frontiers | Frontiers in Computer Science

TYPE Original Research
PUBLISHED 17 July 2023
DOI 10.3389/fcomp.2023.1151150

Check for updates

OPEN ACCESS

EDITED BY
Chathurika S. Wickramasinghe Brahmana,
Capital One, United States

REVIEWED BY
Martin Gjoreski,
University of Italian Switzerland, Switzerland
Antonis Bikakis,
University College London, United Kingdom
Nadine Schlicker,
University of Marburg, Germany

*CORRESPONDENCE
Nicolas Scharowski
✉ nicolas.scharowski@unibas.ch

RECEIVED 25 January 2023
ACCEPTED 27 June 2023
PUBLISHED 17 July 2023

CITATION
Scharowski N, Perrig SAC, Svab M, Opwis K and Brühlmann F (2023) Exploring the effects of human-centered AI explanations on trust and reliance. *Front. Comput. Sci.* 5:1151150. doi: 10.3389/fcomp.2023.1151150

Exploring the effects of human-centered AI explanations on trust and reliance

Nicolas Scharowski*, Sebastian A. C. Perrig, Melanie Svab, Klaus Opwis and Florian Brühlmann

Human-Computer Interaction Research Group, Center for General Psychology and Methodology, Faculty of Psychology, University of Basel, Basel, Switzerland

Transparency is widely regarded as crucial for the responsible real-world deployment of artificial intelligence (AI) and is considered an essential prerequisite to establishing trust in AI. There are several approaches to enabling transparency, with one promising attempt being human-centered explanations. However, there is little research into the effectiveness of human-centered explanations on end-users' trust. What complicates the comparison of existing empirical work is that trust is measured in different ways. Some researchers measure subjective trust using questionnaires, while others measure objective trust-related behavior such as reliance. To bridge these gaps, we investigated the effects of two promising human-centered *post-hoc* explanations, *feature importance* and *counterfactuals*, on trust *and* reliance. We compared these two explanations with a control condition in a decision-making experiment (N = 380). Results showed that human-centered explanations can significantly increase reliance but the type of decision-making (increasing a price vs. decreasing a price) had an even greater influence. This challenges the presumed importance of transparency over other factors in human decision-making involving AI, such as potential heuristics and biases. We conclude that trust does not necessarily equate to reliance and emphasize the importance of appropriate, validated, and agreed-upon metrics to design and evaluate human-centered AI.

KEYWORDS

AI, XAI, HCXAI, trust, reliance, transparency, explainability, interpretability

1. Introduction

It is generally recognized that computers perform specific tasks better than humans, such as numeracy, logical reasoning, or storing information (Solso et al., 2005). But with the recent breakthroughs in artificial intelligence (AI), domains that used to be exclusively associated with human competence and considered computationally unattainable are likewise being challenged by machines. AI has led to improvements in speech recognition, image classification, as well as object detection (LeCun et al., 2015) and is now increasingly used in various everyday applications such as video surveillance, email spam filtering, online customer support, and product recommendations. Because of this general applicability and the potential manifold consequences, voices are being raised that AI should satisfy criteria like fairness, reliability, accountability, and transparency (Ehsan et al., 2021b; ACM FAccT Conference, 2022). The call for transparent AI has led to the multidisciplinary research field of explainable artificial intelligence (XAI), which explores methods and models that make the behaviors, predictions, and decisions of AI transparent and understandable to humans (Lipton, 2018a; Liao et al., 2020). Abdul et al. (2018), as well as Biran and Cotton (2017) have pointed out that the development of transparent systems has long been a research focus, originating from expert systems, intelligent agents, recommender systems, context-aware systems, and other adjacent fields such as automation.

Despite this rich history, current XAI research faces unprecedented challenges as AI is increasingly complex and thus more cumbersome to render transparent (Biran and Cotton, 2017). In the pursuit of ever more accurate predictions, modern AI consists of millions of interdependent values and parameters, resulting in a trade-off between complexity and transparency (Shmueli, 2010; Mittelstadt et al., 2019). Because of this complexity, AI is often characterized by the opaque box paradigm (Suresh et al., 2021), meaning that AI can only be considered in terms of its inputs and outputs without direct observations of its inner workings. This opacity makes it more challenging than ever to ensure fairness, reliability, and accountability, rendering transparency a prerequisite for the other three criteria. Some researchers argue that transparency helps verify and improve the functionality of a system (i.e., for debugging), supports developers in learning from a system (i.e., in generating hypotheses), or is needed to ensure fair and ethical decision-making (Mittelstadt et al., 2019). Others believe that transparency contributes toward building a relationship of trust between humans and AI (Stephanidis et al., 2019), which plays a key role in people's willingness to rely on automated systems (Hoff and Bashir, 2015).

While transparency is generally considered crucial for the effective and responsible real-world deployment of AI, there are various transparency approaches tailored to the algorithm's goal, the context, and the involved stakeholders, such as developers, decision-makers, and end-users (Ehsan et al., 2019; Samek et al., 2019). For end-users, the requirements and purpose of transparency are expected to be distinct (Cheng et al., 2019; Langer et al., 2021; Suresh et al., 2021), and Miller (2019) specified criteria that should be taken into account in order to achieve human-centered explainable AI (HCXAI). Following the notion of Ehsan and Riedl (2020), we understand HCXAI as an approach that puts humans at the center of technology design. Within this framework, not only is it important to conduct user studies that validate XAI methods with ordinary end-users, but also to consider explanations designed to account for human needs. We argue that Miller (2019)'s criteria and the focus on how humans explain decisions to one another are a good starting point for meaningful AI explanations to end-users. However, empirical investigations of the effects of human-centered explanations satisfying these criteria are sparse, and there is mixed evidence about whether transparency is in fact increasing trust (Cramer et al., 2008; Nothdurft et al., 2013; Cheng et al., 2019; Ehsan et al., 2019; Zhang et al., 2020; Poursabzi-Sangdeh et al., 2021). These ambiguous findings may arise from the use of proxy-tasks rather than actual decision-making tasks when evaluating AI systems (Buçinca et al., 2020) and from varying conceptualizations of trust. Studies on XAI appear to define and measure trust differently (Vereschak et al., 2021). Some researchers assess attitudinal trust measures via questionnaires (Buçinca et al., 2020), while others focus on trust-related behavior such as reliance (Poursabzi-Sangdeh et al., 2021). However, research has shown that subjective trust can be a poor predictor of actual reliance (Dzindolet et al., 2003; Miller et al., 2016; Papenmeier et al., 2022). Therefore, it seems particularly important to distinguish between attitudinal and behavioral measures when studying the effect of transparency on trust (Parasuraman and Manzey, 2010; Sanneman and Shah, 2022; Scharowski et al., 2022).

In this study, we focus on explainability as a means of AI transparency. Explainability, in this context, is the process of explaining how an opaque box AI arrived at a particular result or decision after a computation has been performed (i.e., *post-hoc* explanations), without directly revealing the AI's internal mechanisms via visualizations or graphical interfaces, as typically aimed for in clear box AI. Grounded in Miller's work, we identified *feature importance explanations* and *counterfactual explanations* as two promising *post-hoc* explanations for achieving HCXAI. We conducted an online decision-making experiment (N = 380) on Amazon Mechanical Turk (MTurk) to investigate the effect of those two human-centered explanations on end-users' trust and reliance with a control condition. The *Trust between People and Automation Scale* (TPA, Jian et al., 2000) served as an attitudinal measure of AI trust. Reliance on the AI recommendation, captured by *weight of advice*, provided a measure for trust-related behavior. The results suggest that the relationship between transparency and reliance is more nuanced than commonly assumed and emphasize the importance of adequately differentiating between trust and reliance and their respective measurements when evaluating XAI. While transparency did not affect trust, reliance increased through human-centered *post-hoc* explanations, but only for specific decision-making tasks. In the particular context we examined, it appears that the type of decision-making participants were facing (increasing a price vs. decreasing a price) had a greater influence on reliance than how the AI explained its recommendation to the end-users. This suggests that humans display cognitive biases and apply heuristics in decision-making tasks that involve AI recommendations. If biased human decision-making prevails, AI may not support people to reach better decisions. The XAI community should consider potential biases and heuristics for a more nuanced understanding of the human-AI interaction. It remains to be further explored whether measuring attitudinal trust via questionnaires reflects trust-related behavior (i.e., reliance) appropriately and whether heuristics and biases also have an impact on trust. If researchers and practitioners who develop and evaluate AI systems assess only subjective trust, they may not draw valid conclusions about actual AI reliance and vice versa. Given that AI is increasingly utilized to make critical decisions with far-reaching consequences, adopting agreed-upon, validated, and appropriate measurements in XAI is of paramount importance.

2. Related work

2.1. Human-centered explanations

Two closely related terms that are often used interchangeably should be distinguished when referring to AI transparency: explainability and interpretability. While both terms refer to methods for achieving transparency, they differ in their approach to implementing transparency. For Lipton (2018b), interpretability is the information that a system provides about its inner workings and associated with the notion of *clear box* AI, meaning AI whose internal mechanisms are accessible and not concealed. Interpretability is thus achieved by using or designing AI in a way that its decision-making can be directly observed or otherwise visualized. Explainability, on the other hand, implies accepting *opaque box* AI whose internal mechanisms are not readily accessible or understandable, and providing meaningful information by explaining how a specific output or decision was reached after a computation has been carried out. In this sense, explainability is *post-hoc* interpretability (Lipton, 2018b; Ehsan et al., 2019; Miller, 2019; Mohseni et al., 2020).

In addition to this distinction between explainability and interpretability, XAI researchers need to be aware of the varying needs and goals different stakeholders have when interpreting, understanding, and reacting to explanations coming from AI (Suresh et al., 2021). Past research has raised concerns that AI explanations are frequently based on the intuition of researchers, AI developers, and experts rather than addressing the needs of end-users (Du et al., 2019; Miller, 2019). A growing body of work has engaged with this challenge (Ferreira and Monteiro, 2020; Hong et al., 2020; Liao et al., 2020; Ehsan et al., 2021a) and now focuses on more human-centered approaches that align AI explanations with people's needs. Despite these considerations regarding human-centered explanations, previous work on AI transparency has often placed a greater emphasis on interpretability (i.e., model visualization for clear box AI) than on explainability (i.e., *post-hoc* explanations for opaque box AI) (Kulesza et al., 2015; Krause et al., 2016; Cheng et al., 2019; Kocielnik et al., 2019; Lai and Tan, 2019; Poursabzi-Sangdeh et al., 2021). This emphasis has led to focusing on graphical interfaces that allow users to observe and understand the decision-making processes of these models more directly. While *post-hoc* explanations also require some sort of user interface or visualization, they operate at a more abstract level and provide a simplified or approximate representation of the decision-making process rather than direct access to the internal workings of the model, as interpretability seeks to accomplish. However, some researchers have questioned that interpretability approaches are useful to all people equally. Suresh et al. (2021) and Lipton (2018b) argue that explainability might be more reflective of the way that humans are transparent about their own decisions. When it comes to humans, the exact processes by which our brains form decisions and our explanations regarding those decisions are distinct (Lipton, 2018b). Similar to how people provide explanations to one another, AI might explain its decisions without disclosing the computation underlying them. Because of its proximity to how humans reason about their decisions, explainability seems promising to achieve HCXAI if the way humans provide and understand explanations is taken into account.

