.1093/poq/nfu058

Grzymek, V., & Puntschuh, M. (2019). Was Europa über Algorithmen weiß und denkt: Ergebnisse einer repräsentativen Bevölkerungsumfrage. *Impuls Algorithmenethik*. https://doi.org/10.11586/2019006

Hagras, H. (2018). Toward Human-Understandable, Explainable AI. *Computer*, *51*(9), 28–36. https://doi.org/10.1109/MC.2018.3620965

Hair, J. F., Black, Wiliam C., Babin, Barry J., & Anderson, Rolph E. (Eds.). (2010). *Multivariate data analysis* (7th ed). Prentice Hall.

Hayes, A. F. (2022). *Introduction to mediation, moderation, and conditional process analysis: A regression-based approach* (3rd edition). The Guilford Press.

Helm, J. M., Swiergosz, A. M., Haeberle, H. S., Karnuta, J. M., Schaffer, J. L., Krebs, V. E., Spitzer, A. I., & Ramkumar, P. N. (2020). Machine Learning and Artificial Intelligence: Definitions, Applications, and Future Directions. *Current Reviews in Musculoskeletal Medicine*, *13*(1), 69–76. https://doi.org/10.1007/s12178-020-09600-8

High-Level Expert Group on Artificial Intelligence. (2019). *Ethics Guidelines for Trustworthy AI*. European Commission. https://ec.europa.eu/futurium/en/ai-alliance-consultation.1.html

Hoffman, R. R., Mueller, S. T., Klein, G., & Litman, J. (2018). *Metrics for Explainable AI: Challenges and Prospects*. ArXiv. https://doi.org/10.48550/ARXIV.1812.04608

Hollander-Blumoff, R., & Tyler, T. R. (2008). Procedural Justice in Negotiation: Procedural Fairness, Outcome Acceptance, and Integrative Potential: Procedural Fairness, Outcome Acceptance, and Integrative Potential. *Law & Social Inquiry*, *33*(2), 473–500. https://doi.org/10.1111/j.1747-4469.2008.00110.x

Horowitz, M. C., Kahn, L., Macdonald, J., & Schneider, J. (2023). *Adopting AI: How Familiarity Breeds Both Trust and Contempt*. ArXiv. https://doi.org/10.48550/ARXIV.2305.01405

Jordan, S. L., Palmer, J. C., Daniels, S. R., Hochwarter, W. A., Perrewé, P. L., & Ferris, G. R. (2022). Subjectivity in fairness perceptions: How heuristics and self-efficacy shape the fairness expectations and perceptions of organisational newcomers. *Applied Psychology*, *71*(1), 103–128. https://doi.org/10.1111/apps.12313

Knowles, B., Richards, J. T., & Kroeger, F. (2022). *The Many Facets of Trust in AI: Formalizing the Relation Between Trust and Fairness, Accountability, and Transparency*. ArXiv. https://doi.org/10.48550/ARXIV.2208.00681

Köchling, A., & Wehner, M. C. (2020). Discriminated by an algorithm: A systematic review of discrimination and fairness by algorithmic decision-making in the context of HR recruitment and HR development. *Business Research*, *13*(3), 795–848. https://doi.org/10.1007/s40685-020-00134-w

Kocielnik, R., Amershi, S., & Bennett, P. N. (2019). Will You Accept an Imperfect AI?: Exploring Designs for Adjusting End-user Expectations of AI Systems. *Proceedings*

of the 2019 CHI Conference on Human Factors in Computing Systems, 1–14. https://doi.org/10.1145/3290605.3300641

Kordzadeh, N., & Ghasemaghaei, M. (2022). Algorithmic bias: Review, synthesis, and future research directions. *European Journal of Information Systems*, *31*(3), 388–409. https://doi.org/10.1080/0960085X.2021.1927212

Krishnan, R., Martin, X., & Noorderhaven, N. G. (2006). When Does Trust Matter to Alliance Performance? *Academy of Management Journal*, *49*(5), 894–917. https://doi.org/10.5465/amj.2006.22798171

Kulesza, T., Stumpf, S., Burnett, M., Yang, S., Kwan, I., & Wong, W.-K. (2013). Too much, too little, or just right? Ways explanations impact end users' mental models. *2013 IEEE Symposium on Visual Languages and Human Centric Computing*, 3–10. https://doi.org/10.1109/VLHCC.2013.6645235

