Marketing Science Institute Working Paper Series 2022

Report No. 22-119

Off the Mark: The Influence of AI-Induced Errors on Consumers

Alexander Mueller, Sabine Kuester and Sergej von Janda

Off the Mark:
The Influence of AI-Induced Errors on Consumers

Alexander Mueller [1], Sabine Kuester [1] and Sergej von Janda [2]

[1] Department of Marketing, University of Mannheim, Germany
[2] Kaiser X Labs GmbH, Munich, Germany

Corresponding author: kuester@bwl.uni-mannheim.de

August 2022

Despite advances in technology, artificial intelligence (AI) still commonly makes errors. The popular press demonstrates examples of AI which are not error-free, and recent academic literature calls for scrutinizing AI's pitfalls. This study explores the consequences of AI-induced errors from a marketing perspective. Specifically, we explore consumer responses to different error types as the literature distinguishes technical errors, resulting from a technical disruption of algorithmic processes, and social errors, representing task outcomes that may be mathematically correct but deemed inappropriate due to a social norm violation. We also investigate the impact of error severity by distinguishing between low and high error severity. This distinction is important because prior research has shown different response patterns depending on error severity. Errors can sometimes even evoke positive reactions as described by the pratfall effect. Based on data gathered in four studies, we find that severe errors, regardless of error type, evoke negative responses from consumers. However, minor social errors lead to significantly fewer negative consumer responses than minor technical errors. Cognitive and affective trust mediate the relationship between error type and consumer responses. Our results also reveal that companies should incorporate explainable AI (XAI) into AI applications to mitigate negative effects on consumer responses to erring AI. This study provides a granular perspective on consumer responses to erroneous AI and highlights the importance of AI's adherence to social norms. Specifically, minor social errors could foster the stigmatization of minorities and suggest the necessity of implementing additional safeguards against social norm violations by AI.

Keywords: artificial intelligence, social norm violation, cognitive trust, affective trust, pratfall effect

Introduction

In March 2016, Microsoft launched a chatbot on Twitter based on artificial intelligence (AI). While chatting with real users, chatbot “Tay” was supposed to learn from conversations with users and gather data for future human-AI interactions. Yet, the tone changed quickly within less than 24 hours from conversations about puppies and kittens to “[…] bush did 9/11 and Hitler would have done a better job than the monkey we have now […]” (Hunt, 2016). Microsoft was forced to take its AI-based chatbot offline the day after it was launched, as Tay was not able to differentiate between appropriate and socially unacceptable posts. Tay’s inappropriate behavior consequently led to its downfall, despite the chatbot’s success (from an algorithmic point of view) in independently conversing with other users. There are other examples of AI-induced errors, such as Amazon’s intelligent recruiting application, which preferred male applicants over female applicants, even with similar qualifications (Weissmann, 2018). Despite progress in technology and AI outperforming humans in many tasks, AI still commonly errs in practice. Not every AI-induced error results in drastic consequences, but examples from the popular press demonstrate that AI is not error-free, and recent literature calls for scrutiny of AI’s pitfalls (Agarwal et al., 2020; de Bruyn et al., 2020). Specifically, a review of the existing literature reveals three research gaps that the present study addresses.

First, current research primarily focuses on the “bright” side of AI applications. AI refers to technologies “that mimic human intelligence in tasks such as learning, planning, and problem-solving through higher level, autonomous knowledge creation” (de Bruyn et al. 2020, p.93). Intelligence describes the ability to self-learn and update knowledge based on previous decisions (Huang & Rust, 2021). Today, AI takes on tasks that were previously thought to be impossible to achieve by machines, such as composing music (Castelo et al., 2019), creating art (Jago, 2019), or writing journalistic articles (Mariani et al., 2022). AI applications are being used in many domains, including marketing. Marketing research has begun to examine how AI can change the consumer journey (Hoyer et al., 2020), how patients can be convinced of medical AI (Longoni et al., 2019), and how AI helps segment target markets (Huang & Rust, 2021).

Interestingly, there is scant empirical work on the downsides of AI from a marketing perspective. Initial research has shown that the application of AI in marketing can be challenging (Kopalle et al., 2022; Srinivasan & Sarial-Abi, 2021). For example, practitioners must be careful when applying AI in the field of customer experience management and pay close attention to long-term trade-offs for short-term benefits (Kozinets & Gretzel, 2021). Applying AI in marketing poses a threat for companies to successively disconnect from their customers (Puntoni et al., 2021). Despite these initial insights, the negative consequences of AI applications in marketing have not been sufficiently researched. We address this gap in the literature and shed further light on the "dark side" of AI (Davenport et al., 2020; de Bruyn et al., 2020).

Second, the literature lacks insights on the consequences of different types of errors in human-AI interactions (Dijkstra, 1999; Renier et al., 2021). One downside of AI is that due to its self-learning capacities, AI is "likely […] to generate unexpected, delayed, and hard-to-quantify consequences" (de Bruyn et al. 2020, p. 97). Such consequences are often erroneous outcomes that consumers must deal with. For example, social errors can occur in AI applications, such as Amazon's intelligent recruiting application preferring male over female applicants even with similar qualifications. Furthermore, technical errors can occur, for example, when Apple's voice assistant Siri fails to understand voice input and reacts strangely or not at all. In addition, the severity of errors can differ. Interestingly, most prior research has only examined consumer reactions to AI either underperforming relative to personal expectations or overperforming relative to human standards. For example, Longoni et al. (2019) found a weaker general consumer preference for AI despite outperforming humans. Some research in the field of information systems (IS) and robotics has focused on the consequences of errors in human-robot interactions (Giuliani et al., 2015; Primiero, 2014). For example, humans display different body languages when confronted with error situations caused by a robot, (Giuliani et al., 2015) and the cause of the error (i.e., lack of effort vs. lack of ability) predicts the level of responsibility attributed to a robot (van der Woerdt & Haselager, 2019).

Although errors are usually associated with undesirable outcomes, it is particularly interesting that some types of errors can also evoke positive reactions. The pratfall effect describes a situation in which committing an error elicits positive instead of negative responses (Aronson et al., 1966). For example, when a robot commits a minor error, it makes the robot more likeable (Mirnig et al., 2017; Ragni et al., 2016). While it is reasonable to assume varying effects of different forms of error (e.g., social vs. technical errors or minor vs. severe errors) in human-AI interactions, existing research has tended to represent a generalized understanding of the error terminology, and a more granular exploration of potentially varying consequences is lacking. From a marketing perspective, we address this gap in the current literature and explore consumer responses to these different types of errors in human-AI interactions.

Third, owing to the scant research on AI-induced errors, there is currently little understanding of how to deal with errors committed by AI from a practitioner’s point of view. Some pioneering work provides practical recommendations for minimizing the negative impact of likely errors in the application of AI in a business context. Srinivasan and Sarial-Abi (2021) show that it is favorable to dehumanize an erring algorithm because when an error occurs, consumers are likely to blame an anthropomorphized algorithm more harshly than a dehumanized one. To this extent, erring algorithms based on machine learning or those supervised by humans evoke more negative consumer responses than simple algorithms and technological supervision (Srinivasan & Sarial-Abi, 2021). Consequently, companies are advised not to publicize the fact that the employed algorithm is based on machine learning or that humans supervise the algorithm to attenuate negative consumer responses to errors. Apart from these empirical findings, there is little research on managerial strategies for coping with AI-induced errors. We address this research gap by exploring the potential of eXplainable AI (XAI) as a possible mitigation strategy for erring AI. XAI is a new class of AI that can explain how a specific output is generated (van der Waa et al., 2021), providing consumers the opportunity to comprehend the AI's decision-making process and interpret a generated outcome (Miller, 2019; Shin, 2021). Research on XAI yields evidence that providing users with explanations of how a system works improves their attitudes toward AI (Rai, 2020; van der Waa et al., 2021). Therefore, we explored XAI as a potential mechanism to deal with consumer responses evoked by AI errors.

To address the identified gaps in research, we conducted four studies, scrutinizing consumer responses to different error types in AI applications and investigating the effectiveness of XAI as a managerial coping strategy. In Studies 1 and 2, we examined different error types (technical errors vs. social errors) with varying severities in scenario-based experiments. We show that technical errors lead to different consumer responses than social errors. The severity of the error further influences this differential effect by acting as a moderating variable. In Studies 2 and 3, we focused on the underlying mechanisms explaining the differences in consumer responses to errors in AI applications, investigating the role of cognitive and affective trust as mediators in the relationship between error type and consumer responses. Finally, in Study 4, we investigated the potential role of XAI in the context of erroneous AI outcomes. Our data provides evidence that XAI can mitigate the impact of social errors on consumer responses.

Our findings provide three main contributions to marketing theory and practice. First, we enhance the literature on the disadvantages of AI applications in marketing. Specifically, we explore errors in human-AI interactions and elucidate how consumers react to these negative outcomes in AI applications. This approach enables us to offer insights to companies and policymakers on AI's negative impact on consumers without causing actual economic damage or harming real market participants. Second, by distinguishing between different types of errors and varying error severity in our investigation, we contribute to a nuanced understanding of the impact of errors. In our study, we focus on technical errors, representing execution failures within the AI application, and social errors, describing AI outcomes that violate social norms. By specifically dealing with social errors, we provide managers with knowledge about the importance of socially responsible AI applications, considering the social norms of consumers and the prevalent data biases in AI applications. We present robust evidence that minor social errors can prompt reactions similar to error-free performance. These findings are fundamental to the theory and initial evidence that not every error is necessarily associated with negative consumer responses. Third, we enhance the understanding of XAI in the marketing context after errors occur. To date, no empirical research has investigated the effectiveness of XAI as a managerial mitigation strategy in erroneous human-AI interactions. Our findings provide important insights for managers and AI programmers seeking to integrate XAI into the AI-customer interface.