With regard to human-centered explanations, researchers have emphasized the importance of incorporating insights from philosophy, the social sciences, and psychology on how people define, generate, select, evaluate, and present explanations into the field of XAI (Miller, 2019; Mittelstadt et al., 2019). Based on findings from these areas of research, Miller (2019) defined certain criteria for what contributes to a meaningful explanation for people, including *selectivity* (providing the most important reasons for a decision), *contrastivity* (providing contrastive information with a decision), and *sociality* (explaining something in a similar way to how humans explain their actions). Miller (2019) and Mittelstadt et al. (2019) argued that explanations from AI should at least fulfill some of these criteria to be meaningful for end-users. Adadi and Berrada (2018) identified over 17 different transparency approaches that are being proposed in the current XAI literature. Based on Miller (2019)'s criteria, we narrowed down Adadi and Berrada (2018)'s selection and identified two promising human-centered *post-hoc* explanations: *feature importance explanations* and *counterfactual explanations*.

Feature importance explanations. Humans rarely expect a complete explanation for a decision and often select the most important or immediate cause from a sometimes infinite number of reasons (Miller, 2019). As the name suggests, feature importance allows end-users to determine which features are most important for an AI's output. Such explanations thus satisfy the selectivity criterion proposed by Miller (2019) because they show how certain factors influenced a decision. Feature importance explanations have the following notation: "Outcome P was returned because variable V had values (vi, vii, ...) associated with it" (Wachter et al., 2018, p. 9).

Counterfactual explanations. Humans usually ask why a particular decision was made instead of another one (Miller, 2019). In addition to the leading causes of an output, counterfactuals provide contrastive "what-if" statements that help identify what might be changed in the future to achieve a desired output (Mothilal et al., 2020). Counterfactuals combine Miller's selectivity and contrastivity criteria. They are expected to have psychological benefits because they help people act, rather than merely understand, by altering future behavior to achieve a desired outcome (Wachter et al., 2018; Mothilal et al., 2020). Counterfactuals commonly have the following notation: "Outcome P was returned because variable V had values (vi, vii, ...) associated with it. If V had values (vi', vii', ...) instead, outcome P' would have been returned" (Wachter et al., 2018, p.9).

Both explanations also seem to meet Miller (2019)'s sociality criterion. For humans, explanations are a form of social interaction or, more specifically, a transfer of knowledge often presented as part of a conversation between the explainer and the explainee that is subject to the rules of conversation (Hilton, 1990; Miller, 2019). Although Miller (2019) points out that this does not imply that explanations must be given in natural language, we expect natural language explanations to be a promising approach for human-centered explanations because they are accessible and intuitive to humans (Ehsan et al., 2019). De Graaf and Malle (2018) argued that because people attribute human-like traits to artificial agents, they might expect them to provide explanations similar to how humans explain their actions. Szymanski et al. (2021) showed that while end-users prefer visual over textual explanations, they performed significantly worse with the former, and Kizilcec (2016) demonstrated that short textual explanations build subjective trust in an algorithm's decision. There are also jurisdictional reasons for explanations in natural language. They comply with the EU's GDPR (Wachter et al., 2018) and align with the regulatory requirement for automated decision-making to explain decisions in an "easily accessible form, using clear and plain language [...] provided in writing." (European Parliament and Council of the European Union, 2016, article 12). To the best of our knowledge, there is little to no empirical research on the effectiveness of these two human-centered explanations derived from the literature (i.e., Adadi and Berrada, 2018) using Miller's criteria in fostering end-users' trust in AI. Therefore, an empirical investigation into the efficacy of *feature importance explanations* and *counterfactual explanations* seems warranted.

2.2. Trust in XAI

Within the XAI community, researchers define and measure trust in different ways, and there does not appear to be a clear consensus about the desired effect of trust or a clear differentiation of the factors that contribute to trust (Chopra and Wallace, 2003; Mohseni et al., 2020). To provide two examples: Lai and Tan (2019) proposed a spectrum between full human agency and full automation, with varying levels of explanations along this spectrum. In a deception detection task (asking end-users to decide whether a hotel review is genuine or deceptive), they illustrated that heatmaps of relevant instances and example-based explanations improved human performance and increased the trust humans place on the predictions of the AI. Lai and Tan defined trust as the percentage of instances for which humans relied on the machine prediction. In contrast, Cheng et al. (2019) conducted an experiment where participants used different UI interfaces to comprehend an algorithm's decision for university admissions. They showed that revealing the inner workings of an algorithm can improve users' comprehension and found that users' subjective trust, assessed by a 7-point Likert scale, was not affected by the explanation interface. These two empirical studies exemplify how trust is measured differently in XAI research. These discrepancies could be a reason for the inconclusive findings in current XAI literature regarding the effect of transparency on trust. This warrants a more precise definition and rigorous distinction between trust and related concepts, such as reliance, in empirical studies investigating the relationship between transparency and trust.

The differentiation between subjective and objective trust and their measurement in XAI was addressed by Mohseni et al. (2020). They pointed out that subjective trust measures include interviews, surveys, and self-reports via questionnaires, which according to Buçinca et al. (2020) have been the focal points for evaluating AI transparency. For objective measures of trust, Mohseni et al. (2020) proposed users' perceived system competence, understanding, and users' reliance on a system. This distinction between trust and reliance was emphasized by Hartmann (2020). They argued that the everyday use of the word *trust* is misleading when applied to technology and that, in this case, trust must be differentiated from *reliance*. Hartmann (2020) was not the only one that distinguished between trust and reliance as other researchers have shown that attitudinal judgments have an impact on people's intention to rely on automated systems (Cramer et al., 2008; Merritt, 2011); People tend to rely on automation they trust and reject automation they distrust (Lee and See, 2004). This makes trust particularly relevant in the misuse (overreliance Parasuraman and Riley, 1997) and disuse (neglect or underutilization Parasuraman and Riley, 1997) of automation (Hoff and Bashir, 2015; Yu et al., 2017; Stephanidis et al., 2019). To avoid such instances, users' trust needs to be calibrated or warranted. Trust calibration refers to the extent to which the trust that users place in the system is adequate to the system's actual capabilities (Wischnewski et al., 2023). Fostering end users' trust in AI should aim to attain an appropriate level of trust to avoid overreliance or underutilization of AI systems. According to Lee and See (2004) and correspondent with Hoff and Bashir (2015), we define trust in the context of AI as "the attitude that an agent will help achieve an individual's goals in a situation characterized by uncertainty and vulnerability" (Lee and See, 2004, p. 6). This definition encapsulates the notion of uncertainty and vulnerability as proposed by Jacovi et al. (2021) and Mayer et al. (1995), which is the most widely used and accepted definition of trust (Rousseau et al., 1998). Adopting Lee and See's definition and model, we distinguish between trust and reliance and think of trust as an *attitude* and reliance as a *behavior* that follows the level of trust. In their work on metrics for XAI, Hoffman et al. (2019) make a similar distinction when they differentiate between trusting a machine's output and following its advice. In this framework, attitudes and behaviors remain conceptually distinct and do not share a deterministic but a probabilistic relationship (Ajzen and Fishbein, 1980; Körber, 2018). Even if an AI system is trusted, reliance must not necessarily follow (Kirlik, 1993; Körber, 2018), and people may claim to trust an AI system, yet behave in a way that suggests they do not (Miller et al., 2016). This implies that attitudes may not always translate into behaviors. The empirical findings of Dzindolet et al. (2003) support this argument: although some automated decision aids were rated as more trustworthy than others, all were equally likely to be relied upon.

Given these possible contradictions, we think it is useful to conceptualize trust as an antecedent to reliance that guides but does not determine it (Lee and See, 2004). However, this consideration has been insufficiently taken into account in past research (Papenmeier et al., 2022; Scharowski et al., 2022). A rigorous distinction and an accurate conceptualization of trust and reliance are vital for empirical XAI studies since researchers who evaluate AI systems using only subjective measures of AI trust might not draw valid conclusions about actual reliance on AI and vice versa. Arrieta et al. (2020) emphasized that only agreed-upon metrics and their respective measurements allow for meaningful comparisons in XAI research and that without such consistency, any claims are elusive and do not provide a solid foundation. For this reason, we decided to investigate two alternative methodological approaches, namely, measuring attitudinal trust on the one hand and measuring trust-related behavior in terms of reliance on the other hand. This approach is in line with Sanneman and Shah (2022), that recommended using trust scales in conjunction with behavior-based metrics to determine if people appropriately trust *and* use AI systems in response to AI explanations they provide.

3. Empirical investigation

3.1. Research question and hypotheses

We investigated the following research question:

RQ: What effect do human-centered explanations have on end-users' trust and reliance?

To answer this research question, we compared the previously introduced feature importance and counterfactual *post-hoc* explanations with a control in a scenario in which participants had to estimate subleasing prices for different apartments. We employed a mixed study design with a 3 (explanation condition: feature importance vs. counterfactual vs. control) × 2 (type of AI recommendation: increasing price vs. decreasing price). Explanation condition was the between-subject factor, type of recommendation was the within-subject factor. Following Poursabzi-Sangdeh et al. (2021), we focused on the domain of real estate valuation, where machine learning is often used to predict apartment prices. Airbnb (https://airbnb.com) and Zillow (http://zillow.com) are examples of websites that provide price recommendations to end-users in this way. Considering the previous clarifications, we expected that trust and reliance are influenced by human-centered explanations similarly but should be treated as distinct concepts. We, therefore, formulated separate hypotheses for both trust and reliance. We further presumed that feature importance and counterfactual explanations lead to more trust *and* reliance in participants compared to a control condition where no additional explanation was present. Counterfactuals are both selective and contrastive, while feature importance explanations are just selective (Miller, 2019). This makes counterfactuals an even more promising type of human-centered explanation compared to feature importance explanations.