Ladbury, C., Zarinshenas, R., Semwal, H., Tam, A., Vaidehi, N., Rodin, A. S., Liu, A., Glaser, S., Salgia, R., & Amini, A. (2022). Utilization of model-agnostic explainable artificial intelligence frameworks in oncology: A narrative review. *Translational Cancer Research*, *11*(10), 3853–3868. https://doi.org/10.21037/tcr-22-1626

Langer, M., König, C. J., & Hemsing, V. (2020). Is anybody listening? The impact of automatically evaluated job interviews on impression management and applicant reactions. *Journal of Managerial Psychology*, *35*(4), 271–284. https://doi.org/10.1108/JMP-03-2019-0156

Langer, M., & Landers, R. N. (2021). The future of artificial intelligence at work: A review on effects of decision automation and augmentation on workers targeted by algorithms and third-party observers. *Computers in Human Behavior*, *123*, Article 106878. https://doi.org/10.1016/j.chb.2021.106878

Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Sesing, A., & Baum, K. (2021). What do we want from Explainable Artificial Intelligence (XAI)? –

A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artificial Intelligence*, *296*, Article 103473. https://doi.org/10.1016/j.artint.2021.103473

Lee, J. D., & See, K. A. (2004). Trust in Automation: Designing for Appropriate Reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, *5*(1). https://doi.org/10.1177/2053951718756684

Lepri, B., Oliver, N., Letouzé, E., Pentland, A., & Vinck, P. (2018). Fair, Transparent, and Accountable Algorithmic Decision-making Processes: The Premise, the Proposed Solutions, and the Open Challenges. *Philosophy & Technology*, *31*(4), 611–627. https://doi.org/10.1007/s13347-017-0279-x

Lerner, J. S., Li, Y., Valdesolo, P., & Kassam, K. S. (2015). Emotion and Decision Making. *Annual Review of Psychology*, *66*(1), 799–823. https://doi.org/10.1146/annurev-psych-010213-115043

Leventhal, G. S. (1980). What Should Be Done with Equity Theory? In K. J. Gergen, M. S. Greenberg, & R. H. Willis (Eds.), *Social Exchange* (pp. 27–55). Springer US. https://doi.org/10.1007/978-1-4613-3087-5_2

Lind, E. A., Kray, L., & Thompson, L. (2001). Primacy Effects in Justice Judgments: Testing Predictions from Fairness Heuristic Theory. *Organizational Behavior and Human Decision Processes*, *85*(2), 189–210. https://doi.org/10.1006/obhd.2000.2937

Lind, E. A., & van den Bos, K. (2002). When fairness works: Toward a general theory of uncertainty management. *Research in Organizational Behavior*, *24*, 181–223. https://doi.org/10.1016/S0191-3085(02)24006-X

Lindebaum, D., Vesa, M., & den Hond, F. (2020). Insights From "The Machine Stops to Better Understand Rational Assumptions in Algorithmic Decision Making and Its Implications for Organizations. *Academy of Management Review*, *45*(1), 247–263. https://doi.org/10.5465/amr.2018.0181

Liu, B. (2021). In AI We Trust? Effects of Agency Locus and Transparency on Uncertainty Reduction in Human–AI Interaction. *Journal of Computer-Mediated Communication*, *26*(6), 384–402. https://doi.org/10.1093/jcmc/zmab013

Lund, B. D., Wang, T., Mannuru, N. R., Nie, B., Shimray, S., & Wang, Z. (2023). ChatGPT and a new academic reality: Artificial Intelligence-written research papers and the ethics of the large language models in scholarly publishing. *Journal of the Association for Information Science and Technology*, 570–581. https://doi.org/10.1002/asi.24750

Magner, N. R., Johnson, G. G., Sobery, J. S., & Welker, R. B. (2000). Enhancing Procedural Justice in Local Government Budget and Tax Decision Making. *Journal of Applied Social Psychology*, *30*(4), 798–815. https://doi.org/10.1111/j.1559-1816.2000.tb02824.x

Malik, A., Budhwar, P., Mohan, H., & N. R., S. (2022). Employee experience –the missing link for engaging employees: Insights from an MNE 's AI -based HR ecosystem. *Human Resource Management*, *62*(1), 97–115. https://doi.org/10.1002/hrm.22133

Marcinkowski, F., Kieslich, K., Starke, C., & Lünich, M. (2020). Implications of AI (un-)fairness in higher education admissions: The effects of perceived AI (un-)fairness on exit, voice and organizational reputation. *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, 122–130. https://doi.org/10.1145/3351095.3372867