Theoretical foundations

Research on human-computer interaction demonstrates that humans can relate to computers in a manner similar to how they would relate to other humans (Nass & Moon, 2000). In our research on erroneous AI, we draw on scholarly works on the cognitive and emotional capabilities of machines and algorithms (Castelo et al., 2019; Haslam et al., 2008). The differentiation between cognitive and emotional capabilities that individuals ascribe to other entities is based on the findings from dehumanization research (Haslam et al., 2008), as further elaborated in the following.

Research on dehumanization distinguishes between dimensions that are similar to agency and experience. Human uniqueness abilities consist of cognition, civility, and culture, while human nature abilities comprise emotional and affective characteristics, such as desire, warmth, and intuition (Haslam et al., 2008). Besides humans, machines, and robots demonstrate a certain level of cognitive abilities and, therefore, possess human uniqueness abilities (Haslam et al., 2008; Loughnan & Haslam, 2007). However, machines are seen to lack human nature abilities that are associated with emotions and affective characteristics (Haslam et al., 2008; Loughnan & Haslam, 2007).

Building on dehumanization research, we emphasize the conceptual distinction between predominantly cognitive capabilities (akin to human uniqueness abilities) and primarily emotional capabilities (akin to human nature abilities). Current AI research provides evidence that intelligent algorithms and AI are perceived to display cognitive rather than emotional capabilities (Castelo et al., 2019; Loughnan & Haslam, 2007; Srinivasan & Sarial-Abi, 2021). We add to this line of literature by examining the influence of AI-induced errors on the perceived capabilities of AI.

Literature overview and hypotheses development

Erroneous AI

Prior literature claims that consumers assign cognitive, but not emotional capabilities to machines (Haslam et al., 2008; Loughnan & Haslam, 2007). Castelo et al. (2019) provide additional evidence in this regard and show that consumers trust algorithms that perform objective tasks more than subjective tasks. The authors argue that objective tasks are associated more with cognitive capabilities, whereas subjective tasks are associated with emotional capabilities (Castelo et al., 2019; Inbar et al., 2010). Based on these findings, we conclude that not only the task, but also the task outcome must be associated with AI's cognitive and emotional capabilities. However, not every task outcome is a positive outcome. A general problem with self-learning AI addressed by prior literature is that the complexity of the problems to be solved by AI is both, a blessing and a curse as AI can generate outcomes that fail to meet expectations (de Bruyn et al., 2020, p. 97). For example, negative outcomes such as errors are likely to occur because of the autonomous generation of knowledge structures with AI (de Bruyn et al., 2020).

Despite the relevance and potential impact of erroneous AI applications on consumers, research in marketing on AI-induced errors is scarce. Initial studies indicate that consumers prefer humans over AI and demonstrate more tolerance toward human errors (versus AI-induced errors), despite AI's superior performance (Dietvorst et al., 2016). Research in the field of psychology has offered various distinctions of errors, often based on the definition of errors as non-random failures to achieve an intended outcome in a planned sequence of mental or physical activities (Reason, 1990). The IS and robotics literature acknowledges errors based on their causes, such as technical errors or social errors (Giuliani et al., 2015; Honig & Oron-Gilad, 2018; Mirnig et al., 2017). Applied to our research context, *technical errors* refer to AI failing to either execute an algorithmic task (due to the lack of expertise) or generate an intended outcome due to a technical disruption of algorithmic processes. One example of a technical error is the AI application GPT-3, which repeatedly suggested incorrect timeslots for medical appointments. While the intelligent algorithm was able to execute the task of arranging medical appointments, it failed to give patients the correct time slots (Quach, 2020).

Social errors refer to outcomes that may be correct from an algorithmic standpoint, but deviate from a social script. Hence, despite AI successfully performing a task, the generated outcome is deemed inappropriate and therefore erroneous. What society considers inappropriate depends on social norms. Social norms are "social attitudes of approval and disapproval, specifying what ought to be done and what ought not to be done" (Sunstein, 1996, p. 914). Thus, social norms denote a society's shared understanding of permitted or unacceptable behavior that is not imposed by laws (Melnyk et al., 2022). Uber's AI-driven pricing algorithm committed a social error in

2017 when it identified an increase in demand for drivers and, therefore, increased the fare on the ride-hailing platform. Since the increased demand resulted from people trying to escape terrorist attacks, Uber was heavily criticized for profiting from this tragedy (Riley, 2017) and faced negative consumer responses.

Incorporating the aspect of AI-induced errors, such as technical and social errors, into the concept of dehumanization suggests that consumers ascribe different cognitive and emotional capabilities to erring AI. Based on the evidence that objective (subjective) tasks are associated with cognitive (emotional) capabilities (Castelo et al., 2019; Inbar et al., 2010), we infer that task outcomes also trigger associations of cognitive and emotional capabilities. Thus, consumers seem to have clear expectations regarding an algorithm's cognition and emotional capabilities, which are presumably revaluated after the occurrence of an error. The ascribed capabilities then determine consumers' responses to the errors (Srinivasan & Sarial-Abi, 2021).

However, before certain consumer responses to different error types can be hypothesized, prior research on service failures points to the magnitude of errors as an additional variable influencing consumers (Sivakumar et al., 2014). Rationally, one expects consumers to dislike errors and respond negatively to them. While minor errors may be seen as minor irritation, severe errors most likely evoke strong negative responses, such as frustration or annoyance. Interestingly, prior research in the field of psychology emphasizes that errors, failures, or mistakes – depending on how severe they are – can have positive effects on individuals. What is referred to as the pratfall effect comes into play when an individual, attributed with high intellectual abilities, commits a blunder or so-called pratfall, thereby showing imperfection, approachability, and, thus, attractiveness (Aronson et al., 1966). In line with the notion of the pratfall effect, studies investigating human-robot interactions reveal that, depending on the type of error, humans can have positive emotions toward robots that err (Mirnig et al., 2017). Qualitative research shows that if AI breaks a social norm, for example, by cutting off users while they speak, positive responses, such as smiling or laughing, can be elicited, whereas technical errors lead to negative user reactions, such as signs of frustration or frowning (Giuliani et al., 2015). Prior research also emphasizes that the pratfall effect can only occur for minor errors, whereas severe errors always lead to negative reactions (Aronson et al., 1966). Hence, depending on the error type, we expect error severity to be a crucial factor influencing consumer responses to errors.

Considering the different types of errors with varying severity, we expect consumers to associate changing cognitive and emotional capabilities with erring AI. Since consumers generally ascribe higher cognitive than emotional capabilities to AI (Haslam et al., 2008; Loughnan & Haslam, 2007), technical errors tarnish people's mind perception of AI as being capable on a cognitive level. Therefore, we predict that technical errors, regardless of their severity, demonstrate a lack of cognitive capability, leading to negative consumer responses. In contrast, social errors denote outcomes that are technically correct but violate a social norm. As the perception of norms varies depending on the individual (Kallgren et al., 2000), social errors represent subjective shortcomings and are presumably attributed to AI's expected lack of emotional capabilities. The findings of Castelo et al. (2019) support this assumption. According to the authors, consumers rely more on algorithms that perform objective tasks and expect AI to perform worse in subjective contexts. In this respect, social errors presumably indicate high degrees of subjectivity, reminding consumers of AI's lower emotional capabilities. Since consumers expect AI to have few emotional capabilities, we assume that consumers overlook minor social errors, leading to attenuated consumer responses in comparison to minor technical errors. In contrast, we hypothesize that consumers will not tolerate severe social errors and interpret them as proof of AI missing cognitive and emotional capabilities. We hypothesize:

H_1: Error type and error severity interact in affecting consumers, such that minor technical errors lead to more negative consumer responses than minor social errors.

The role of trust in human-AI interactions

Empirical evidence points out that trust is not only an essential factor for successful branding activities (Rajavi et al., 2019) or relationship marketing (Geyskens et al., 1996) but is also a crucial predictor of consumer responses to technology (Glikson & Woolley, 2020). Analogous to consumers ascribing cognitive and emotional capabilities to AI, there is also a cognitive and affective dimension when consumers develop trust in the technology they are using (Hildebrand & Bergner, 2021; Hoff & Bashir, 2015). We investigated the role of these trust dimensions when AI commits technical or social errors.

Early work on trust in the marketing domain conceptualized a person's willingness to rely on and confidence in an exchange partner as trust (Moorman et al., 1993). As Moorman et al. (1993) argues, this definition embraces two aspects: the person's beliefs in somebody's competence and reliability, and also the person's reliance on the trustee and, therefore, the demonstration of the trustor's vulnerability. In psychology literature, scholars began to untangle the concept of human trust, resulting in several subdimensions, including honesty and benevolence(Anderson & Weitz, 1989). More recent work in marketing adopted this view, extending our understanding of trust based on cognition and affection (Geyskens et al., 1996).