For these reasons, the specific hypotheses were:

H_1 The experimental condition *feature importance* will lead to higher reliance compared to the control.

H_2: The experimental condition *counterfactuals* will lead to higher reliance compared to the control.

H_3: *Counterfactuals* will lead to higher reliance compared to *feature importance*.

H_4: The experimental condition *feature importance* will lead to higher trust compared to the control.

H_5: The experimental condition *counterfactuals* will lead to higher trust compared to the control.

H_6: *Counterfactuals* will lead to higher trust compared to *feature importance*.

4. Method

4.1. Measures

The independent variable was *condition*, with the two levels feature importance and counterfactuals, as well as a third level without explanations, which served as a control. We used two measures as dependent variables to account for the aforementioned distinction between trust as an attitude and reliance as trust-related behavior. On the one hand, we wanted to determine if people relied on the AI and changed their behavior after being presented with an explanation. This behavior change was captured by the parameter *Weight of Advice* (WOA), which stems from the literature on taking advice (Harvey and Fischer, 1997). WOA has the following notation:

$$\mathbf{WOA} = \frac{T2 - T1}{R - T1} \tag{1}$$

In Equation (1), R is defined as the model's recommendation, $T1$ is the participant's initial estimate of the apartment's price before seeing R, and $T2$ is the participant's final estimate of the apartment's price after seeing R. WOA measures the degree to which people change their behavior and move their initial estimate toward the advice. WOA is equal to 1 if the participant's final prediction matches the AI recommendation and equal to 0.5 if they average their initial prediction with the AI recommendation. A WOA of 0 occurs when a participant ignores the AI recommendation ($T1 = T2$), and a negative WOA signifies that a participant discounted the recommendation completely and moved further away from the recommendation. WOA can be viewed as a percentage of how much people weigh the received advice (i.e., the AI recommendation), and this straightforward interpretation is an advantage of this reliance measurement. While WOA has been used in the past by researchers in XAI as an alternative trust measurement (Logg et al., 2019; Mucha et al., 2021; Poursabzi-Sangdeh et al., 2021), it has never been explicitly referred to as reliance and thus clearly differentiated from trust to the best of our knowledge.

On the other hand, we chose the TPA (Jian et al., 2000) to measure trust because the scale's underlying definition of trust is compatible with the one we adopted from Lee and See (2004). Furthermore, the TPA is an established measure in HCI (Hoffman et al., 2019). Several other scales evaluating AI have adapted items from the TPA (e.g., Hoffman et al., 2019), and its psychometric quality has been evaluated multiple times (Spain et al., 2008; Gutzwiller et al., 2019). Jian et al. (2000) treated trust and distrust as opposite factors extending along a single dimension of trust. The scale is a seven-point Likert-type scale (ranging from 1: "not at all" to 7: "extremely") and consists of 12 items. Five items for distrust (i.e., *"The system is deceptive."*, *"The system behaves in an underhanded manner."*, *"I am suspicious of the system's intent, action or, outputs."*, *"I am wary of the system."*, *"The system's actions will have a harmful or injurious outcome."*). The seven remaining items for trust (i.e., *"I am confident in the system."*, *"The system provides security."*, *"The system has integrity."*, *"The system is dependable."*, *"The system is reliable."*, *"I can trust the system."*, *"I am familiar with the system."*). We used the scale in its original form, except for prefixing the word "AI" to the word "system," e.g., "I have confidence in the AI system."

4.2. Experiment

We carried out a one-factor between-subjects design online experiment[1] over Amazon Mechanical Turk (MTurk, http://mturk.com). The experiment was implemented through the online survey tool Limesurvey (http://limesurvey.org).

4.2.1. Participants

A total of 913 participants were initially recruited over MTurk, and 798 of them fully completed the survey. Only workers from the USA with a human-intelligence-task (HIT) approval of 95% and at least 100 approved HITs were allowed to participate in the experiment. Workers who completed the task conscientiously were reimbursed with 1.50 US dollars and a bonus of 0.30 US dollars for their participation. Several criteria were applied during data cleaning to ensure data quality. Participants who failed to provide a correct answer ($n = 36$) for the bogus item ("This is an attention check. Please choose 7 here") or for one of three control questions ("In this survey, you had to tell us for how much money you would sell a house to a company"; "In this survey, we asked you to indicate in which U.S. state you currently live"; "In this survey, you got price recommendations from a good friend") were removed ($n = 310$). We also excluded participants that showed unrealistic WOA's ($n = 72$). Following prior research (Gino and Moore, 2007; Logg et al., 2019), we defined unrealistic WOA as being ≤ -1 and ≥ 2. For the data analysis, 380 participants remained. The sample was predominantly male (61%) and had an average age of 37 years ($M = 37.03$, $SD = 10.15$, $min = 18$, $max = 69$). A majority of the participants (68%) possessed a higher-educational degree (i.e., a bachelor's degree, master's degree, or PhD).

4.2.2. Procedure and task

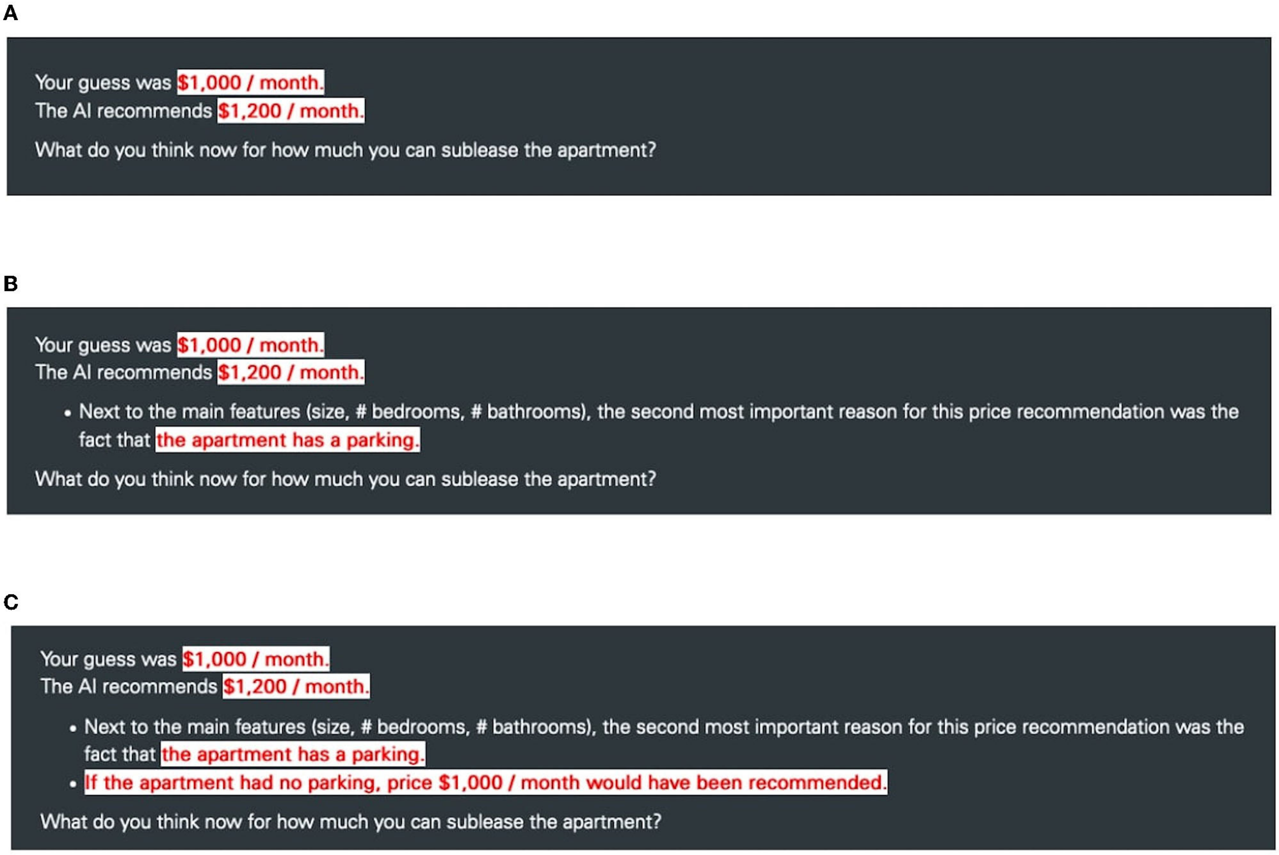
After providing informed consent, participants were introduced to the study and their task. They were asked to imagine a scenario where their goal was to sublease six different apartments on a subleasing website. Based on the apartment's features and amenities (e.g., number of bedrooms, distance to public transit), they had to estimate an initial subleasing price ($T1$). After estimating $T1$, an alleged AI from the website provided a computed price recommendation (R). In reality, the price recommendation was based on an algorithm introduced as an AI. Participants were informed that they would be more likely to find a subleaser by deciding on a lower price but would consequently receive less profit. If they decided on a higher price, they would be less likely to find a subleaser but potentially receive more profit. They were told that the AI's goal was to help them find the optimal price to successfully find a subleaser with a reasonable profit. How exactly this price recommendation was calculated by the algorithm will be discussed in the next section. Figure 1 shows how the price recommendation and respective explanations were presented to the participants. A list of all the explanations used for each type of recommendation can be found in the online repository[1]. The output was designed to appear as if the subsequent explanation was an extension of the preceding ones. The most relevant outputs were presented as a console output in order to make the stimuli more convincing (see Figure 1). After seeing the AI recommendation, participants could decide if they desired to approach it or not and settled on a final subleasing price ($T2$). This deliberate choice by the participants to either rely on the AI recommendation or not makes our experiment an actual decision-making task rather than a proxy task (Buçinca et al., 2020). In proxy-tasks, the focus lies on how well participants can simulate the model's decisions (Buçinca et al., 2020). In actual decision-making tasks, people's choices involve systematic thinking errors (biases) and mental shortcuts (heuristics) (Tversky and Kahneman, 1974), as it is up to the participants to decide whether and how to use the AI (i.e., reliance on the AI).