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An Integrative Model of Organizational Trust. *The Academy of Management Review*, *20*(3), 709–734. https://doi.org/10.2307/258792

Merritt, S. M. (2011). Affective Processes in Human–Automation Interactions. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *53*(4), 356–370. https://doi.org/10.1177/0018720811411912

Miller, T. (2019). Explanation in artificial intelligence: Insights from the social sciences. *Artificial Intelligence*, *267*, 1–38. https://doi.org/10.1016/j.artint.2018.07.007

Miller, T., Howe, P., & Sonenberg, L. (2017). *Explainable AI: Beware of Inmates Running the Asylum Or: How I Learnt to Stop Worrying and Love the Social and Behavioural Sciences*. ArXiv. https://doi.org/10.48550/ARXIV.1712.00547

Morse, L., Teodorescu, M. H. M., Awwad, Y., & Kane, G. C. (2022). Do the Ends Justify the Means? Variation in the Distributive and Procedural Fairness of Machine Learning Algorithms. *Journal of Business Ethics*, *181*(4), 1083–1095. https://doi.org/10.1007/s10551-021-04939-5

Nass, C., & Moon, Y. (2000). Machines and Mindlessness: Social Responses to Computers. *Journal of Social Issues*, *56*(1), 81–103. https://doi.org/10.1111/0022-4537.00153

Newman, D. T., Fast, N. J., & Harmon, D. J. (2020). When eliminating bias isn't fair: Algorithmic reductionism and procedural justice in human resource decisions. *Organizational Behavior and Human Decision Processes*, *160*, 149–167. https://doi.org/10.1016/j.obhdp.2020.03.008

Noble, S. M., Foster, L. L., & Craig, S. B. (2021). The procedural and interpersonal justice of automated application and resume screening. *International Journal of Selection and Assessment*, *29*(2), 139–153. https://doi.org/10.1111/ijsa.12320

Nørskov, S., Damholdt, M. F., Ulhøi, J. P., Jensen, M. B., Ess, C., & Seibt, J. (2020). Applicant Fairness Perceptions of a Robot-Mediated Job Interview: A Video Vignette-Based Experimental Survey. *Frontiers in Robotics and AI*, *7*, Article 586263. https://doi.org/10.3389/frobt.2020.586263

Nourani, M., Kabir, S., Mohseni, S., & Ragan, E. D. (2019). The Effects of Meaningful and Meaningless Explanations on Trust and Perceived System Accuracy in Intelligent Systems. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, 7*, 97–105. https://doi.org/10.1609/hcomp.v7i1.5284

Omrani, N., Rivieccio, G., Fiore, U., Schiavone, F., & Agreda, S. G. (2022). To trust or not to trust? An assessment of trust in AI-based systems: Concerns, ethics and contexts. *Technological Forecasting and Social Change*, *181*, Article 121763. https://doi.org/10.1016/j.techfore.2022.121763

OpenAI. (2023). *GPT-4 Technical Report*. ArXiv. https://doi.org/10.48550/ARXIV.2303.08774

Ötting, S. K., & Maier, G. W. (2018). The importance of procedural justice in Human–Machine Interactions: Intelligent systems as new decision agents in organizations. *Computers in Human Behavior*, *89*, 27–39. https://doi.org/10.1016/j.chb.2018.07.022

Ouchchy, L., Coin, A., & Dubljević, V. (2020). AI in the headlines: The portrayal of the ethical issues of artificial intelligence in the media. *AI & SOCIETY*, *35*(4), 927–936. https://doi.org/10.1007/s00146-020-00965-5

Pal, S., Sharma, L., & Jangid, M. (2022). Trust and Technology Acceptance in Crossing-Decision of Pedestrian in Automated Vehicles. *2022 8th International Conference on Advanced Computing and Communication Systems (ICACCS)*, 609–614. https://doi.org/10.1109/ICACCS54159.2022.9785157

Pierson, E. (2018). *Demographics and Discussion influence views on algorithmic fairness*. ArXiv. https://doi.org/10.48550/ARXIV.1712.09124

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*, 1135–1144. https://doi.org/10.1145/2939672.2939778

Roy, C., Nourani, M., Honeycutt, D. R., Block, J. E., Rahman, T., Ragan, E. D., Ruozzi, N., & Gogate, V. (2021). Explainable activity recognition in videos: Lessons learned. *Applied AI Letters, 2*(4). https://doi.org/10.1002/ail2.59