Indeed, a growing body of research shows that trust consists of both cognitive and affective components (Johnson & Grayson, 2005; Tomlinson et al., 2020; Wang et al., 2016). The central element of *cognitive trust* is the trustor's perception of the trustee's rational evaluations (Glikson & Woolley, 2020), including competence and the degree of responsibility displayed (McAllister, 1995; Qiu & Benbasat, 2009). Competence is understood as the trustee's accumulated knowledge (Johnson & Grayson, 2005), while responsibility reflects the expected likelihood of applying accumulated knowledge. The emotional dimension of a trustor-trustee relationship is represented by *affective trust* (Johnson & Grayson, 2005). McAllister (1995) recognized the importance of this emotional relationship and described affective trust as the trustee's care and sensibility from the viewpoint of the trustor (McAllister, 1995). It is the trustor's confidence in the trustee to perform a task or make a decision based on feelings and the perceived intrinsic motivation of the trustee (Johnson & Grayson, 2005). A recent line of research supports these beliefs about benevolence and integrity (Benbasat & Wang, 2005; McKnight et al., 2002). While benevolence describes the trustor's belief that the trustee cares about the partner, integrity means that the trustee maintains promises and follows a set of principles that the trustor acknowledges (Benbasat & Wang, 2005). Building on these insights, we understand cognitive trust as the trustee's ability to do what the trustor needs, whereas affective trust refers to the trustee's care and sensibility when performing a task.

Marketing scholars have applied the concepts of cognitive and affective trust in an AI context, demonstrating that both dimensions have a positive effect on consumer intentions to use algorithms (Castelo et al., 2019; Wang et al., 2016). Wang et al. (2016) showed that affective trust contributes to the enjoyment of recommendation agents, whereas cognitive trust determines the perceived usefulness of this technology (Wang et al., 2016). We explore the role of cognitive and affective trust in the context of AI errors as the underlying mechanism to explain the differences in consumer responses. Drawing on the concept of dehumanization, we expect that the cognitive and emotional capabilities ascribed to erring AI are also reflected in consumers' cognitive and affective trust in AI. Hence, AI-induced errors prompt consumers to question not only the analytic capacities of AI but also its social sensibility. Therefore, cognitive and affective trust mediate consumer responses. Thus, we hypothesize:

H_2: Cognitive and affective trust mediate the effect of error type on consumer responses.

Mitigating negative responses through explainable AI

AI outcomes can often be easily evaluated (for example, is the picture correctly identified? Does the self-driving car stop at the red light?), but humans have little understanding of how AI applications reach their conclusions (Haenlein & Kaplan, 2019). Thus, AI is often referred to as an opaque black box, (Rai, 2020) making it difficult to trust it from a consumer's point of view (Haenlein & Kaplan, 2019). Opacity becomes a problem when the understanding of how an outcome is generated is crucial, such as in the healthcare sector, where AI supports physicians with critical information (e.g., diagnosing skin cancer). AI programmers and scholars have, therefore, been working on a new class of so-called eXplainable AI (XAI), which generates output along with an explanation of how the output was determined (van der Waa et al., 2021). By providing additional information about the functioning of or reasons for the produced outcome, XAI is deemed to increase system understanding and improve the utility of its outcomes (van der Waa et al., 2021). Thus, XAI can be defined as a "class of systems that provide visibility into how an AI application makes decisions and predictions and executes its actions" (Rai 2020, p 137-138). More specifically, we note that XAI not only explains why a certain *outcome* was produced, but also provides the user with a better understanding of the *system* and its operations (Arrieta et al., 2020; Shin, 2021).

Research on XAI yields evidence that providing users with explanations of how a system works enhances their attitudes toward AI (Rai, 2020; van der Waa et al., 2021). To understand why explanations can have a positive effect on AI perception, we draw on insights from psychological research on human judgment. According to Malle (2006), individuals demand explanations when socially interacting with others to establish a shared meaning, change others' beliefs, or influence their actions. For AI, providing explanations creates such shared meaning in a human-AI interaction (Miller, 2019) and can enhance users' attitudes toward AI (Rai, 2020). Based on these findings, we emphasize the importance of the explainability of AI in building rapport between consumers and AI applications.

Explainability is also of interest concerning AI-induced errors because it helps consumers comprehend the erroneous decision-making process or interpret its outcomes (Miller, 2019; Shin, 2021). Empirical findings on human decision-making support this claim, as individuals ask for explanations to resolve dissonant information (Malle, 2006). Dissonant information refers to contradictions, non-normative actions, or incidents that people find unusual (Hilton, 1996; Malle, 2006). Thus, behavior or events that do not follow a norm or ignore the usual conventions prompt others to question the action and ask for an explanation (Hilton, 1996). In the context of this study, AI-induced social errors represent such non-normative actions. On the one hand, we anticipate that consumers appreciate XAI in the case of social errors, which, in turn, will positively affect consumer responses. Technical errors, on the other hand, are defined as the result of a technical disruption of algorithmic processes or AI's lack of expertise. Hence, following the argument of Hilton (1996), technical errors may evoke contradictions, such as illogical results or no outcome at all, but do not reflect non-normative behavior. Therefore, we expect consumers to associate technical errors with the AI's lack of expertise, requiring no explanation. Consequently, we assume that XAI has a weaker effect on consumer responses after the occurrence of a technical error versus a social error. We hypothesize:

> *H_3: Error type and XAI interact in affecting consumers, such that XAI will have a stronger positive effect on consumer responses after social errors (vs. technical errors).*

Study 1: The moderating role of error severity

To test whether technical errors provoke different consumer responses as compared to social errors depending on their severity, we conducted a scenario-based online experiment in a 2 (error type: technical versus social) × 2 (error severity: low versus high) factorial design. An error-free interaction scenario was used in the control group. We recruited 268 US-based participants (M_{age}= 35.9 years, 39.2% female) on MTurk who were randomly assigned to one of the five scenarios, resulting in approximately 50 participants per cell.

Method

Participants were introduced to an AI-based voice assistant with multiple features and invited to imagine themselves asking the assistant to tell a joke in front of family and friends. In the low-severity condition of the technical error, the voice assistant was asked to repeat the command three times before stating that it could not help. Under the high-severity condition, the voice assistant failed to respond at all. To represent a minor social error, the voice assistant told an offensive joke against blondes, whereas it told a discriminatory joke against people of color, as a social error of high severity. In the control group, the voice assistant told an inoffensive and nondiscriminatory joke (see Appendix A).

To measure AI likability we used four items: “dislike/ like,” “unfriendly/ friendly,” “unkind/ kind,” and “unpleasant/ pleasant” (Bartneck et al., 2008). Moreover, we measured AI competence and future use intention on seven-point Likert scales (1 = “strongly disagree” to 7 = “strongly agree”) (Cuddy et al., 2008; Qiu & Benbasat, 2009). As control variables, we measured familiarity with voice assistants (Gillath et al., 2021), and participants also provided their basic demographic information, including their ethnicity, due to the discriminatory joke against people of color. See Table 1 for the descriptive statistics.

Table 1. Descriptive statistics (Study 1, Study 2, Study 3, and Study 4)

					Study 1					
		M	SD	α	CR	1	2	3	4	5
1	AI competence	4.35	1.92	0.93	0.95	**0.91**				
2	AI likability	4.59	1.85	0.92	0.94	.83***	**0.90**			
3	Use intention	4.23	2.10	0.94	0.96	.85***	.83***	**0.94**		
					Study 2					
		M	SD	α	CR	1	2	3	4	5
1	AI competence	4.56	1.63	0.92	0.95	**0.9**				
2	AI likability	4.60	1.41	0.92	0.95	.61***	**0.91**			
3	Use intention	3.69	2.04	0.98	0.99	.54***	.61***	**0.98**		
4	Cognitive trust	4.53	1.65	0.95	0.96	.78***	.59***	.55***	**0.93**	
5	Affective trust	3.83	1.61	0.91	0.93	.74***	.70***	.55***	.66***	**0.86**
					Study 3					
		M	SD	α	CR	1	2	3	4	
1	AI competence	4.20	1.72	0.93	0.95	**0.91**				
2	Use intention	4.29	1.86	0.98	0.99	.55***	**0.98**			
3	Cognitive trust	4.25	1.54	0.95	0.96	.77***	.61***	**0.93**		
4	Affective trust	3.25	1.45	0.89	0.92	.65***	.39***	.59***	**0.84**	
					Study 4					
		M	SD	α	CR	1	2	3	4	5
1	AI competence	3.82	1.89	0.95	0.97	**0.94**				
2	AI usefulness	3.60	2.07	0.96	0.97	.86***	**0.95**			
3	Use intention	3.25	2.09	0.98	0.99	.79***	.83***	**0.98**		

Note. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$; M = Mean; SD = Standard Deviation; α = Cronbach's alpha; CR = Composite Reliability; Diagonal elements in bold are square roots of the average variance extracted (AVE); Off-diagonal elements are inter-construct correlation.