To evoke a certain degree of uncertainty, participants were told that they would be reimbursed based on their performance. Uncertainty is a defining characteristic of trust (Mayer et al., 1995) and has been referred to as a necessary prerequisite of human trust in AI that has been lacking in current XAI studies (Jacovi et al., 2021). Participants were informed that for good estimations, the top 10% would be paid an additional bonus of 0.30 US dollars. In actuality, every participant received the bonus, regardless of their performance. In order to better control for the price disparity between urban and rural regions, participants were asked to indicate what US state they are currently living in (e.g., Colorado) to ascertain their state capital (e.g., Denver). The objective was to make the estimate easier for the participants and to make the AI more persuasive since it was claimed that the AI would likewise base its price recommendations on data collected in that state capital. After an example that showed how the apartments and their amenities would be presented to them, participants could start with the actual task. Once the task was completed, participants had to fill out the TPA (Jian et al., 2000) and give some demographic information. Participants were debriefed at the end of the study and informed that the AI was an algorithm introduced as an AI that did not use participants' state capitals for its recommendations. To ascertain that our algorithm was convincing, we asked participants before and after the interaction how certain they were about the AI's prediction. ("How certain are you that the AI can make an accurate recommendation for a sublease price?") on a ten-point Likert-type scale (ranging from 1: "not certain at all" to 10: "absolutely certain").

4.2.3. Stimuli

The apartments that participants had to evaluate were real apartments retrieved from the website Zillow (http://zillow.com) in May 2020. To create some variability, we selected six different apartments of different sizes and price ranges: two small-sized apartments (500–750 square feet), two medium-sized apartments (751–1,000 square feet), and two large-sized apartments (1,001–1,250 square feet). Figure 2 shows an example of how apartments were presented to participants. Features and amenities were collected directly from the website Zillow, whenever available. If not available, a random value within a reasonable range was chosen

[1] Data, R-scripts, detailed EFA results, experimental materials (including all questions and tasks), as well as a flowchart of the inclusion and exclusion process of the experiment are available under https://osf.io/bs6q3/.



A

Your guess was $1,000 / month.
The AI recommends $1,200 / month.

What do you think now for how much you can sublease the apartment?

B

Your guess was $1,000 / month.
The AI recommends $1,200 / month.

- Next to the main features (size, # bedrooms, # bathrooms), the second most important reason for this price recommendation was the fact that the apartment has a parking.

What do you think now for how much you can sublease the apartment?

C

Your guess was $1,000 / month.
The AI recommends $1,200 / month.

- Next to the main features (size, # bedrooms, # bathrooms), the second most important reason for this price recommendation was the fact that the apartment has a parking.
- If the apartment had no parking, price $1,000 / month would have been recommended.

What do you think now for how much you can sublease the apartment?

FIGURE 1
Examples how the different explanations for the conditions **(A)** Control, **(B)** Feature Importance, and **(C)** Counterfactual were presented to the participants. For this apartment, the AI recommended increasing the initial price $T1$.

for continuous variables (e.g., distance to public transit between 0.1 and 2.0 miles), and a random choice for dichotomous variables was made (e.g., elevator YES/NO). All participants were presented with the same stimuli, that is, identical apartments and features. What differed was the price recommendation and the explanation that accompanied it (see Figure 1).

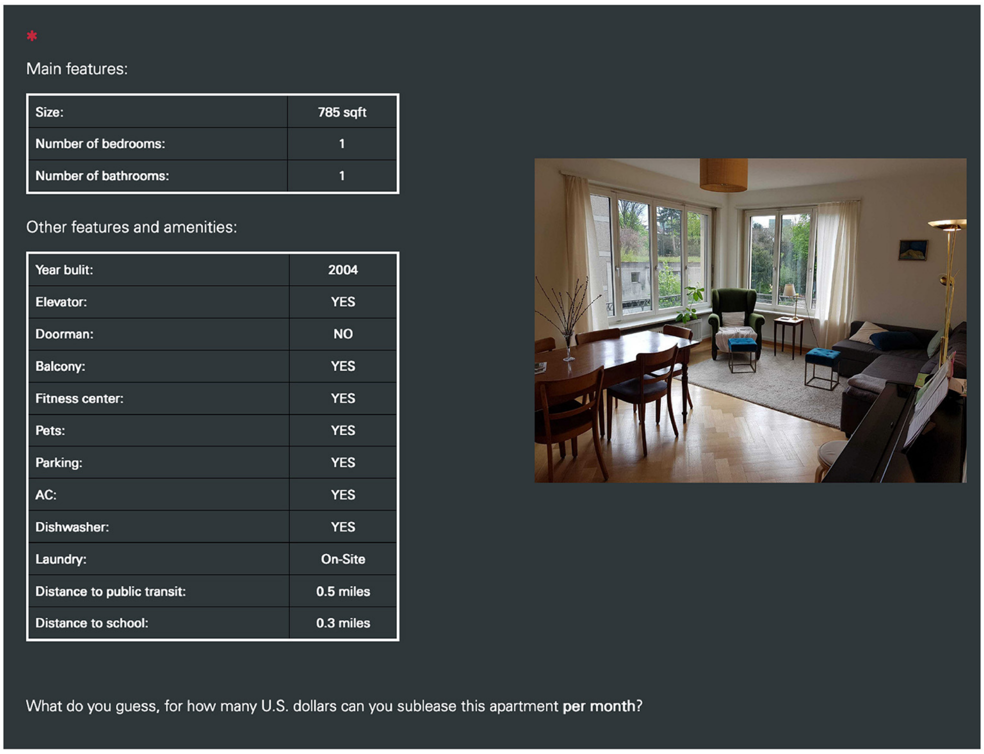
The price recommendation from the algorithm introduced as an AI was designed to pick a random number between 10 and 20. This random number was then transformed into percentages and either added or subtracted to the initial subleasing price ($T1$), which led to a random deviation between 10 and 20 percent. This deviation seemed substantial enough that subjects did not entirely adopt the recommendation, but it was also subtle enough not to appear unrealistic and that it seemed possible that the features and amenities could account for the discrepancy. With this procedure, we ensured that no participant could estimate the price accurately since there was no "true price." Defining a ground truth has been a limitation of past studies (Poursabzi-Sangdeh et al., 2021). If, by pure chance, a participant estimates the "true price," the interpretation of WOA becomes meaningless since $T1$ and R are equal. The absence of a "true price" imposes the decision on participants either to rely or not to rely on the AI. By determining a relative deviation between 10 and 20 percent from participants' initial price estimate, we furthermore controlled for the system's accuracy since it was shown to have a significant effect on people's trust (Yin et al., 2019; Zhang et al., 2020).

It was randomly assigned that for three of the six apartments, the algorithm introduced as an AI recommended decreasing the initial price, $T1$ (e.g., if $T1$ was $1,000 and the random number 17, the AI recommendation was $830), and for the other three apartments, the recommendation was to increase $T1$ (e.g., if $T1$ was $1,000 and the random number 17, the AI recommendation was $1,170). By doing this, the AI informed participants that their initial price estimates were either too low or too high, which made it possible to compare AI recommendations to *increase* $T1$ with recommendations to *decrease* $T1$. We were interested in this comparison because prior research from Kliegr et al. (2021) and Wang et al. (2019) led us to postulate that the AI's recommendation to increase or decrease the initial price might also influence participants' decision-making and consequently their reliance on the AI.

5. Results

5.1. Descriptive statistics

On average, participants across all conditions approached the AI recommendation, resulting in a positive WOA ($M = 0.69$,



*

Main features:

Size:	785 sqft
Number of bedrooms:	1
Number of bathrooms:	1

Other features and amenities:

Year bulit:	2004
Elevator:	YES
Doorman:	NO
Balcony:	YES
Fitness center:	YES
Pets:	YES
Parking:	YES
AC:	YES
Dishwasher:	YES
Laundry:	On-Site
Distance to public transit:	0.5 miles
Distance to school:	0.3 miles

What do you guess, for how many U.S. dollars can you sublease this apartment **per month**?

FIGURE 2
Example of how an apartment was presented to participants in the online experiment. Note that the image of the apartment depicted in the screenshot was replaced with a similar image for this publication due to potential copyright issues.

SD = 0.36). The TPA showed average overall ratings (M = 5.01, SD = 0.86), high ratings for trust (M = 4.98, SD = 1.05), and lower ratings for distrust (M = 2.95, SD = 1.62). Across all three conditions, the certainty that the AI could make an accurate prediction increased from pre- to post-interaction (M_{Δ} = 0.36, SD = 1.53) and was rated at a high level after the interaction (M = 7.59, SD = 1.60). The inspection of the average estimated prices also confirmed our classifications into the apartment categories "small," "medium," and "large" ($M_{small} = \$1,091$, $M_{medium} = \$1,286$, $M_{large} = \$1,449$). From this, we concluded that the assigned task was a compelling one. Table 1 includes descriptive statistics for the experiment.

5.2. Reliance—WOA

To address $\mathbf{H_1}$, $\mathbf{H_2}$, and $\mathbf{H_3}$, corresponding contrasts were created. The first contrast made it possible to determine if the feature importance condition was significantly different from the control (planned contrast 1: feature importance explanation vs. control for answering $\mathbf{H_1}$). By defining two other contrasts, it was possible to examine if the counterfactual condition was significantly different from the control (planned contrast 2: counterfactual explanation vs. control for answering $\mathbf{H_2}$) and if the counterfactual condition was significantly different from the feature importance condition (planned contrast 3: counterfactual explanation vs. feature importance explanation for answering $\mathbf{H_3}$). The effect of the three contrasts on WOA was analyzed by employing linear mixed-effect models (LMEMs) using the *lme4* package (Bates et al., 2015) for R (version 4.2.2.). We report β-estimates, their 95% confidence interval, t-values, and the corresponding p-values. Our models contained two fixed effects: the contrasts and the difference of the recommendation to *increase* or *decrease* $T1$. Under the assumption that the stimuli and conditions had varying random effects for different participants, we introduced a random intercept (id) in the model. The utilized model had the following specifications:

$$WOA \sim 1 + Contrast1 + Recommendation + (1|id)$$

For this model, the first contrast (feature importance explanation vs. control) was not significant [$\beta = 0.01$, 95% CI β[-0.01, 0.03], $t_{(378)} = 0.64$, $p = 0.53$], while the difference between recommendations to increase or decrease $T1$ was highly significant [$\beta = 0.05$, 95% CI $\beta[0.03, 0.07]$, $t_{(1,899)} = 3.62$, $p < 0.001$]. The second contrast (counterfactual explanation vs. control) did not return any significant results [$\beta = 0.02$, 95% CI β[-0.01, 0.04], $t_{(378)} = 1.33$, $p = 0.19$], and neither did the third contrast (counterfactual explanation vs. feature importance explanation) [$\beta = 0.01$, 95% CI β[-0.01, 0.03], $t_{(378)} = 0.63$, $p = 0.53$]. However, in all three models, the recommendation type (increasing a price vs. decreasing a price) returned highly significant results with β-estimates ranging between 95% CI [0.03, 0.07]. Comparing this model to a model without the recommendation term confirmed that its inclusion was justified since it significantly

TABLE 1 Descriptive statistics of the conducted experiment with the mean (*M*), standard deviation (*SD*), and median (*Mdn*) for *WOA*, the TPA, and *AI* certainty.