Schaubroeck, J., May, D. R., & Brown, F. W. (1994). Procedural justice explanations and employee reactions to economic hardship: A field experiment. *Journal of Applied Psychology*, *79*(3), 455–460. https://doi.org/10.1037/0021-9010.79.3.455

Schoeffer, J., Machowski, Y., & Kuehl, N. (2021). *Perceptions of Fairness and Trustworthiness Based on Explanations in Human vs. Automated Decision-Making*. ArXiv. https://doi.org/10.48550/ARXIV.2109.05792

Schroeder, S. A., & Fulton, D. C. (2017). Voice, Perceived Fairness, Agency Trust, and Acceptance of Management Decisions Among Minnesota Anglers. *Society & Natural Resources, 30*(5), 569–584. https://doi.org/10.1080/08941920.2016.1238987

Shareef, M. A., Kumar, V., Dwivedi, Y. K., Kumar, U., Akram, M. S., & Raman, R. (2021). A new health care system enabled by machine intelligence: Elderly people's trust or losing self control. *Technological Forecasting and Social Change, 162*, Article 120334. https://doi.org/10.1016/j.techfore.2020.120334

Shepard, R. N. (1987). Toward a Universal Law of Generalization for Psychological Science. *Science, 237*(4820), 1317–1323. https://doi.org/10.1126/science.3629243

Shin, D. (2021). The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. *International Journal of Human-Computer Studies, 146*, Article 102551. https://doi.org/10.1016/j.ijhcs.2020.102551

Shulner-Tal, A., Kuflik, T., & Kliger, D. (2022). Fairness, explainability and in-between: Understanding the impact of different explanation methods on non-expert users' perceptions of fairness toward an algorithmic system. *Ethics and Information Technology, 24*(1), 2. https://doi.org/10.1007/s10676-022-09623-4

Shulner-Tal, A., Kuflik, T., & Kliger, D. (2023). Enhancing Fairness Perception – Towards Human-Centred AI and Personalized Explanations Understanding the Factors Influencing Laypeople's Fairness Perceptions of Algorithmic Decisions. *International Journal of Human–Computer Interaction*, 1–28. https://doi.org/10.1080/10447318.2022.2095705

Starke, C., Baleis, J., Keller, B., & Marcinkowski, F. (2022). Fairness perceptions of algorithmic decision-making: A systematic review of the empirical literature. *Big Data & Society*, *9*(2), 1–16. https://doi.org/10.1177/20539517221115189

Stevens, J. (2009). *Applied multivariate statistics for the social sciences* (5th ed). Routledge.

Suen, H.-Y., Chen, M. Y.-C., & Lu, S.-H. (2019). Does the use of synchrony and artificial intelligence in video interviews affect interview ratings and applicant attitudes? *Computers in Human Behavior*, *98*, 93–101. https://doi.org/10.1016/j.chb.2019.04.012

Szymanski, M., Millecamp, M., & Verbert, K. (2021). Visual, textual or hybrid: The effect of user expertise on different explanations. *26th International Conference on Intelligent User Interfaces*, 109–119. https://doi.org/10.1145/3397481.3450662

Tešić, M., & Hahn, U. (2022). Can counterfactual explanations of AI systems' predictions skew lay users' causal intuitions about the world? If so, can we correct for that? *Patterns*, *3*(12), Article 100635. https://doi.org/10.1016/j.patter.2022.100635

Thau, S., Bennett, R. J., Mitchell, M. S., & Marrs, M. B. (2009). How management style moderates the relationship between abusive supervision and workplace deviance: An uncertainty management theory perspective. *Organizational Behavior and Human Decision Processes, 108*(1), 79–92. https://doi.org/10.1016/j.obhdp.2008.06.003

Thibaut, J. W., & Walker, L. (1975). *Procedural justice: A psychological analysis*. L. Erlbaum Associates, Hillsdale.