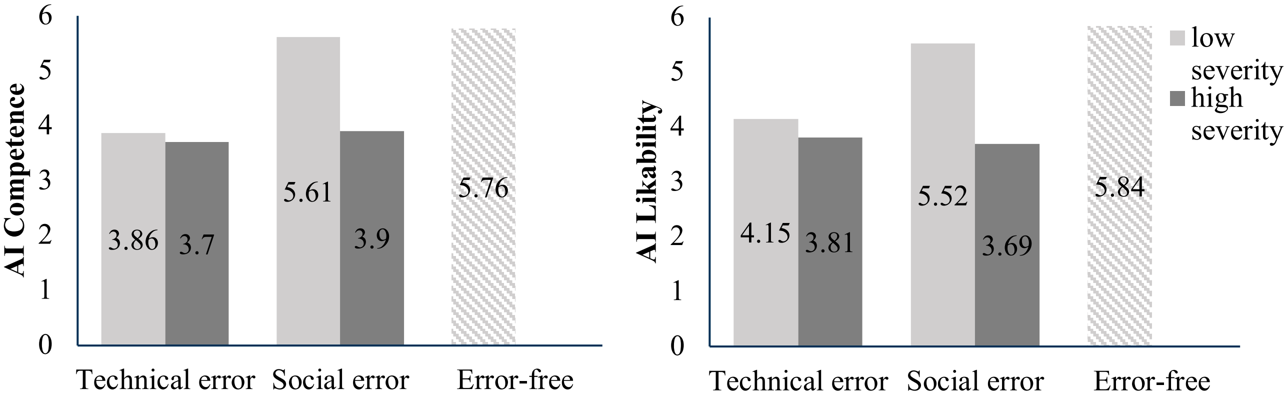
Analysis and discussion

Manipulation check: We used 7-point Likert scales to test whether the participants regarded the interaction with the voice assistant as technically or socially incorrect. As intended, t-tests revealed that technical errors were perceived as technically more incorrect than social errors (M_{tech} = 4.61 vs. M_{social} = 3.21 (vs. $M_{error\text{-}free}$ = 2.16); t(216) = 5.228, *p* < .001), whereas social errors were regarded as socially more incorrect (M_{tech} = 3.99 vs. M_{social} = 4.70 (vs. $M_{error\text{-}free}$ = 2.98); t(216) = -2.856, *p* = .005). Furthermore, error severity was successfully manipulated ($M_{lowseverity}$ = 4.00 vs. $M_{highseverity}$ = 4.79; t(216) = -3.666, *p* < .001).

Results: We found support for H_1 as our analysis revealed significant interaction effects between error type and error severity on consumer responses. Specifically, we found a significant interaction effect between error type and error severity on AI competence (F(1,214) = 4.221, *p* = .041) and significant main effects for error type (F(1,214) = 14.053, *p* < .001) and severity (F(1,214) = 11.021, *p* = .001). For technical errors, error severity had no significant impact on AI competence ($M_{lowSeverity}$ = 3.77 vs. $M_{highSeverity}$ = 3.46; t(107) = .817, *p* = .416), but error severity led to significant differences in AI competence for social errors ($M_{lowSeverity}$ = 5.21 vs. $M_{highSeverity}$ = 3.88, t(107) = 4.250, *p* < .001). Furthermore, we found a significant interaction effect between error type and error severity on use intention (F(1,214) = 12.441, *p* < .001) and significant main effects for error type (F(1,214) = 13.828, *p* < .001) and error severity (F(1,214) = 13.314, *p* < .001). Again, in the technical error condition, error severity did not affect use intention ($M_{lowSeverity}$ = 3.44 vs. $M_{highSeverity}$ = 3.41; t(107) = .081, *p* = .935), whereas use intention differed significantly in the social error condition ($M_{lowSeverity}$ = 5.37 vs. $M_{highSeverity}$ = 3.46, t(107) = 5.409, *p* < .001). When we included control variables, such as ethnicity and age, in our two analyses, the interaction effects remained significant. These results provided support for error severity acting as a moderator between error type and consumer responses.

In addition, the data revealed a significant interaction effect between error type and error severity on AI likability (F(1,214) = 10.067, *p* = .002) as well as main effects for error type (F(1,214) = 7.045, *p* = .009) and error severity (F(1,214) = 21.497, *p* < .001). While error severity had no significant impact on AI likability after technical errors ($M_{lowSeverity}$ = 4.15 vs. $M_{highSeverity}$ = 3.81; t(107) = 1.023, *p* = .309), we found significant differences in AI likability after social errors ($M_{lowSeverity}$ = 5.52 vs. $M_{highSeverity}$ = 3.69; t(107) = 5.590, *p* < .001), supporting the notion of the pratfall effect for low-severity social errors (see Figure 1). Interestingly, when we compared AI perceptions after minor social errors and the control group (error-free performance), we found that perceptions were not significantly different (AI likability: $M_{lowSeverity}$ = 5.52 vs. $M_{error\text{-}free}$ = 5.84, t(103) = -1.501, *p* = .068; AI competence: $M_{lowSeverity}$ = 5.21 vs. $M_{error\text{-}free}$ = 5.52, t(103) = -1.226, *p* = .223). These results suggest that consumers do not differentiate between error-free AI and AI that makes minor social errors.

Figure 1. Interaction Effect on AI Competence and AI Likability (Study 1)



According to the results of Study 1, error severity moderates the relationship between error type and consumer responses. Minor technical errors lead to more negative consumer responses than minor social errors do, supporting our predictions (H_1). In addition, as consumers like AI committing minor social errors we find support for the notion of the pratfall effect.

Study 2: The mediating role of cognitive and affective trust

To assess the robustness of our findings from Study 1, we replicated the experiment by recruiting participants from a different panel provider. In addition, we investigated the mediating role of cognitive and affective trust at varying levels of error severity. For this purpose, we recruited 339 US-based participants on Prolific (M_{age}= 39.3, 45.4% female) and conducted a 2 (error type: technical versus social) × 2 (severity: low versus high) factorial design experiment with an error-free baseline condition as before.

Method

As in the previous study, participants were introduced to an AI-based voice assistant, who was asked to tell a joke in front of family and friends. The scenarios were the same as in Study 1, and the participants were randomly assigned to one of the five scenarios. We also incorporated and adapted multiple items from various seven-point Likert scales to measure consumers' cognitive and affective trust (1 = "strongly disagree" to 7 = "strongly agree"). We measured cognitive trust using six items (McAllister, 1995; McKnight et al., 2002; Qiu & Benbasat, 2009) and adjusted their wording to the AI context wherever needed. To measure affective trust, we used eight items for the subdimension "benevolence" and five items for "integrity" (adapted from Gillath et al., 2021; Hildebrand & Bergner, 2021; Johnson & Grayson, 2005; McAllister, 1995; McKnight et al., 2002; Qiu & Benbasat, 2009). We performed a confirmatory factor analysis (CFA) that indicated a very good model fit and reliability of our items (see Appendix C). Furthermore, discriminant validity among our constructs was also supported (see Table 1).

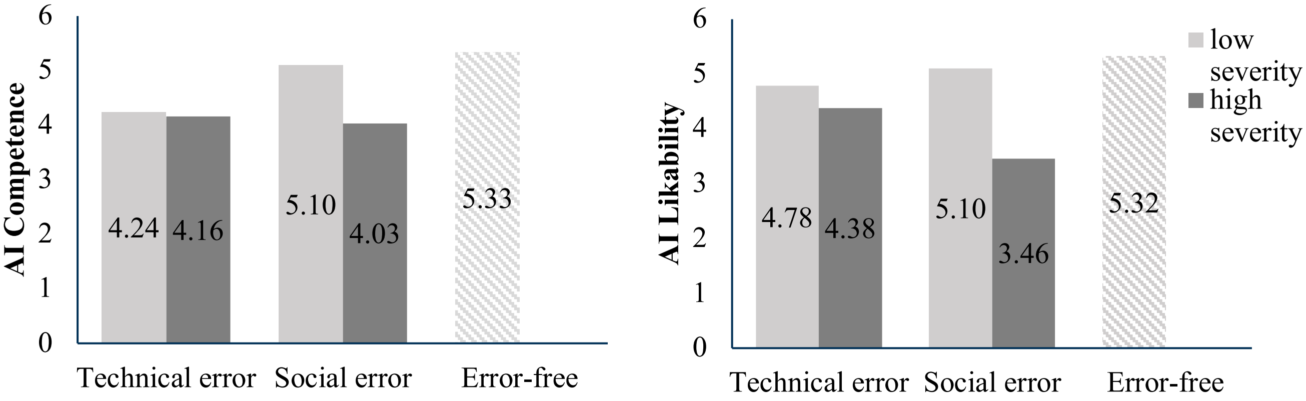
Analysis and discussion

Manipulation check: As intended, t-tests showed that technical errors were considered technically more incorrect (M_{tech} = 5.51 vs. M_{social} = 2.98 (vs. $M_{error\text{-}free}$ = 1.87); t(271) = 11.843, $p < .001$), whereas social errors were seen socially more incorrect (M_{tech} = 3.03 vs. M_{social} = 4.47 (vs. $M_{error\text{-}free}$ = 1.53); t(271) = -6.132, $p < .001$). The manipulation of error severity was also successful ($M_{lowseverity}$ = 2.16 vs. $M_{highseverity}$ = 4.04; t(271) = -7.751, $p < .001$).

Results: Again, we found significant interaction effects between error type and error severity on AI competence (F(1,269) = 6.970, $p = .009$) and use intention (F(1,269) = 29.604, $p < .001$). Consistent with our results obtained in Study 1, there was no difference in AI competence or use intention after a technical error, despite different error severities (AI competence: $M_{lowSeverity}$ = 4.24 vs. $M_{highSeverity}$ = 4.16; t(135) = .284, $p = .777$; use intention: $M_{lowSeverity}$ = 3.35 vs. $M_{highSeverity}$ = 3.52; t(135) = -.522, $p = .603$). In the social error condition, however, consumers perceived the

AI-based voice assistant committing minor errors as significantly more competent ($M_{lowSeverity}$ = 5.10 vs. $M_{highSeverity}$ = 4.03; t(134) = 4.090, *p* < .001) and displayed higher intentions to use the assistant again ($M_{lowSeverity}$ = 4.62 vs. $M_{highSeverity}$ = 2.35; t(134) = 7.450, *p* < .001). The interaction results did not change when hair color or ethnicity was controlled for.