	Control			Feature importance			Counterfactual		
	(*n* = 130)			(*n* = 143)			(*n* = 107)		
	M	*SD*	*Mdn*	*M*	*SD*	*Mdn*	*M*	*SD*	*Mdn*
Weight of Advice									
Increase price	0.69	0.38	0.69	0.68	0.36	0.69	0.67	0.34	0.67
Decrease price	0.69	0.38	0.69	0.73	0.40	0.79	0.78	0.37	0.79
TPA (Jian et al., 2000)									
Overall	5.01	0.91	4.88	4.95	0.84	4.83	5.10	0.84	5.08
Trust	5.11	1.06	5.29	4.82	1.07	5.00	5.03	1.00	5.14
Distrust	3.12	1.73	2.60	2.88	1.58	2.40	2.82	1.54	2.20
AI certainty									
Pre-interaction	7.24	1.60	8.00	7.20	1.57	8.00	7.27	1.64	8.00
Post-interaction	7.55	1.74	8.00	7.48	1.61	8.00	7.80	1.37	8.00

improved the model fit [$\chi^2_{(1)} = 13.05$, $p < 0.001$]. To better understand the relationship between explanations and the type of recommendation, we created a visualization (see Figure 3).

Depending on the type of recommendation, the condition effect was different, meaning that whether the AI recommended *increasing* or *decreasing* the initial subleasing price $T1$, influenced the way that explanations affected WOA. For recommendations to increase, the explanations had a negligible effect on WOA, but for recommendations to decrease, the effect was substantial. In our case, the effect of explanations cannot be readily understood without considering the different type of AI recommendation. We therefore divided the data into two subsets. One subset contained the three apartments with the *recommendation to increase* $T1$, the other subset contained the three apartments with the *recommendation to decrease* $T1$. We then executed the specified model again, but the term "recommendation" was naturally omitted as a fixed effect. For the subset that contained the recommendations to decrease $T1$, the second contrast (counterfactual explanation vs. control) was significant [$\beta = 0.04$, 95% $CI\ \beta[0.01, 0.08]$, $t_{(378)} = 2.31$, $p = 0.02$]. The β-estimates indicate that on average, counterfactual explanations increased WOA by an approximated 4% compared to the control that received no explanations. Note that feature importance explanations likewise increased WOA by 2% compared to the control, but this difference was not significant for the 0.05 significance level [$\beta = 0.02$, 95% $CI\ \beta[-0.01, 0.05]$, $t_{(378)} = 1.10$, $p = 0.27$]. However, explanations had no effect on WOA when the AI recommended increasing the price estimate (Figure 3). LMEMs are quite robust against violations of distributional assumptions (Schielzeth et al., 2020). We nevertheless checked the residuals of WOA values for normal distribution via quantile–quantile plots (Q–Q plots) to determine if the residual variance was equal across conditions (homoscedasticity) and also checked the multicollinearity assumption. The normality distribution seemed to be satisfied, with some deviation from normality at the tails, which indicates that more data is located at the extremes. Levene's test indicated equal variances [$F_{(2)} = 0.57$, $p = 0.56$] that did not differ between the conditions, and a multicollinearity check revealed low correlation between the model terms.

5.3. Trust—Trust between people and automation scale

To identify the underlying structure of the TPA, we performed an exploratory factor analysis (EFA) using MinRes and rotated with the Oblimin method (see footnote 1). Parallel analysis and very simple structure (VSS) indicated two factors, which is in line with previous research (Spain et al., 2008). The first five items loaded on one factor (with 0.79 — 0.89), and the other seven loaded on a second factor (0.56 − 0.85), which corresponded accurately to the trust/distrust items of the scale. Internal consistency for the first five items (i.e., distrust) was excellent ($\alpha = 0.92$, 95% CI[0.90, 0.93], $\omega = 0.92$, 95% CI[0.90, 0.93]) and good for the seven trust items ($\alpha = 0.88$, 95% CI[0.86, 0.90], $\omega = 0.88$, 95% CI[0.86, 0.90]) according to George and Mallery (2019). To test $\mathbf{H_4}$ (feature importance leads to higher trust compared to the control), $\mathbf{H_5}$ (counterfactuals lead to higher trust compared to the control) and $\mathbf{H_6}$ (counterfactuals lead to higher trust compared to feature importance), we intended to perform two types of one-way analyses of variance (ANOVAs), once using the overall mean score, and once using mean scores for the trust and distrust factors. However, visual inspection of the distribution and a Shapiro–Wilk test ($W = 0.97$, $p < 0.001$) revealed a non-trivial violation of the normality assumption. Thus, the ANOVA results might not have been interpretable and meaningful. Under these circumstances, a non-parametric Kruskal–Wallis test (Kruskal and Wallis, 1952) was carried out as it does not assume a normal distribution of the residuals. The results of the Kruskal–Wallis test showed that the overall mean ratings for the TPA were not significantly different between the conditions [$H_{(2)} = 1.54$, $p = 0.46$]. The same was

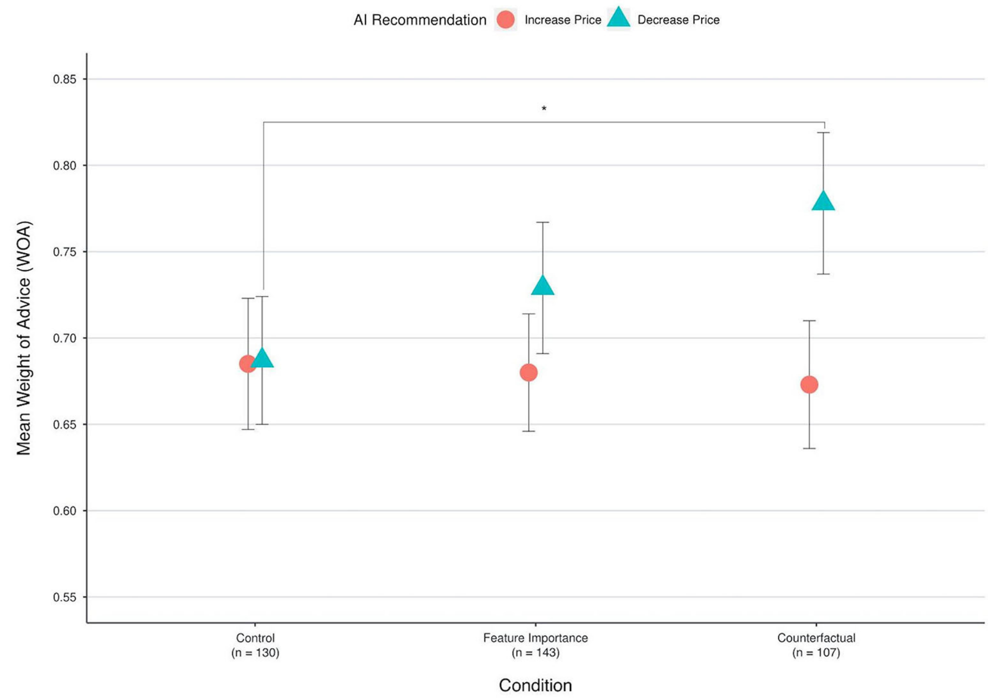


FIGURE 3
Mean weight of advice for each condition and type of recommendation. Note that the y-axis is scaled to better visualize the effect. The error bars depict 95% confidence intervals. *Statistically significant difference with $p < 0.05$.

true for the factors trust [$H_{(2)} = 5.06$, $p = 0.08$] and distrust [$H_{(2)} = 2.03$, $p = 0.36$]. Since the omnibus Kruksal–Wallis was not significant, we did not perform further *post-hoc* tests. Figure 4 captures the similar trust ratings for the two experimental conditions and the control.

6. Discussion

The experiment reported in this study demonstrates that participants generally rely on AI recommendations in low-stake decision-making tasks—in this case, receiving AI recommendations to find an optimal price for subleasing an apartment. Regardless of the different experimental manipulations, on average, participants displayed high overall Weight of Advice scores ($M = 0.67 - 0.69$, $SD = 0.25 - 0.36$). A WOA of 0.70 signifies that participants adopted 70% of the AI recommendations when updating their prior beliefs to form their final estimate. This finding supports the idea that people generally rely on AI (Logg et al., 2019).

The results further demonstrated that under certain conditions, explainability significantly increases AI reliance. However, in the context of our study, the effect of human-centered explanations depended on the type of decision-making the participants had to engage in. We presented participants with two kinds of recommendations: for the first type, an algorithm introduced as an AI recommended that participants *increase* their initially estimated apartment price. For the second type, participants were advised to *decrease* their initial price. The results of the experiment indicate that when the AI recommended increasing the price, human-centered explanations did not affect reliance. By contrast, in the case of recommendations to decrease the price, providing counterfactual explanations affected WOA significantly. Participants in the counterfactual condition where the AI recommended decreasing $T1$ relied up to 9 percentage points more on the recommendation than participants in the control that received the recommendation to increase their price. Therefore, the findings support the second hypothesis (**H_2**: counterfactuals will lead to higher reliance compared to the control) only for decision-making tasks where the AI recommended decreasing the price. The first hypothesis (**H_1**: feature importance will lead to higher reliance compared to the control) and the third hypothesis (**H_3**: counterfactuals will lead to higher reliance compared to feature importance) are not supported for either type of recommendation. We conclude that counterfactual explanations can significantly increase reliance but only under certain conditions. However, there was no significant difference between the two *post-hoc* explanations, although counterfactuals are arguably more human-centered since they additionally fulfill Miller (2019)'s *contrastivity* criterion.