Tsuchiya, A., Miguel, L. S., Edlin, R., Wailoo, A., & Dolan, P. (2005). Procedural Justice in Public Healthcare Resource Allocation: *Applied Health Economics and Health Policy*, *4*(2), 119–127. https://doi.org/10.2165/00148365-200504020-00006

van den Bos, K., & Lind, E. A. (2002). Uncertainty management by means of fairness judgments. In *Advances in Experimental Social Psychology* (Vol. 34, pp. 1–60). Elsevier. https://doi.org/10.1016/S0065-2601(02)80003-X

van den Bos, K., Lind, E. A., & Wilke, H. A. M. (2001). The Psychology of Procedural and Distributive Justice Viewed from the Perspective of Fairness Heuristic Theory. In R. Cropanzano (Ed.), *Justice in the workplace: From theory to practice* (pp. 49–66). Lawrence Erlbaum Associates Publishers.

van den Bos, K., Wilke, H. A. M., & Lind, E. A. (1998). When do we need procedural fairness? The role of trust in authority. *Journal of Personality and Social Psychology*, *75*(6), 1449–1458. https://doi.org/10.1037/0022-3514.75.6.1449

Vereschak, O., Bailly, G., & Caramiaux, B. (2021). How to Evaluate Trust in AI-Assisted Decision Making? A Survey of Empirical Methodologies. *Proceedings of the ACM on Human-Computer Interaction*, *5*(CSCW2), 1–39. https://doi.org/10.1145/3476068

Viklund, M., & Sjöberg, L. (2008). An Expectancy-Value Approach to Determinants of Trust. *Journal of Applied Social Psychology*, *38*(2), 294–313. https://doi.org/10.1111/j.1559-1816.2007.00306.x

Wang, R., Harper, F. M., & Zhu, H. (2020). *Factors Influencing Perceived Fairness in Algorithmic Decision-Making: Algorithm Outcomes, Development Procedures, and Individual Differences.* ArXiv. https://doi.org/10.48550/ARXIV.2001.09604

Wang, X., Guchait, P., Lee, J., & Back, K.-J. (2019). The importance of psychological safety and perceived fairness among hotel employees: The examination of antecedent and outcome variables. *Journal of Human Resources in Hospitality & Tourism*, *18*(4), 504–528. https://doi.org/10.1080/15332845.2019.1626964

Wierzbicki, A. (2010). Theory of Trust and Fairness. In J. Kacprzyk (Ed.), *Trust and Fairness in Open, Distributed Systems* (Vol. 298, pp. 11–69). Springer Berlin Heidelberg. https://doi.org/10.1007/978-3-642-13451-7_2

Woods, S. A., Ahmed, S., Nikolaou, I., Costa, A. C., & Anderson, N. R. (2020). Personnel selection in the digital age: A review of validity and applicant reactions, and future research challenges. *European Journal of Work and Organizational Psychology*, *29*(1), 64–77. https://doi.org/10.1080/1359432X.2019.1681401

Yang, S. C.-H., Folke, T., & Shafto, P. (2022). *A Psychological Theory of Explainability*. ArXiv. https://doi.org/10.48550/ARXIV.2205.08452

Yeomans, M., Shah, A., Mullainathan, S., & Kleinberg, J. (2019). Making sense of recommendations. *Journal of Behavioral Decision Making*, *32*(4), 403–414. https://doi.org/10.1002/bdm.2118

Zhang, L., & Amos, C. (2023). Dignity and use of algorithm in performance evaluation. *Behaviour & Information Technology*, 1–18. https://doi.org/10.1080/0144929X.2022.2164214

Zhang, Y., Song, K., Sun, Y., Tan, S., & Udell, M. (2019). *"Why Should You Trust My Explanation?" Understanding Uncertainty in LIME Explanations*. ArXiv. https://doi.org/10.48550/ARXIV.1904.12991

Złotowski, J., Proudfoot, D., Yogeeswaran, K., & Bartneck, C. (2015). Anthropomorphism: Opportunities and Challenges in Human–Robot Interaction. *International Journal of Social Robotics*, *7*(3), 347–360. https://doi.org/10.1007/s12369-014-0267-6