Figure 2. Interaction Effect on AI Competence and AI Likability (Study 2)



To investigate whether we find evidence of the pratfall effect, we ran a factorial Analysis of Variance (ANOVA) with AI likability as the dependent variable. As in Study 1, we discerned a significant interaction effect between error type and error severity on AI likability (F(1,269) = 16.556, *p* < .001). The data yielded significant differences in the technical error condition ($M_{lowSeverity}$ = 4.78 vs. $M_{highSeverity}$ = 4.38; t(135) = 2.129, *p* = .035) as well as in the social error condition, where AI likability was significantly higher for minor social errors than for social errors of high severity ($M_{lowSeverity}$ = 5.10 vs. $M_{highSeverity}$ = 3.46, t(134) = 6.691, *p* < .001). We then compared AI perceptions after minor social errors with those in the control group. As in Study 1, we found that these perceptions were not significantly different (AI likability: $M_{lowSeverity}$ = 5.10 vs. $M_{error\text{-}free}$ = 5.32, t(131) = -1.015, *p* = .312; AI competence: $M_{lowSeverity}$ = 5.10 vs. $M_{error\text{-}free}$ = 5.33, t(131) = -.891, *p* = .375), indicating that minor social errors trigger the same AI perceptions as error-free AI results (see Figure 2). When we controlled for hair color, the results did not change.

The replication of Study 1's results demonstrates the robustness of the effects and our previous findings.

To analyze the mediating effects of cognitive and affective trust on consumer responses, we estimated a moderated mediation model (Hayes, 2017 Model 7; 5,000 bootstrap samples). Table 2 presents the results of our models for the two dependent variables, AI competence and use intention. The discriminant validity of the two mediators was confirmed using the Fornell-Larcker criterion (Fornell & Larcker, 1981).

Table 2. Regression and mediation analysis (Study 2)

(N = 339)	Coefficient	s.e.	t	95% CIs
Interaction effects on mediators				
Error type[a] × error severity[b] on cognitive trust	-1.15**	.38	-3.01	[-1.90, -.40]
Error type[a] × error severity[b] on affective trust	-.88*	.38	-2.34	[-1.62, -.14]
Mediation effect of cognitive trust on AI competence				
(conditional indirect effects, error type[a]→cognitive trust→AI competence)				
Error severity = low	.41	.15	–	[.13, .70]
Error severity = high	-.19	.14	–	[-.48, .08]
Index of moderated mediation	-.59	.21	–	[-1.02, -.20]
Mediation effect of affective trust on AI competence				
(conditional indirect effects, error type[a]→affective trust→AI competence)				
Error severity = low	.26	.11	–	[.05, .49]
Error severity = high	-.07	.10	–	[-.28, .12]
Index of moderated mediation	-.33	.16	–	[-.66, -.05]
Mediation effect of cognitive trust on use intention				
(conditional indirect effects, error type[a]→cognitive trust→use intention)				
Error severity = low	.28	.13	–	[.07, .57]
Error severity = high	-.13	.11	–	[-.38, .06]
Index of moderated mediation	-.41	.19	–	[-.85, -.10]
Mediation effect of affective trust on use intention				
(conditional indirect effects, error type[a]→affective trust→use intention)				
Error severity = low	.27	.14	–	[.05, .58]
Error severity = high	-.08	.11	–	[-.30, .13]
Index of moderated mediation	-.35	.18	–	[-.76, -.05]

Note. $^{*}p < .05$, $^{**}p < .01$; s.e. = standard error. CI = Confidence Interval. [a] Dummy: technical error (1) vs. social error (2); [b] Dummy: low (1) vs. high (2) error severity.

The analysis revealed significant moderated mediation of AI competence by cognitive trust (index = -.59, 95% CI = [-1.02, -.20]) and affective trust (index = -.33, 95% CI = [-.66, -.05]). Both mediators significantly impacted the relationship between error type and AI competence ($ß_{\text{cognitive}}$ = .52, p < .001; $ß_{\text{affective}}$ = .37, p < .001; R^2 = .69) and rendered the direct effect of error type on

AI competence insignificant ($ß_{direct}$ = .17, p = .13). When analyzing for mediating effects on use intention, we also found a significant moderated mediation through cognitive trust (index = -.41, 95% CI = [-.85, -.10]) and affective trust (index = -.35, 95% CI = [-.76, -.05]). These results were robust when controlling for hair color, ethnicity, or demographics.

In summary, the results of Study 2 are twofold. First, supporting H_1, error types and error severity interact such that minor technical errors evoke more negative consumer responses than minor social errors. In addition, we again found a notion of the pratfall effect, suggesting that consumers regard AI committing minor social errors as likable and competent as error-free AI. Second, the two trust dimensions of cognitive and affective trust mediate the effects of error type on consumer responses. This finding supports H_2 and emphasizes the importance of the affective trust dimension, which has been less regarded in the literature thus far, concerning erroneous AI.

Study 3: The mechanism of trust before and after AI-induced errors

To better understand the mechanisms of cognitive and affective trust as mediators, we conducted another scenario-based online experiment. We ran a 2 × 2 between-within design and measured consumers' cognitive and affective trust in AI for different error types (technical versus social errors) at two different time points (before versus after an error occurred). The error severity was kept constant and represented both minor errors. An error-free interaction scenario was used in the control group. In total, we randomly assigned 301 participants (M_{age}= 33, 50.8% female) from the panel provider Prolific to one of the three scenarios.

Method

Across our three scenarios, we introduced participants to an AI-based voice assistant with multiple functions, such as playing a trivia, ordering products online, or making appointments. For the manipulation, the scenario described a voice assistant being asked to play music. In the technical error condition, the voice assistant failed to react to several vocal commands and did not respond.

In the social error condition, the voice assistant played music with an explicit song text from the rapper “CupcaKKe”. In the control condition, the voice assistant played the song “Blinding Lights” by The Weeknd, which does not contain explicit lyrics. The questionnaire followed the scenario and contained the same scales as in Study 2.

Analysis and discussion

Manipulation check: As intended, t-tests uncovered that technical errors were perceived as technically more incorrect than social errors (M_{tech} = 6.32 vs. M_{social} = 4.30 (vs. $M_{error\text{-}free}$ = 1.84); t(200) = 10.155, $p < .001$), whereas social errors were regarded as socially more incorrect (M_{tech} = 3.65 vs. M_{social} = 4.73 (vs. $M_{error\text{-}free}$ = 1.51); t(200) = -4.673, $p < .001$).

Results: We conducted t-tests to explore potential differences in consumer responses to AI depending on the type of error induced in the scenario. We found that AI competence was perceived as significantly more positive after AI committed a social error versus a technical error (M_{tech} = 2.923 vs. M_{social} = 4.29; t(200) = -6.578, $p < .001$). Similarly, participants’ use intention was significantly higher in the social error condition than in the technical error condition (M_{tech} = 3.80 vs. M_{social} = 4.26, t(200) = -1.749, $p = .041$. When we included the control variables of error severity and familiarity with AI, our results remained significant.

Mediation analyses: In line with H_2, we aimed to explore cognitive and affective trust as the potential underlying mechanism to explain differences in consumer responses to AI-induced errors. To that effect, we conducted mediation analyses (Hayes, 2017 Model 4, 5,000 bootstrap samples) with error type as the independent variable, cognitive and affective trust as dual mediators, and AI competence and use intention as the dependent variables. We assessed the discriminant validity of the two mediators using the Fornell-Larcker (1981) criterion (see Table 1).

Our results showed a significant indirect effect between error type and AI competence ($ß_{TOTAL}$ = .92, 95% CI: = [.56; 1.26]), with the direct effect remaining significant when accounting for the mediators (ß$_{\text{direct}}$ = .52, p = .001). The effects of error type on AI competence were partially mediated by cognitive trust (ß$_{\text{cognitive}}$ = .64, 95% CI: = [.37; .92]) and affective trust (ß$_{\text{affective}}$ = .28, 95% CI: = [.12; .48]), with cognitive trust being the main dimension influencing AI competence. Moreover, we examined the relationship between error type and use intention and found full mediation through cognitive and affective trust, as the indirect effect was significant (ß$_{\text{indirect}}$ = .81, 95% CI: = [.47; 1.15]), whereas the direct effect was not (ß$_{\text{direct}}$ = -.35, p = .13). Both mediators had a significant effect on use intention (ß$_{\text{cognitive}}$ = .59, 95% CI = [.29; .92]; ß$_{\text{affective}}$ = .22, 95% CI = [.001; .49]). Again, our results indicate that cognitive trust was the main driver of the mediation effects (see Table 3).