The experiment illustrated that the decision-making task with regards to increasing or decreasing a price had a significant effect on reliance. Regardless of providing explanations, participants consistently relied more on AI recommendations to decrease prices than recommendations to increase them (see Figure 3). This seems counterintuitive at first glance since one might expect that participants would always embrace the prospect of obtaining a higher subleasing price. We argue that the two

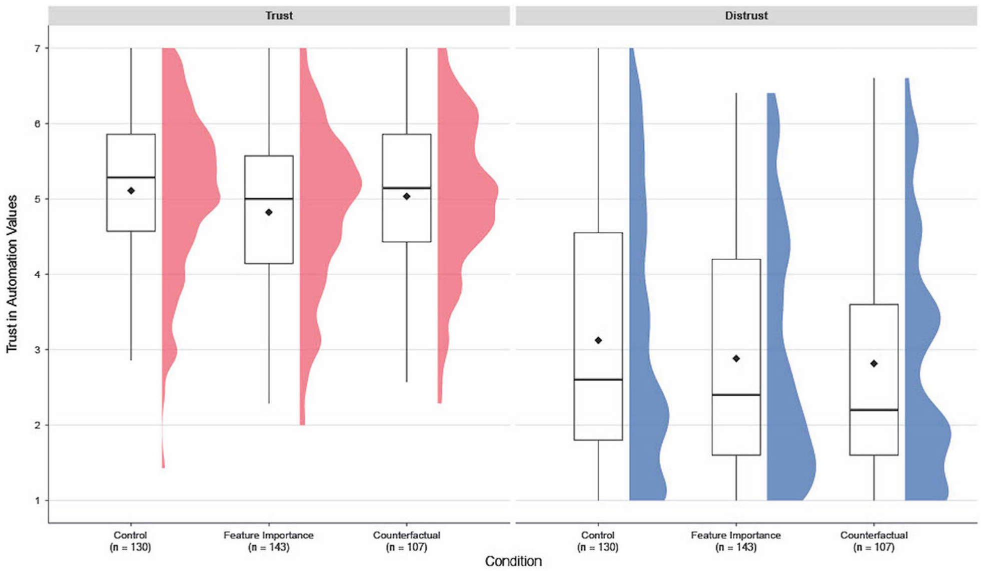


FIGURE 4
Boxplots and density plots for the Trust in Automation factors trust and distrust for each condition. Black rhombuses indicate mean values. Items 1–5 of the TPA code for distrust, items 6–12 code for trust.

types of recommendations should be thought of as two distinct decision-making tasks and our results demonstrate how cognitive biases may affect humans in their decision-making involving AI as proposed by Kliegr et al. (2021). The well-studied concept of *loss aversion* by Tversky and Kahneman (1991) could account for this discrepancy and serve as an explanation attempt for our findings. Loss aversion suggests that, psychologically, people assign more utility to losses than to gains (Tversky and Kahneman, 1991). In practical terms, this means that the dissatisfaction experienced by a person who loses $100 is greater than the satisfaction experienced by a person who gains $100. Our study design seems to satisfy the preconditions for a possible loss-aversion effect: when participants received a recommendation to increase their initial price estimate, they were likely concerned that this potential price increase would cause an unsuccessful sublease. The prospect of getting more money (gain) mattered less in this decision-making task than the possibility of not being able to sublease at all (loss). A recommendation to decrease the initial price may not have induced loss aversion in participants. When not being confronted with loss aversion, explanations seemed to convince participants that demanding less money was the right decision to successfully sublease the apartment, compared to the control where no additional explanation was present for the recommendation. This interpretation suggests that human-centered AI explanations can have an effect on reliance, but only for decision-making tasks where other contributing factors such as loss aversion are absent. Substantial work has been published about biased AI-training data but little about humans' cognitive biases and heuristics when exposed to AI. A notable exception is the work of Lu and Yin (2021) that showed how people use heuristics and base their reliance on the level of agreement between the machine learning model and themselves when performance feedback was limited. Moreover, Wang et al. (2019) proposed a framework of how human reasoning should inform XAI to mitigate possible cognitive biases, and a recent review by Kliegr et al. (2021) explored to what extent biases affect human understandings of interpretable machine-learning models. We present empirical findings suggesting that the XAI community should account for possible biases and heuristics to develop genuinely human-centered explanations. Inherent biases and heuristics may be so hardwired in people that AI explanations are not convincing enough to disprove non-optimal human decision-making. If that is the case, AI may not help users to reach better decisions in circumstances where human intuition becomes too tempting for their judgment. While the interpretation of the present results under the perspective of loss aversion requires further investigation, our findings highlight the importance of biases and heuristic end-users can exhibit in actual decision-making tasks rather than proxy tasks. These biases and heuristics may result in irrational and non-optimal choices, which in turn affect the measured variables of interest, including trust and trust-related behavior (Wang et al., 2019; Kliegr et al., 2021). In the context of our study, the type of decision-making participants faced had a greater effect on reliance than the explanation provided by the AI. This suggests that factors other than explainability are crucial when designing human-centered AI. Cognitive biases and heuristics, such as loss aversion (Tversky and Kahneman, 1991), framing (Tversky and Kahneman, 1981), or confirmation bias (Wason, 1960), could potentially undermine AI explanations.

Concerning trust, no significant differences were found for human-centered explanations (Figure 4). Therefore, we reject the fourth hypothesis (**H$_4$**: feature importance will lead to higher trust compared to the control), the fifth hypothesis (**H$_5$**: counterfactuals will lead to higher trust compared to the control), and the sixth hypothesis (**H$_6$**: counterfactuals will lead to higher trust compared to feature importance). While the effect of human-centered explanations on reliance depended on the nature of the AI recommendation (increasing or decreasing the initial estimated apartment price), this dependence and the potential effect of cognitive biases and heuristics remain to be explored for trust. We operationalized trust as an attitude toward the AI system and consequently assessed users' trust after the entire task while we observed reliance from trial-to-trial. Given the present study, our results do not indicate a consistent effect of human-centered explanations on trust. Thus, our findings are in line with other research that provided mixed evidence regarding the effect of transparency on trust (Cramer et al., 2008; Nothdurft et al., 2013; Cheng et al., 2019; Ehsan et al., 2019; Zhang et al., 2020; Poursabzi-Sangdeh et al., 2021). However, the conceptual distinction between trust and reliance carries significant implications for XAI evaluation and uncovers two potential challenges. First, if researchers only assess attitudinal trust via questionnaires, they could falsely assume that people will not rely on an AI system. Second, if only trust-related behavior (i.e., reliance) is measured, researchers might incorrectly deduce that people necessarily trust the system in question. Consequently, researchers and practitioners have to answer the challenging question of which of the two measures to account for and investigate when evaluating AI. Whether trust translates into reliance is more nuanced than often assumed and depends on an interaction between the operator, the automation, and the situation (Körber, 2018). It remains to be determined which factors other than trust drive AI reliance (e.g., system accuracy, perceived usefulness, cognitive biases) or whether current measurement tools originally designed to assess trust in automation also accurately capture AI trust.

Furthermore, we initially chose to measure overall trust, expecting a single-factor structure as proposed by Jian et al. (2000). However, investigating the scale's factor structure through exploratory factor analysis before interpreting the results implied a two-factor structure, as previously observed outside the AI context (Spain et al., 2008). This change in the theoretical structure led us to use the TPA to measure trust and distrust as two distinct factors in the analysis. While in our study, levels of trust and distrust behaved as expected across the three conditions (i.e., high trust and low distrust), it is plausible to assume that there are situations where the difference between people's trust and distrust in an AI is more nuanced. Attitudes are often seen as lying along one continuum, as was initially proposed for the TPA, but past research has argued that positive and negative attitudes can co-occur (Priester and Petty, 1996). For example, when smokers try to quit smoking, they can have a simultaneously negative and positive view toward cigarettes (Cacioppo and Berntson, 1994). Thus, it may be necessary to distinguish between trust and distrust in studies that aim to investigate ambivalent attitudes toward AI. This distinction can be accounted for when using the two-factor structure for the TPA. On the other hand, a single-factor structure comes with the risk of oversimplifying situations and losing important nuance when using the TPA to measure trust in AI. Overall, our findings emphasize that researchers must carefully differentiate attitude from behavior and choose appropriate evaluation metrics for human-centered AI accordingly.

7. Limitations and future work

We conducted an online decision-making experiment in a domain-specific task. Future work should broaden the scope and focus on domains other than real estate to investigate if the findings of this study are transferable to different scenarios and AI systems used in practice to increase external validity. While decision-making experiments on MTurk allow high control over confounding variables and are comparable to those in laboratory settings, even in low-stakes scenarios (Amir et al., 2012), future studies could focus on high-stakes decisions to evoke uncertainty where a more tangible loss depends on the participants' decision to trust AI.

Participants were presented with AI recommendations that were expected to seem reasonably trustworthy since the recommendations were formed based on the participants' initial estimates. However, explainability might have a greater impact on trust and reliance if the recommendations were not credible or showed greater deviations from the initial estimate. Past research has shown that people calibrate their trust based on the system's capabilities (Lee and See, 2004; Zhang et al., 2020), and often fail to rely on algorithms after learning that they are imperfect, a phenomenon called algorithm aversion (Dietvorst et al., 2015). By providing explanations, people may better understand AI errors and factors that influence those errors (Dzindolet et al., 2003). Future research could investigate more untrustworthy recommendations by gradually reducing the capabilities of the system (e.g., by decreasing the system's accuracy) and examining how explainability affects reliance and trust in cases where the AI objectively performs poorly or when there is a clear disagreement between the end-user and the AI.

The study design did not allow for a clear distinction between *dispositional trust*, *situational trust*, and *learned trust*, as suggested by Hoff and Bashir (2015). In addition, by measuring trust as an attitude toward the AI system after task completion, examining the effects of bias and heuristics at the level of individual trials was not possible. Trust is a dynamic process in the human–AI interaction, and we expect that trust changes as time passes in the interaction between end-users and AI-based systems. We recommend that future studies investigate the varying manifestations of trust because they are critical for a comprehensive understanding of the human–AI interaction. Researchers could measure AI trust before and after participants are exposed to an AI system and compare the reported trust scores (learned trust). Alternatively, they could expose participants to AI recommendations while inducing different emotional valences (situational trust). Future research could also investigate the relationship between AI trust and reliance from the perspective of the technology acceptance model (Davis, 1989), which can be seen as a further development of Ajzen and Fishbein (1980)'s work.

8. Conclusion

We conducted an empirical experiment demonstrating that human-centered explanations as a means for transparent AI increase reliance for specific decision-making tasks. While this provides some evidence that human-centered *post-hoc* explanations can be an opportunity for more transparent AI, our findings emphasize that the effect of transparency on reliance and trust is more nuanced than commonly assumed.