Appendix A

Main Study

Table A1

Decision-making Scenario for each Condition

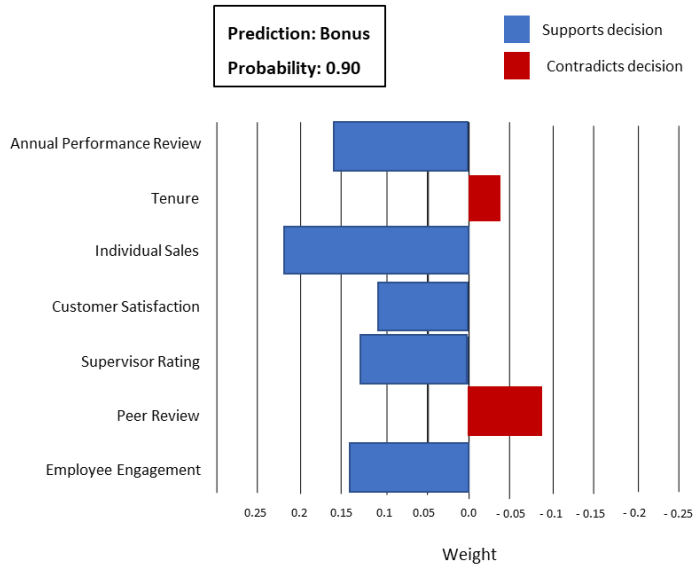
HDM condition	ADM condition
Company X just went through the process of making its end of the year bonus allocation. In order to determine whether each employee should receive a bonus or not, Company X relied on the manager, who took into account a variety of factors. After the manager made a series of deliberations, they determined whether each employee should receive a bonus or not.	Company X just went through the process of making its end of the year bonus allocation. In order to determine whether each employee should receive a bonus or not, Company X relied on an algorithm (a computerized decision-making tool) that took into account a variety of factors. After the algorithm made a series of computations, it determined whether each employee should receive a bonus or not.

Note. Original decision scenario is from Newman et al., (2020).

Figure A1

Visual Explanation from the Main Study

The following image shows an example of how the algorithm/manager makes a decision for one particular employee. Whether the employee receives a bonus or not is based on the factors below and the numbers represent the weight (or how much impact) a factor has on the final decision. The Probability in the upper box indicates the confidence of the algorithm/manager that the decision is correct.



Appendix B

Pilot Study

Figure B1

Visual Explanation Version A from the Pilot Study

The following image shows an example of how the algorithm makes a decision for one particular employee. Whether the employee receives a bonus or not is based on the factors below and the numbers represent the weight a factor has on the final decision. The prediction probabilities in the upper left corner indicate the confidence of the algorithm that the decision is correct.

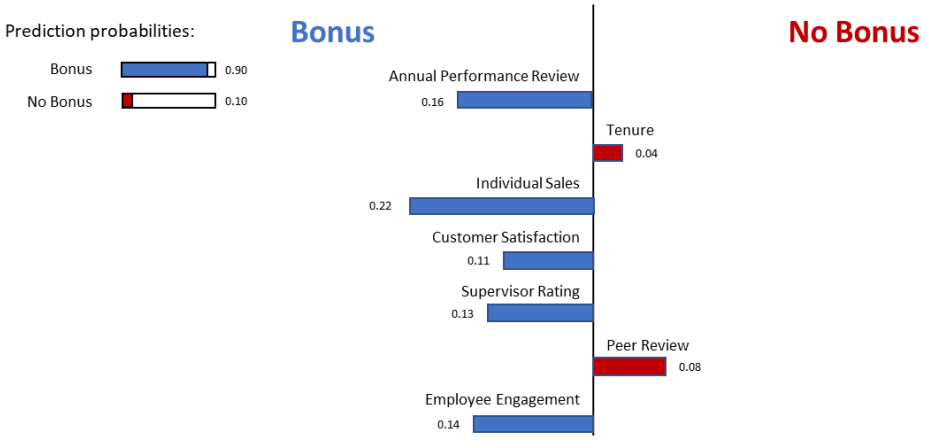


Figure B2

Visual Explanation Version B from the Pilot Study

The following image shows an example of how the algorithm makes a decision for one particular employee. Whether the employee receives a bonus or not is based on the factors below and the numbers represent the weight a factor has on the final decision. The Probability in the upper box indicates the confidence of the algorithm that the decision is correct.