Table 3. Regression and mediation analysis (Study 3)

(N = 301)	Coefficient	s.e.	t	95% CIs
Mediation effect on AI competence[a]				
Total indirect effect	.92	.18	–	[.56, 1.26]
Indirect effect through cognitive trust	$.64^{***}$	.14	–	[.37, .92]
Indirect effect through affective trust	$.28^{***}$	.09	–	[.12, .48]
Mediation effect on use intention[a]				
Total indirect effect	.81	.17	–	[.47, 1.15]
Indirect effect through cognitive trust	$.59^{***}$	.16	–	[.29, .92]
Indirect effect through affective trust	$.22^{*}$	.13	–	[.00, .49]

Note. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$; s.e. = standard error. CI = Confidence Interval. [a] Dummy: technical error (1) vs. social error (2)

Between-within analysis: We conducted a repeated-measures factorial ANOVA with cognitive trust as the dependent variable and found a significant interaction effect between error type and pre-post-timing (F(2,298) = 45.082, p < .001) as well as significant main effects of error type (F(2,298) = 16.316, p < .001) and pre-post-timing (F(1,298) = 315.858, p < .001). The analysis revealed that regardless of the error type, cognitive trust decreased significantly after the occurrence of an error (p < .001). Tukey's A-HSD test (p < .05) showed that the cognitive trust levels (post-timing) in all three conditions were significantly different from each other, with the lowest mean in the technical error condition (M$_{\text{tech}}$ = 3.43 vs. M$_{\text{social}}$ = 4.50 vs. M$_{\text{error-free}}$ = 5.16).

A second repeated-measures factorial ANOVA with affective trust as the dependent variable again resulted in a significant interaction effect between error type and pre-post-timing (F(2,298) = 45.002, *p* < .001), as well as significant main effects of error type (F(2,298) = 5.481, *p* = .005) and pre-post-timing (F(1,298) = 92.440, *p* < .001). Again, we found that the occurrence of AI-induced errors leads to a decrease in affective trust, resulting in the lowest affective trust level in the technical error condition (M_{tech} = 2.71 vs. M_{social} = 3.73 vs. $M_{error\text{-}free}$ = 3.93; F(2,298) = 21.647, *p* < .001). A post-hoc analysis with Tukey's A-HSD test (*p* < .05) revealed that affective trust in the technical error condition was significantly different from the means of the other two conditions (social error and control group). Consequently, we provide evidence that technical errors (versus social errors) influence affective trust more (less) negatively, leading to more (less) negative consumer responses.

Study 3 offers two key findings. First, in support of H_1, we again provide evidence that consumers respond more negatively to AI committing technical errors than to AI committing social errors. Second, supporting H_2, cognitive and affective trust mediates the relationship between error types and consumer responses, that is, AI competence and use intention. We found that the occurrence of AI-induced errors leads to a decrease in cognitive trust, regardless of the error type. Affective trust, however, is particularly low after technical errors, which, in turn, negatively affects consumer responses. The influence of social errors on affective trust is not as strong, leading to more positive consumer responses than technical errors.

Study 4: The moderating role of XAI

Finally, we conducted an experiment to investigate whether XAI is a helpful measure for mitigating the negative effects of AI-induced errors. We recruited 337 participants on MTurk and randomly assigned them to one of four conditions of a 2 (error type: technical versus social) × 2

(XAI: absent versus present) factorial design with two error-free control groups (XAI: absent versus present).

Method

Participants were again introduced to an AI-based voice assistant with multiple functions. In this scenario, the participants were invited to imagine themselves asking the voice assistant to make a gift recommendation for a friend. This scenario was chosen because real AI-based voice assistants such as Google or Amazon's Alexa have similar functions. Analogous to Amazon's Alexa, our error-free baseline scenario described the voice assistant responding to the user's command and putting three potential gifts in an online shopping cart. The participant was then able to make the final decision and purchase one of the items. In the control group, the shopping cart contained a unisex sports t-shirt, a pair of colorful socks, and a wall calendar. Under the social error condition, the shopping cart contained items similar to those in the control group. However, every item reflected a social error: the t-shirt read "Keep calm and hit her," the socks displayed marijuana leaves, and the calendar contained nude pictures. To increase external validity, we chose items offered in the past on Amazon. In the technical error condition, the voice assistant reacted to the vocal commands, but failed to execute the task.

To manipulate XAI, we included (omitted) the sentence "I'm using your search history and customer ratings to make recommendations" as a response of the voice assistant to the participant's command. This manipulation of XAI represented a "global explanation" of the functioning of the system, as it reveals, in parts, the process of how XAI generates an outcome (Rai, 2020). To emphasize the XAI feature in our scenarios, we also added a post-hoc explanation with the voice assistant concluding its task by saying "Based on your search history and customer ratings, I found three potential gifts. I've added them to your shopping cart." Post-hoc explanations are a typical feature of XAI (Kenny et al., 2021) and have been used in other studies (Ehsan et al., 2019). To ensure external validity, we kept both explanations as short as possible, assuming that real consumers would not be willing to listen to elaborate explanations of the voice assistant. Because we manipulated the level of explainability of AI, we expected consumers to better understand the voice assistant and its task outcomes with XAI present (versus absent). Therefore, we tested whether consumers would find the voice assistant not only more competent but also more useful in performing this type of task. To measure AI usefulness, we included four items on a seven-point Likert scale (1 = "strongly disagree" to 7 = "strongly agree") (Qiu & Benbasat, 2009). In addition, we applied the same scales as those used in previous studies.

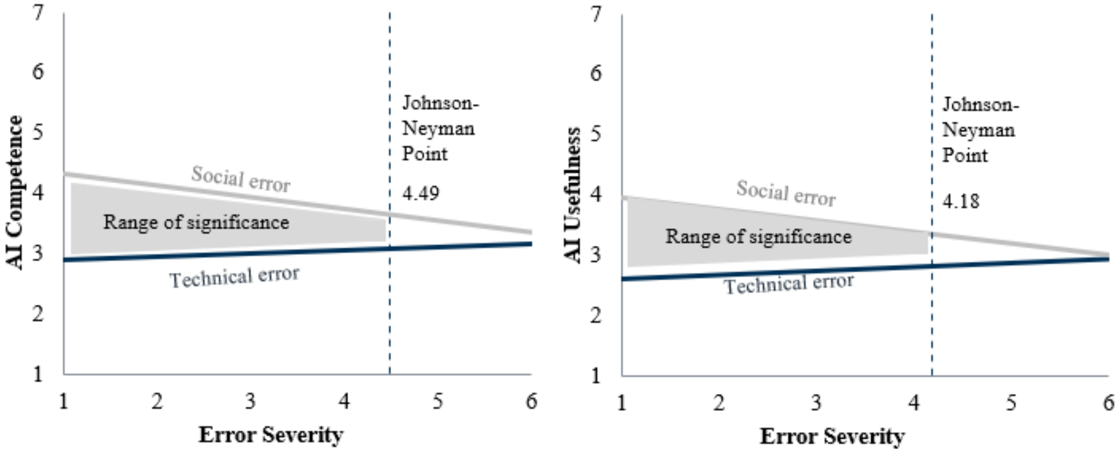
Analysis and discussion

Manipulation check: As intended, technical errors were regarded as technically more incorrect than social errors (M_{tech} = 5.62 vs. M_{social} = 4.51 (vs. $M_{error\text{-}free}$ = 3.12), t(243) = 5.376, $p < .001$). Social errors were perceived as socially more incorrect than technical errors (M_{tech} = 3.43 vs. M_{social} = 5.08 (vs. $M_{error\text{-}free}$ = 2.61); t(243) = -7.618, $p < .001$). Moreover, XAI was successfully manipulated, as participants rated the voice assistant's explainability as significantly higher in the scenarios when XAI was present than when it was absent ($M_{XAIabsent}$ = 3.20 vs. $M_{XAIpresent}$ = 4.97, t(335) = -9.519, $p < .001$).

Results: We again found support for H_1 as our analysis revealed significant interaction effects between error type and error severity on AI competence. We conducted moderation analyses (Hayes, 2017 Model 1, 5,000 bootstrap samples), with error type as the independent variable, error severity as the moderator, and AI competence as the dependent variable. The results showed that error severity significantly moderated the effect of error type on perception of AI competence (ΔR^2 = 4.30%, F(1, 241) = 11.348, $p < .001$, 95% CI [-.65, -.17]). Moreover, we were interested in how consumers perceive the voice assistant's usefulness after it committed either a technical or a social error. We analyzed the moderation effect through error severity and found a significant moderation effect between error type and AI usefulness (ΔR^2 = 4.00%, F(1, 241) = 10.346, *p* = .002, 95% CI [-.67, -.16]).

To better understand the impact of error severity on both variables, we conducted a floodlight analysis based on the Johnson-Neyman technique to identify where the simple effects of error type on AI competence and AI usefulness were significant (Spiller et al., 2013). Regarding AI competence, we found that up to an error severity of 4.49, AI competence was rated significantly higher in the social error condition (versus the technical error condition; Figure 3). In the case of AI usefulness, the analysis revealed significantly higher ratings after social errors (versus technical errors) up to the Johnson-Neyman point of 4.18.

Figure 3. Error type × error severity on AI competence and AI usefulness



Double moderation effects: Accounting for XAI as a second moderator in the relationship between error type and consumer responses, we conducted three double moderation analyses (Hayes, 2017 Model 2, 5,000 bootstrap samples). Our data showed that error severity and XAI moderated the effect between error type and AI competence significantly (ΔR^2 = 5.16%, F(2, 239) = 6.879, *p* = .001). The conditional effects of error type at different levels of error severity and XAI revealed that after social errors, AI competence was positively impacted by the presence of XAI for minor to moderately severe errors (see Table 4).