The type of decision-making task (increasing vs. decreasing a price) had a greater influence on end-users' reliance than how the AI explained its decision did. We argue that humans may exhibit cognitive biases and apply heuristics to decision-making tasks that involve AI. So far, the discussion around bias has focused primarily on biased data and prejudice due to incorrect assumptions in the machine-learning process. The implications of potential biases and heuristics when humans are presented with explanations from AI have received little attention in the current XAI debate. Both researchers and practitioners need to be aware of such dynamics in the design for truly human-centered AI, as poor partnership between people and automation will become increasingly costly and consequential (Lee and See, 2004).

In order to draw valid conclusions from experiments, XAI researchers need to be cautious when measuring the human side of the human–AI interaction. Conceptualizing trust as an attitude and reliance as a trust-related behavior might lead to divergent results. Our study also confirmed a two-factor structure (trust and distrust) for the TPA, as previously reported outside the AI context. Given the importance of AI, as it is increasingly used to make critical decisions with far-reaching implications, meaningful evaluations in XAI research require agreed-upon metrics and appropriate measurements that have been empirically validated.

Data availability statement

The datasets presented in this study can be found in online repositories. The names of the repository/repositories and accession number(s) can be found here: https://osf.io/bs6q3/.

Ethics statement

Ethical review and approval was not required for the study on human participants in accordance with the local legislation and institutional requirements. The patients/participants provided their written informed consent to participate in this study.

Author contributions

NS and FB contributed to the conception and design of the study and implemented the online study. NS collected the data and wrote the first draft. NS, FB, and SP performed the statistical analysis. All authors contributed to the manuscript revision, read, and approved the submitted version.

Funding

This research was financed entirely by our research group and the publication fund of the University of Basel for open access; we received no additional funding.

Acknowledgments

Special thanks to Zgjim Memeti and Léane Wettstein.

Conflict of interest

The authors declare that the research was conducted in the absence of any commercial or financial relationships that could be construed as a potential conflict of interest.

Publisher's note

All claims expressed in this article are solely those of the authors and do not necessarily represent those of their affiliated organizations, or those of the publisher, the editors and the reviewers. Any product that may be evaluated in this article, or claim that may be made by its manufacturer, is not guaranteed or endorsed by the publisher.

References

Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y., and Kankanhalli, M. (2018). "Trends and trajectories for explainable, accountable and intelligible systems: an HCI research agenda," in *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI '18* (New York, NY: ACM), 1–18.

ACM FAccT Conference (2022). "ACM conference on fairness, accountability, and transparency 2022 (ACM FAccT 2022) call for papers," in *ACM Conference on Fairness, Accountability, and Transparency 2022* (New York, NY).

Adadi, A., and Berrada, M. (2018). Peeking inside the black-box: a survey on explainable artificial intelligence (XAI). *IEEE Access* 6, 52138–52160. doi: 10.1109/ACCESS.2018.2870052

Ajzen, I., and Fishbein, M. (1980). *Understanding Attitudes and Predicting Social Behavior*. Englewood Cliffs, NJ: Prentice-Hall.

Amir, O., Rand, D. G., and Gal, Y. K. (2012). Economic games on the internet: the effect of $1 stakes. *PLoS ONE* 7, e31461. doi: 10.1371/journal.pone.0031461

Arrieta, A. B., Diaz-Rodriguez, N., Ser, J. D., Bennetot, A., Tabik, S., Barbado, A., et al. (2020). Explainable artificial intelligence (XAI): concepts, taxonomies, opportunities and challenges toward responsible AI. *Inform. Fusion* 58, 82–115. doi: 10.1016/j.inffus.2019.12.012

Bates, D., Mochler, M., Bolker, B., and Walker, S. (2015). Fitting linear mixed-effects models using lme4. *J. Stat. Softw.* 67, 1–48. doi: 10.18637/jss.v067.i01

Biran, O., and Cotton, C. (2017). "Explanation and justification in machine learning: a survey," in *IJCAI-17 Workshop on Explainable AI (XAI)* (Melbourne), 8–13.

Buçinca, Z., Lin, P., Gajos, K. Z., and Glassman, E. L. (2020). "Proxy tasks and subjective measures can be misleading in evaluating explainable AI systems," in *Proceedings of the 25th International Conference on Intelligent User Interfaces, IUI '20* (New York, NY: ACM), 454–464.

Cacioppo, J. T., and Berntson, G. G. (1994). Relationship between attitudes and evaluative space: a critical review, with emphasis on the separability of positive and negative substrates. *Psychol. Bull.* 115, 401–423.

Cheng, H.-F., Wang, R., Zhang, Z., O'Connell, F., Gray, T., Harper, F. M., et al. (2019). "Explaining decision-making algorithms through UI: strategies to help non-expert stakeholders," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, CHI '19* (New York, NY: ACM), 1–12.

Chopra, K., and Wallace, W. A. (2003). "Trust in electronic environments," in *Proceedings of the 36th Annual Hawaii International Conference on System Sciences, HICSS '03* (Los Alamitos, CA: IEEE Computer Society Press, IEEE Computer Society Press), 10–15.

Cramer, H., Evers, V., Ramlal, S., Someren, M., Rutledge, L., Stash, N., et al. (2008). The effects of transparency on trust in and acceptance of a content-based art recommender. *User Model. User Adapt. Interact.* 18, 455–496. doi: 10.1007/s11257-008-9051-3

Davis, F. D. (1989). Perceived usefulness, perceived ease of use, and user acceptance of information technology. *MIS Q.* 13, 319–340.

De Graaf, M. M. A., and Malle, B. F. (2018). "People's judgments of human and robot behaviors: a robust set of behaviors and some discrepancies," in *Companion of the 2018 ACM/IEEE International Conference on Human-Robot Interaction, HRI '18* (New York, NY: ACM), 97–98.

Dietvorst, B. J., Simmons, J. P., and Massey, C. (2015). Algorithm aversion: people erroneously avoid algorithms after seeing them err. *J. Exp. Psychol. Gen.* 144, 114–126. doi: 10.1037/xge0000033

Du, M., Liu, N., and Hu, X. (2019). Techniques for interpretable machine learning. *Commun. ACM* 63, 68–77. doi: 10.1145/3359786

Dzindolet, M. T., Peterson, S. A., Pomranky, R. A., Pierce, L. G., and Beck, H. P. (2003). The role of trust in automation reliance. *Int. J. Hum. Comput. Stud.* 58, 697–718. doi: 10.1016/S1071-5819(03)00038-7

Ehsan, U., Liao, Q. V., Muller, M., Riedl, M. O., and Weisz, J. D. (2021a). "Expanding explainability: towards social transparency in AI systems," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (New York, NY: ACM), 1–19.

Ehsan, U., and Riedl, M. O. (2020). "Human-centered explainable AI: towards a reflective sociotechnical approach," in *HCI International 2020-Late Breaking Papers: Multimodality and Intelligence: 22nd HCI International Conference, HCII 2020* (Copenhagen), 449–466.

Ehsan, U., Tambwekar, P., Chan, L., Harrison, B., and Riedl, M. O. (2019). "Automated rationale generation: a technique for explainable ai and its effects on human perceptions," in *Proceedings of the 24th International Conference on Intelligent User Interfaces, IUI '19* (New York, NY: ACM), 263–274.

Ehsan, U., Wintersberger, P., Liao, Q. V., Mara, M., Streit, M., Wachter, S., et al. (2021b). "Operationalizing human-centered perspectives in explainable AI," in *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (New York, NY: Association for Computing Machinery), 1–6.

European Parliament and Council of the European Union (2016). *Regulation (EU) 2016/679 of the European Parliament and of the Council of 27 April 2016 on the Protection of Natural Persons With Regard to the Processing of Personal Data and on the Free Movement of Such Data, and Repealing Directive 95/46/EC (General Data Protection Regulation).*

Ferreira, J. J., and Monteiro, M. S. (2020). "What are people doing about XAI user experience? A survey on ai explainability research and practice," in *Design, User Experience, and Usability. Design for Contemporary Interactive Environments*, eds A. Marcus and E. Rosenzweig (Cham: Springer International Publishing), 56–73.

George, D., and Mallery, P. (2019). *IBM SPSS Statistics 26 Step by Step: A Simple Guide and Reference, 16th Edn.* New York, NY: Routledge.

Gino, F., and Moore, D. A. (2007). Effects of task difficulty on use of advice. *J. Behav. Decis. Mak.* 20, 21–35. doi: 10.1002/bdm.539

Gutzwiller, R. S., Chiou, E. K., Craig, S. D., Lewis, C. M., Lematta, G. J., and Hsiung, C.-P. (2019). "Positive bias in the "trust in automated systems survey"? An examination of the Jian et al. (2000) scale," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 217–221.

Hartmann, M. (2020). *Vertrauen - Die unsichtbare Macht.* Berlin: Fischer Verlag.

Harvey, N., and Fischer, I. (1997). Taking advice: accepting help, improving judgment, and sharing responsibility. *Organ. Behav. Hum. Decis. Process.* 70, 117–133.

Hilton, D. J. (1990). Conversational processes and causal explanation. *Psychol. Bull.* 107, 65–81.

Hoff, K. A., and Bashir, M. (2015). Trust in automation: integrating empirical evidence on factors that influence trust. *Hum. Fact.* 57, 407–434. doi: 10.1177/0018720814547570

Hoffman, R. R., Mueller, S. T., Klein, G., and Litman, J. (2019). Metrics for explainable AI: challenges and prospects. *arXiv preprint arxiv: 1812.04608.* doi: 10.48550/arXiv.1812.04608

Hong, S. R., Hullman, J., and Bertini, E. (2020). "Human factors in model interpretability: industry practices, challenges, and needs," in *Proceedings of the ACM on Human-Computer Interaction* (New York, NY).

Jacovi, A., Marasović, A., Miller, T., and Goldberg, Y. (2021). "Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI," in *Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency, FAccT '21* (New York, NY: ACM), 624–635.

Jian, J.-Y., Bisantz, A. M., and Drury, C. G. (2000). Foundations for an empirically determined scale of trust in automated systems. *Int. J. Cogn. Ergon.* 4, 53–71. doi: 10.1207/S15327566IJCE0401_04

Kirlik, A. (1993). Modeling strategic behavior in human-automation interaction: why an "aid" can (and should) go unused. *Hum. Fact.* 35, 221–242.

Kizilcec, R. F. (2016). "How much information? Effects of transparency on trust in an algorithmic interface," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (New York, NY: ACM), 2390–2395.