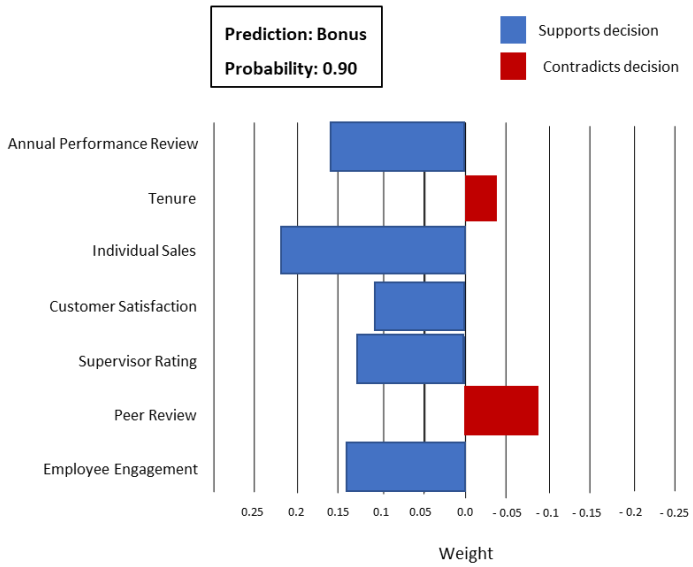
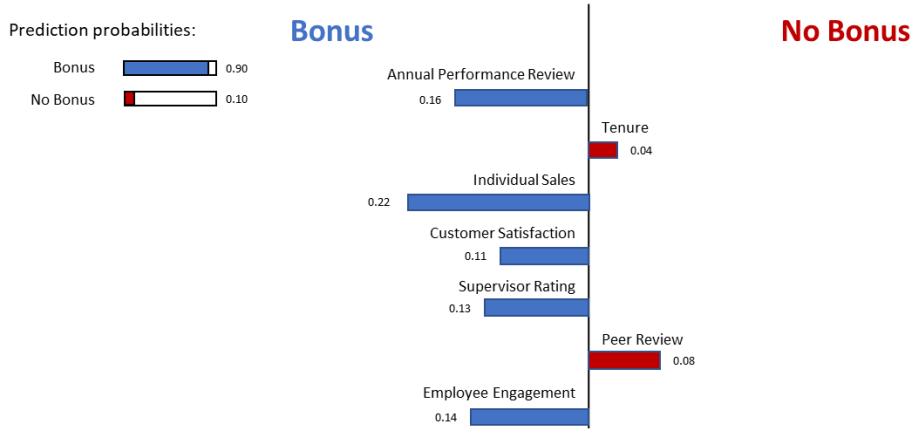


Figure B3

Visual Explanation Version A with a Case Example from the Pilot Study

The following image shows an example of how the algorithm makes a decision for one particular employee. Whether the employee receives a bonus or not is based on the factors below and the numbers represent the weight a factor has on the final decision. The prediction probabilities in the upper left corner indicate the confidence of the algorithm that the decision is correct.

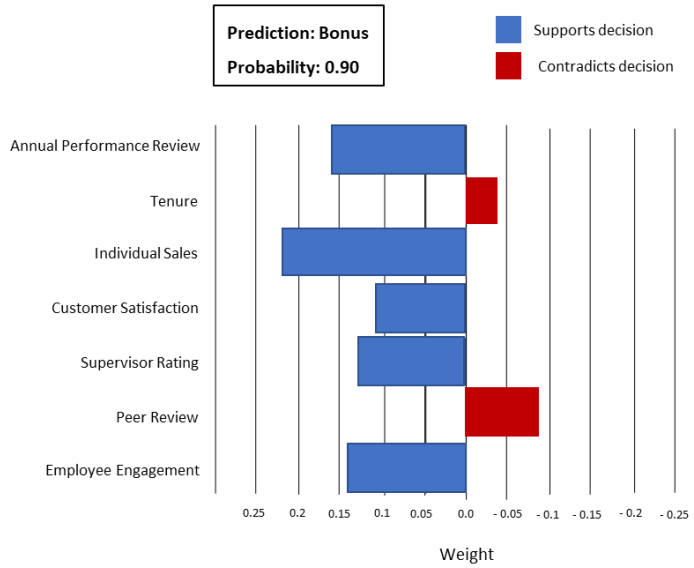


Example: For this case the algorithm predicts that the employee should receive a bonus with a 90% confidence that this decision is correct. The most important factors that support this decision are individual sales and annual performance review which have a weight of 0.22 and 0.16 respectively. The factors tenure and peer review in this case have a negative impact on the decision.

Figure B4

Visual Explanation Version B with a Case Example from the Pilot Study

The following image shows an example of how the algorithm makes a decision for one particular employee. Whether the employee receives a bonus or not is based on the factors below and the numbers represent the weight a factor has on the final decision. The Probability in the upper box indicates the confidence of the algorithm that the decision is correct.



Example: For this case the algorithm predicts that the employee should receive a bonus with a 90% confidence that this decision is correct. The most important factors that support this decision are individual sales and annual performance review which have a weight of 0.22 and 0.16 respectively. The factors tenure and peer review in this case have a negative impact on the decision.