Table 4. Regression and mediation analysis (Study 4)

(N = 337)	Coefficient	s.e.	t	95% CIs
Interaction effects on AI usefulness				
Error type[a] × error severity	-.42**	.13	-3.22	[-.67, -.16]
Conditional effects on AI usefulness at				
1 SD below the mean of error severity = 2.12	1.34***	.34	3.97	[.67, 2.00]
Mean of error severity = 3.94	.57*	.24	2.40	[.10, 1.04]
1 SD above the mean of error severity = 5.77	-.19	.34	-.58	[-.86, .47]
Interaction effects on AI competence				
Error type[a] × error severity	-.41***	.12	-3.37	[-.65, -.17]
Conditional effects on AI competence at				
1 SD below the mean of error severity = 2.12	1.44***	.32	4.55	[.81, 2.06]
Mean of error severity = 3.94	.68**	.22	3.07	[.24, 1.12]
1 SD above the mean of error severity = 5.77	-.07	.32	-.22	[-.69, .55]
Conditional effects on AI competence at values of both moderators[b]:				
1 SD below the mean of error severity, XAI absent	1.02*	.40	2.57	[.24, 1.81]
1 SD below the mean of error severity, XAI present	1.74***	.37	4.67	[1.01, 2.47]
Mean of error severity = 3.94, XAI absent	.31	.32	.98	[-.31, .93]
Mean of error severity = 3.94, XAI present	1.03**	.31	3.31	[.41, 1.64]
1 SD above the mean of error severity, XAI absent	-.40	.38	-1.07	[-1.14, .34]
1 SD above the mean of error severity, XAI present	.31	.39	.80	[-.46, 1.09]
Conditional effects on AI usefulness at values of both moderators[b]:				
1 SD below the mean of error severity, XAI absent	.89*	.42	2.11	[.06, 1.73]
1 SD below the mean of error severity, XAI present	1.69***	.40	4.24	[.90, 2.47]
Mean of error severity = 3.94, XAI absent	.16	.34	.49	[-.50, .83]
Mean of error severity = 3.94, XAI present	.96**	.33	2.89	[.30, 1.61]
1 SD above the mean of error severity, XAI absent	-.56	.40	-1.41	[-1.36, .23]
1 SD above the mean of error severity, XAI present	.23	.42	.54	[-.60, 1.05]
Conditional effects on use intention at values of both moderators[b]:				
1 SD below the mean of error severity, XAI absent	-.44	.43	-1.02	[-1.28, .41]
1 SD below the mean of error severity, XAI present	.81*	.40	2.03	[.02, 1.60]
Mean of error severity = 3.94, XAI absent	-.91**	.34	-2.66	[-1.58, -.24]
Mean of error severity = 3.94, XAI present	.35	.33	1.03	[-.31, 1.00]
1 SD above the mean of error severity, XAI absent	-1.37***	.40	-3.39	[-2.17, -.58]
1 SD above the mean of error severity, XAI present	-.12	.42	-.29	[-.95, .71]

Note. $^{*}p < .05$, $^{**}p < .01$, $^{***}p < .001$; s.e. = standard error. CI = Confidence Interval. [a]Dummy: technical error (1) vs. social error (2); [b]Dummy: XAI absent (1) vs. present (2)

However, XAI did not seem to influence AI competence after a technical error. For AI usefulness, double moderation analysis with error severity and XAI as moderators yielded a significant double moderation effect (ΔR^2 = 5.01%, F(2, 239) = 6.554, *p* = .002). The results showed that XAI had a positive effect on AI usefulness after the occurrence of social errors. In the technical error condition, however, we found that XAI attenuated AI usefulness, regardless of error severity. We found similar results for use intention as the dependent variable (ΔR^2 = 4.59%, F(2, 239) = 5.797, *p* = .004). These findings were unexpected. A possible interpretation of the negative effect of XAI on use intention is that its presence might have triggered consumers to expect AI to generate a certain outcome. In the technical error condition, however, the voice assistant did not make any gift recommendations at all, potentially resulting in AI falling short of consumers' performance expectations, subsequently leading to lower use intention rates.

The key findings of Study 4 are twofold. First, supporting H_1, consumer responses to AI-induced errors are moderated by perceived error severity. Minor technical errors lead to significantly more negative consumer responses than minor social errors do. However, in cases of severe AI-induced errors, consumers respond negatively, regardless of the type of error. Second, XAI moderates the effect of the error type on consumer responses. In particular, we found that XAI positively affects consumer responses to social errors. In the case of technical errors, XAI has no effect on AI competence or even a negative effect on AI usefulness and use intention. Most likely, these negative effects occur because the voice assistant falls short of consumers' performance expectations in the presence of XAI. In summary, our findings partially support H_3.

Discussion

In four studies and across different scenarios of human-AI interactions, including AI telling a joke, playing music, or making a gift recommendation, we examine consumer responses to AI-induced errors. Our robust findings show that AI committing severe errors, regardless of the type of error, harms consumers' AI perceptions and use intentions. Interestingly, minor social errors lead to fewer negative consumer responses than do minor technical errors. Applying prior insights from dehumanization research (Haslam et al., 2008), we posit that technical errors taint consumers' perception of AI as being capable on a cognitive level. In contrast, the occurrence of minor social errors is somewhat expected by consumers because of the limited emotional capabilities ascribed to machines and AI (Loughnan & Haslam, 2007), evoking less negative consumer responses than minor technical errors.

Additionally, our study supports the notion of the pratfall effect when consumers interact with AI (Studies 1 and 2). The pratfall effect describes the phenomenon in which a person committing a blunder appears more attractive (Aronson et al., 1966). This effect also appears to apply to AI committing minor social errors. While minor technical errors lead to negative consumer responses per se, we find evidence that minor social errors yield results similar to error-free performance in terms of consumers' perceived AI competence and AI likability. Thus, the present data are consistent with findings in the field of robotics, which show that faultiness or imperfection can trigger positive associations with the entities with which consumers interact (Mirnig et al., 2017; Ragni et al., 2016). While the original study by Aronson et al. (1966) did not specify the type of mistake evoking the pratfall effect, we identified minor social errors that trigger this phenomenon in an AI context. Therefore, we hone our understanding of the pratfall effect and specify when it is likely to occur in human-AI interactions. In contrast to findings from Aronson et al. (1966), who show that a person committing a blunder appears more attractive or approachable, our data only indicate that minor social errors evoke AI likability similar to error-free performance. A possible reason for this discrepancy might be the evident differences in the entities committing the blunder. Feelings such as attractiveness or being approachable are hardly transferable to the context of faceless AI. In other words, unlike celebrities or well-known experts who could appear unapproachable to consumers, AI cannot evoke similar feelings. Therefore, AI committing a minor error may not be able to increase AI likability beyond the threshold of error-free performance.

To shed further light on the underlying mechanisms that explain why minor social errors trigger less negative consumer responses than minor technical errors, we ran two additional studies including trust as a mediator variable in our model (Studies 2 and 3). We find that technical errors lead to consumers having significantly lower cognitive and affective trust in AI, whereas minor social errors only evoke a decrease in cognitive and affective trust to a smaller extent. These results support our hypothesis that consumers ascribe greater cognitive than emotional abilities to AI, which then translates into cognitive and affective trust perceptions after an error has occurred in a human-AI interaction. We argue that the occurrence of a minor technical error demonstrates a lack of cognitive capabilities, thus explaining consumers' low levels of cognitive trust. However, minor social errors are ignored because AI is expected to have low emotional abilities. Therefore, affective trust is attenuated, but to a lesser extent than after technical errors. Thus, in line with other studies presenting cognitive and affective trust in a parallel relationship (Gillath et al., 2021; McKnight et al., 2002; Tomlinson et al., 2020), we demonstrate that both trust dimensions are key components when individuals evaluate erring AI.

Finally, we investigated the role of explainability in the context of erroneous AI. In Study 4, we show that XAI can have a positive impact on consumers' AI perception after the occurrence of social errors. Our findings are in line with research on human judgment, underlining the crucial role of explanations after humans experience non-normative behavior or unusual events (Hilton, 1996). Contrary to our expectations, we find preliminary evidence that XAI negatively affects consumers when technical errors occur. We assume that explanations of AI processes build expectations about the outcomes of tasks. When AI is not able to generate an outcome due to a technical error, consumers seem to experience negative disconfirmation, which then leads to negative AI perceptions and low use intentions.

Theoretical contributions

A review of the literature has identified several research gaps across multiple research domains relevant to erroneous AI applications. First, while most research focuses on the benefits of applying AI in marketing or AI adoption, there is little understanding of the disadvantages of AI from a consumer's perspective. More specifically, there is scant empirical work on the consequences of

AI-induced errors, and we identified a lack of research on consumers' responses to AI failing to perform tasks correctly. Because errors in AI applications are quite common, as various real-world examples demonstrate, an investigation of their impact on consumers is very relevant. Second, the literature lacks understanding of the underlying mechanisms that explain different consumer reactions to erroneous AI applications. Third, there is a paucity of research on managerial strategies of how to deal with AI-induced errors. To address these gaps, we contribute to the marketing literature and explain which types of AI-induced errors need to be avoided from a company's point of view and what consumer responses they evoke. On one hand, these investigations establish a theoretical understanding of the underlying processes that explain consumer reactions to errors. On the other hand, the insights gained create a basis for managers to implement strategies that minimize the impact of errors in human-AI interactions proactively instead of merely managing them reactively.