Kliegr, T., Stepan Bahnik, and Furnkranz, J. (2021). A review of possible effects of cognitive biases on interpretation of rule-based machine learning models. *Artif. Intell.* 295, 103458. doi: 10.1016/j.artint.2021.103458

Kocielnik, R., Amershi, S., and Bennett, P. N. (2019). "Will you accept an imperfect AI? Exploring designs for adjusting end-user expectations of AI systems," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (New York, NY: ACM), 1–14.

Körber, M. (2018). "Theoretical considerations and development of a questionnaire to measure trust in automation," in *Proceedings of the 20th Congress of the International Ergonomics Association, IEA '18* (Cham: Springer International Publishing), 13–30.

Krause, J., Perer, A., and Ng, K. (2016). "Interacting with predictions: visual inspection of black-box machine learning models," in *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems* (New York, NY: ACM), 5686–5697.

Kruskal, W. H., and Wallis, W. A. (1952). Use of ranks in one-criterion variance analysis. *J. Am. Stat. Assoc.* 47, 583–621.

Kulesza, T., Burnett, M., Wong, W.-K., and Stumpf, S. (2015). "Principles of explanatory debugging to personalize interactive machine learning," in *Proceedings of the 20th International Conference on Intelligent User Interfaces, IUI '15* (New York, NY: ACM), 126–137.

Lai, V., and Tan, C. (2019). "On human predictions with explanations and predictions of machine learning models: a case study on deception detection," in *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* '19* (New York, NY: ACM), 29–38.

Langer, M., Oster, D., Speith, T., Kästner, L., Hermanns, H., Schmidt, E., et al. (2021). What do we want from explainable artificial intelligence (XAI)? A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. *Artif. Intell.* 296, 103473. doi: 10.1016/j.artint.2021.103473

LeCun, Y., Bengio, Y., and Hinton, G. (2015). Deep learning. *Nature* 521, 436–444. doi: 10.1038/nature14539

Lee, J. D., and See, K. A. (2004). Trust in automation: designing for appropriate reliance. *Hum. Fact.* 46, 50–80. doi: 10.1518/hfes.46.1.50.30392

Liao, Q. V., Gruen, D., and Miller, S. (2020). "Questioning the AI: informing design practices for explainable ai user experiences," in *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems, CHI '20* (New York, NY: ACM), 1–15.

Lipton, Z. C. (2018a). The mythos of model interpretability. *Commun. ACM* 61, 36–43. doi: 10.1145/3233231

Lipton, Z. C. (2018b). The mythos of model interpretability: in machine learning, the concept of interpretability is both important and slippery. *Queue* 16, 31–57. doi: 10.1145/3236386.3241340

Logg, J. M., Minson, J. A., and Moore, D. A. (2019). Algorithm appreciation: people prefer algorithmic to human judgment. *Organ. Behav. Hum. Decis. Process.* 151, 90–103. doi: 10.1016/j.obhdp.2018.12.005

Lu, Z., and Yin, M. (2021). "Human reliance on machine learning models when performance feedback is limited: heuristics and risks," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (New York, NY: ACM), 1–16.

Mayer, R. C., Davis, J. H., and Schoorman, F. D. (1995). An integrative model of organizational trust. *Acad. Manage. Rev.* 20, 709–734.

Merritt, S. M. (2011). Affective processes in human–automation interactions. *Hum. Fact.* 53, 356–370. doi: 10.1177/0018720811411912

Miller, D., Johns, M., Mok, B., Gowda, N., Sirkin, D., Lee, K., et al. (2016). "Behavioral measurement of trust in automation: the trust fall," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Washington, DC), 1849–1853.

Miller, T. (2019). Explanation in artificial intelligence: insights from the social sciences. *Artif. Intell.* 267, 1–38. doi: 10.1016/j.artint.2018.07.007

Mittelstadt, B., Russell, C., and Wachter, S. (2019). "Explaining explanations in AI," in *Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* '19* (New York, NY: ACM), 279–288.

Mohseni, S., Zarei, N., and Ragan, E. D. (2020). A multidisciplinary survey and framework for design and evaluation of explainable AI systems. *arXiv preprint arxiv:1811.11839.*

Mothilal, R. K., Sharma, A., and Tan, C. (2020). "Explaining machine learning classifiers through diverse counterfactual explanations," in *Proceedings of the 2020*

Conference on Fairness, Accountability, and Transparency, FAT '20* (New York, NY: ACM), 607–617.

Mucha, H., Robert, S., Breitschwerdt, R., and Fellmann, M. (2021). "Interfaces for explanations in human-AI interaction: proposing a design evaluation approach," in *Extended Abstracts of the 2021 CHI Conference on Human Factors in Computing Systems* (New York, NY: ACM), 1 –6.

Nothdurft, F., Heinroth, T., and Minker, W. (2013). "The impact of explanation dialogues on human-computer trust," in *Proceedings, Part III, of the 15th International Conference on Human-Computer Interaction. Users and Contexts of Use* (Berlin; Heidelberg: Springer-Verlag), 59–67.

Papenmeier, A., Kern, D., Englebienne, G., and Seifert, C. (2022). It's complicated: the relationship between user trust, model accuracy and explanations in AI. *ACM Trans. Comput. Hum. Interact.* 29, 1–33. doi: 10.1145/3495013

Parasuraman, R., and Manzey, D. H. (2010). Complacency and bias in human use of automation: an attentional integration. *Hum. Fact.* 52, 381–410. doi: 10.1177/0018720810376055

Parasuraman, R., and Riley, V. (1997). Humans and automation: use, misuse, disuse, abuse. *Hum. Fact.* 39, 230–253.

Poursabzi-Sangdeh, F., Goldstein, D. G., Hofman, J. M., Wortman Vaughan, J. W., and Wallach, H. (2021). "Manipulating and measuring model interpretability," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, CHI '21* (New York, NY: ACM), 1–52.

Priester, J. R., and Petty, R. E. (1996). The gradual threshold model of ambivalence: relating the positive and negative bases of attitudes to subjective ambivalence. *J. Pers. Soc. Psychol.* 71, 431.

Rousseau, D. M., Sitkin, S. B., Burt, R. S., and Camerer, C. (1998). Not so different after all: a cross-discipline view of trust. *Acad. Manage. Rev.* 23, 393–404.

Samek, W., Montavon, G., Vedaldi, A., Hansen, L. K., and Müller, K.-R. (2019). *Explainable AI: Interpreting, Explaining and Visualizing Deep Learning.* Berlin: Springer.

Sanneman, L., and Shah, J. A. (2022). The situation awareness framework for explainable AI (safe-AI) and human factors considerations for XAI systems. *Int. J. Hum. Comput. Interact.* 38, 1772–1788. doi: 10.1080/10447318.2022.2081282

Scharowski, N., Perrig, S. A., von Felten, N., and Brühlmann, F. (2022). "Trust and reliance in XAI-distinguishing between attitudinal and behavioral measures," in *CHI TRAIT Workshop* (New York, NY).

Schielzeth, H., Dingemanse, N. J., Nakagawa, S., Westneat, D. F., Allegue, H., Teplitsky, C., et al. (2020). Robustness of linear mixed-effects models to violations of distributional assumptions. *Methods Ecol. Evol.* 11, 1141–1152. doi: 10.1111/2041-210X.13434

Shmueli, G. (2010). To explain or to predict? *Stat. Sci.* 25, 289–310. doi: 10.1214/10-STS330

Solso, R. L., MacLin, M. K., and MacLin, O. H. (2005). *Cognitive Psychology.* Auckland: Pearson Education.

Spain, R. D., Bustamante, E. A., and Bliss, J. P. (2008). "Towards an empirically developed scale for system trust: take two," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting* (Los Angeles, CA: Sage Publications), 1335–1339.

Stephanidis, C., Salvendy, G., Antona, M., Chen, J. Y. C., Dong, J., Duffy, V. G., et al. (2019). Seven HCI grand challenges. *Int. J. Hum. Comput. Interact.* 35, 1229–1269. doi: 10.1080/10447318.2019.1619259

Suresh, H., Gomez, S. R., Nam, K. K., and Satyanarayan, A. (2021). "Beyond expertise and roles: a framework to characterize the stakeholders of interpretable machine learning and their needs," in *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (New York, NY: ACM), 1–16.

Szymanski, M., Millecamp, M., and Verbert, K. (2021). "Visual, textual or hybrid: the effect of user expertise on different explanations," in *26th International Conference on Intelligent User Interfaces, IUI '21* (New York, NY: ACM), 109–119.

Tversky, A., and Kahneman, D. (1974). Judgment under uncertainty: heuristics and biases. *Science* 185, 1124–1131.

Tversky, A., and Kahneman, D. (1981). The framing of decisions and the psychology of choice. *Science* 211, 453–458.

Tversky, A., and Kahneman, D. (1991). Loss aversion in riskless choice: a reference-dependent model. *Q. J. Econ.* 106, 1039–1061.

Vereschak, O., Bailly, G., and Caramiaux, B. (2021). "How to evaluate trust in AI-assisted decision making? A survey of empirical methodologies," in *Proceedings of the ACM on Human-Computer Interaction* (New York, NY: ACM), 1–39.

Wachter, S., Mittelstadt, B., and Russell, C. (2018). Counterfactual explanations without opening the black box: automated decisions and the GDPR. *Harvard J. Law Technol.* 31, 841–887. doi: 10.2139/ssrn.3063289

Wang, D., Yang, Q., Abdul, A., and Lim, B. Y. (2019). "Designing theory-driven user-centric explainable AI," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (New York, NY: ACM), 1–15.

Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Q. J. Exp. Psychol.* 12, 129–140.

Wischnewski, M., Krämer, N., and Müller, E. (2023). "Measuring and understanding trust calibrations for automated systems: a survey of the state-of-the-art and future directions," in *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI '23* (New York, NY: Association for Computing Machinery).

Yin, M., Wortman Vaughan, J., and Wallach, H. (2019). "Understanding the effect of accuracy on trust in machine learning models," in *Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems* (New York, NY: ACM), 1–12.

Yu, K., Berkovsky, S., Taib, R., Conway, D., Zhou, J., and Chen, F. (2017). "User trust dynamics: an investigation driven by differences in system performance," in *Proceedings of the 22nd International Conference on Intelligent User Interfaces, IUI '17* (New York, NY: ACM), 307–317.

Zhang, Y., Liao, Q. V., and Bellamy, R. K. E. (2020). "Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making," in *Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency, FAT* '20* (New York, NY: ACM), 295–305.