Appendix C

Scales

Table C1

Items of the Perceived Procedural Fairness Scale

Original Items
1. In my opinion, the outcome of the algorithm's/manager's decision was fair.
2. The process by which the algorithm/manager made this decision was fair.
3. I am satisfied with the way in which the algorithm/manager made the decision.
4. The algorithm/manager made this decision in an unbiased and neutral manner.
5. The algorithm/manager treated all employees with dignity and respect in making this decision.

Note. Original scale is from Conlon et al., (2004).

Table C2

Items of the Trust in Decision maker Scale

Original Items	Adapted Items
1. I believe the AWD (Automatic Weapons Detector) is a competent performer.	1. I believe the manager/algorithm is a competent performer.
2. I trust the AWD.	2. I trust the manager/algorithm.
3 I have confidence in the advice given by the AWD.	3. I have confidence in the decision given by the manager/algorithm.
4. I can depend on the AWD.	4. I can depend on the manager/algorithm.
5. I can rely on the AWD to behave in consistent ways	5. I can rely on the manager/algorithm to behave in consistent ways.
6. I can rely on the AWD to do its best every time I take its advice.	6. I can rely on the manager/algorithm to do their/its best every time they/it makes a decision.

Note. Original scale is from Merritt (2011).

Table C3

Items of the Explainability Scale

Original Items	Adapted Items
1. I found algorithm are easily understandable.	1. I found the algorithm's/manager's decision process easily understandable.
2. I think the algorithm services are interpretable.	2. I think the algorithm's/manager's decision process is interpretable.
3. I can figure out the internal mechanics of a machine learning. I hope that algorithm can be clearly explainable.	3. I can figure out the internal mechanics of the algorithm's/manager's decision process.
	4. I think the algorithm's/manager's decision process is clearly explained.

Note. Original scale is from Shin (2021).

Table C4

Items of the AI literacy Scale

Items	Anchor
1. I can make use of programming to solve a problem.	(1) *Strongly disagree*, (2) *Disagree*, (3) *Somewhat disagree*, (4) *neither disagree nor agree*, (5) *Somewhat agree*, (6) *Agree*, (7) *Strongly agree*
2. I can understand statistical concepts like "error".	(1) *Strongly disagree*, (2) *Disagree*, (3) *Somewhat disagree*, (4) *neither disagree nor agree*, (5) *Somewhat agree*, (6) *Agree*, (7) *Strongly agree*
3. I understand how my navigation system calculates my time of arrival.	(1) *Strongly disagree*, (2) *Disagree*, (3) *Somewhat disagree*, (4) *neither disagree nor agree*, (5) *Somewhat agree*, (6) *Agree*, (7) *Strongly agree*
4. I understand how my email provider's spam filter works.	(1) *Strongly disagree*, (2) *Disagree*, (3) *Somewhat disagree*, (4) *neither disagree nor agree*, (5) *Somewhat agree*, (6) *Agree*, (7) *Strongly agree*
5. I understand how the recommendation system of internet platforms like Amazon, Google, or Facebook work.	(1) *Strongly disagree*, (2) *Disagree*, (3) *Somewhat disagree*, (4) *neither disagree nor agree*, (5) *Somewhat agree*, (6) *Agree*, (7) *Strongly agree*
6. How much programming knowledge do you have?	(1) *No knowledge*, (2) *A little knowledge – I know basic concepts in programming*, (3) *Some knowledge – I have coded a few programs before*, (4) *A lot of knowledge – I code programs frequently*
7. How much knowledge of computer algorithms do you have?	(1) *No knowledge*, (2) *A little knowledge – I know basic concepts in algorithms*, (3) *Some knowledge – I used algorithms before*, (4) *A lot of knowledge – I apply algorithms frequently to my work or I create algorithms*
8. How much statistical knowledge do you have?	(1) *No knowledge*, (2) *A little knowledge – I know basic concepts in statistics*, (3) *Some knowledge – I have used statistics a few times before*, (4) *A lot of knowledge – I use statistics frequently*

Note. Original scale is from Wang et al., (2020).

Table C5

Items of the Explanation Satisfaction Scale

Original Items
1. From the explanation, I understand how the algorithm works.
2. This explanation of how the algorithm works is satisfying.
3. This explanation of how the algorithm works has sufficient detail.
4. This explanation of how the algorithm works seems complete.
5. This explanation of how the algorithm works tells me how to use it.
6. This explanation of how the algorithm works is useful to my goals.
7. This explanation of the algorithm shows me how accurate the algorithm is.
8. This explanation lets me judge when I should trust and not trust the algorithm

Note. Original scale is from Hoffman et al., (2018).