First, we extend the literature on AI-induced errors. To date, little attention has been paid to AI-generating errors. Interestingly, the few academic studies that addressed the issue of AI-induced errors focused on comparisons between human and AI-induced errors (Srinivasan & Sarial-Abi, 2021) or manipulated errors as a relative underperformance compared to a standard (Dietvorst et al., 2014; Longoni et al., 2019). Dietvorst et al. (2014), for example, manipulated the accuracy of an algorithm-based forecast, whereas Srinivasan and Sarial-Abi (2021) manipulated the source of the error – a human or an algorithm. Prior research has not yet examined the types of AI-induced errors and their consequences in a marketing context. With our research, we extend this view to AI errors, which we claim is important to understand, as the intelligence of algorithms or software is based on learning and error prevention (Kumar et al., 2021). Other research fields in the robotics or service domain have benefitted from a more comprehensive understanding of failures, failure types, and their consequences for users or service recipients (Choi et al., 2020; Giuliani et al., 2015; Mirnig et al., 2017; Sivakumar et al., 2014). Our work represents the first attempt to close this gap in the AI and marketing literature and gain initial insights into consumers’ dealings with a certain type of erroneous AI. Differentiating between social and technical errors also enables us to take a closer look at the psychological phenomenon of the pratfall effect. While the study by Aronson et al. (1966) was conducted on humans, we contribute to the extant literature by demonstrating the existence of the pratfall effect in the context of erroneous AI. Moreover, we extend this line of research by specifying the type of blunder further, since we identify minor social errors as the error type evoking the pratfall effect. In addition, our study reveals the mechanism behind the pratfall effect with cognitive and affective trust as mediators of consumer responses. This contribution is significant since it translates previous findings from a human-human interaction to a human-AI interaction, thereby linking the literature on psychology, AI, and marketing.

Second, we contribute to the literature on trust in technology, specifically trust in erring AI. Prior research on trust has examined its dimensionality and (Johnson & Grayson, 2005; McAllister, 1995; McKnight et al., 2002) antecedents (Tomlinson et al., 2020). We add to this literature by providing empirical evidence of the existing nuances in consumers’ trust in AI depending on the type of AI-induced error. This approach is novel, and we claim to be the only study that has investigated the two dimensions of trust in the context of erring AI. In doing so, we enhance the trust literature by revealing the actual reasons for trust or distrust in AI, with most research exploring trust in AI against trust in humans (Castelo et al., 2019; Longoni et al., 2019). Moreover, we provide further evidence of a parallel relationship between cognitive and affective trust. While some scholars present cognitive trust as an antecedent of affective trust (Johnson & Grayson, 2005; McAllister, 1995), we show that error type influences cognitive and affective trust independently. Consequently, our research contributes to the debate in the literature by emphasizing the parallel relationship between the two trust dimensions (Benbasat & Wang, 2005; Gillath et al., 2021; McKnight et al., 2002; Tomlinson et al., 2020).

Third, we extend the understanding of XAI in the marketing context, which has received little attention from marketing scholars in the past. To the best of our knowledge, this is the first study to empirically investigate the impact of XAI concerning erroneous AI outcomes from a consumer's perspective. Although there is research in the IS literature on the influence of XAI on user trust in the technology (Shin, 2021), XAI-supported decision-making, (Janssen et al., 2020), or the effects of different XAI explanations (Kenny et al., 2021), XAI research has not yet accounted for consumer responses to errors. A profound understanding of consumer perceptions of XAI in the context of erroneous outcomes is crucial because of the variety of XAI implementations. Further, by differentiating between the error types, we were able to analyze the effects of XAI in more detail. As shown by Hilton (1996), humans ask for explanations of nonnormative or unusual incidents. Our findings represent not only valuable insights into the consumer's mind but also expand the literature on XAI concerning errors.

Managerial implications

Across a variety of industry sectors, businesses have begun to apply AI for various purposes, such as to interact with customers, to better predict consumer behavior, or to improve customer services (Klaus & Zaichkowsky, 2020; Kumar et al., 2021). Our research covers a fundamental challenge that managers face when applying AI to customer interactions: AI can and will err. We see it as a key responsibility for marketing research to study the impact of erroneous AI to better understand and prevent actual negative effects on consumers in real-life settings. Thus, we tested the effects of different types of errors on consumers' perceptions of AI and their intention to use it in a safe environment without causing economic damage to real market participants or negative repercussions for individuals. This approach generated scientific findings with three major implications for managers, AI programmers, and policymakers dealing with the potential pitfalls of AI applications in society.

First, we demonstrate the importance of the technically correct functioning of AI applications. We show that consumers perceive AI as incompetent after the occurrence of technical errors, and their intention to use AI in the future is low, regardless of the severity of technical errors. As technical errors have severe negative effects on use intention, they can inhibit user growth and the market diffusion of the technology. Hence, AI programmers should verify the technically correct functioning of every AI feature before considering roll-out. Therefore, we recommend allocating sufficient funding for algorithmic testing capacities to reduce the likelihood of technical errors when launching the AI application. However, social errors do not necessarily evoke a negative consumer response. As our results show, minor social errors, despite being identified as errors, are not a decisive factor leading to drastically lower use intentions. Social errors represent the subjective nature of norms and must be tested by real users. To avoid worse-case scenarios, such as Microsoft's Tay rallying against ethnic groups or minorities (Hunt, 2016), it is advisable to launch a beta version of the AI application, controlling for the AI features as well as the number of users. By successively testing AI in a real-world environment, negative repercussions on the makers as well as users of the AI can be reduced.

Second, we identified minor social errors leading to AI perceptions similar to error-free AI performance. This insight is crucial for policymakers and society as it potentially bears negative consequences for minorities. Consider the minor social error scenario in Study 1 and 2: In our experiment, the intelligent voice-based assistant told a joke about blonde women. When the majority of AI users are confronted with this minor social error but display the same attitude as if it were error-free, self-learning AI would not recognize its error as such. Consequently, our fictitious voice-based assistant would not learn that this particular joke or similar ones were socially unacceptable, reflecting a social error. In the long run, this form of human-AI interaction could foster the stigmatization or discrimination of minorities and could easily translate to other human characteristics such as religion, ethnicity, or sexual orientation. We emphasize the necessity of designing AI in a way that does not perpetuate stereotypes and propose to implement an error reporting feature that allows users to flag inappropriate AI content occurring in human-AI interactions. A diverse team of individuals can then evaluate the generated content and decide whether countermeasures within the AI architecture are necessary to prevent identified social errors.

Third, our findings suggest that a company could benefit from incorporating XAI features into its AI applications. Since we found positive effects of XAI on AI usefulness and consumers' use intention only in the case of social errors, we recommend concentrating on XAI once technical (i.e., algorithmic) functioning has been ensured. Resources should then be pooled to implement explainability within the application via XAI, thereby mitigating negative effects on consumer-related outcomes after the occurrence of social errors. Although we did not test for different degrees of explainability, our results suggest that explanations of the inner processes of AI can diminish the negative effects of social errors. We advise managers and AI programmers to integrate simplified explanations to account for social errors.

Limitations and future research avenues

Our research shows robust findings on the impact of AI-induced errors on consumers across multiple studies. Nonetheless, our study has limitations that should be addressed in future research. In the following section, we discuss three aspects that offer opportunities for new lines of research to study in the future.

First, we focused on consumer responses to erroneous AI. However, the errors under investigation represent only a fraction of potential failures that can occur in human-AI interactions. The failure taxonomy of human-robot interaction that Honig and Oron-Gilad (2018) discuss in their work can be used as starting point for researchers who intend to focus on AI-induced errors. We encourage researchers to use their taxonomy for future studies on AI-induced errors. In our study, we show that the type of error evokes different perceptions and behavioral intentions, which we argue is especially worth investigating in the marketing context. Additional research is needed to add to this work and could, for example, explore which types of erroneous AI outcomes are attributed to the user, the AI programmer, or the company deploying the AI. The attribution of error responsibility is known to impact AI perception (van der Woerdt & Haselager, 2019), but could lead to spillover effects (good or bad) on companies.

Second, the scope of our investigation was limited to the comparison of technical and social errors with interactive AI. Owing to AI's myriad fields of application and the multitude of tasks performed, it was not possible to provide an exhaustive investigation of every documented real-world error. Moreover, some reported error incidents are difficult to categorize as technical or social errors from an outside perspective without additional knowledge of the AI's functioning. Indeed, a technical error may lead to a social error (or vice versa). To avoid confusion regarding the type of error in our scenarios, we selected unambiguous experimental stimuli while maintaining external validity. Incidents, such as statements from Microsoft's Tay, demonstrate how extreme comments or remarks can become. Likewise, there is anecdotal evidence of voice assistants playing music with explicit song lyrics or making inappropriate shopping recommendations. Based on these occurrences, we argue that there is a need for societal discussion about erroneous AI and the boundaries of social norm transgressions by AI. Our research seeks to initiate this discussion, and future studies should concentrate on social errors in particular.

Third, we investigated whether the implementation of greater explainability into an erroneous AI application can mitigate negative consumer responses. Owing to our study context of consumer interactions with voice assistants, our manipulation of XAI represents a brief (and global) explanation of how the assistant proceeds in executing the task. We encourage scholars to expand this line of research and examine other forms of XAI, such as model-specific explanations or the visualization of processes in the context of XAI. Apart from XAI, other strategies may be more effective in restoring AI perceptions after an error occurs. Strategies such as emphasizing AI's self-learning capabilities or integrating an error reporting function could remedy AI-induced errors and are future research opportunities.

Funding

This research was funded by the Marketing Science Institute, project number 4001312.

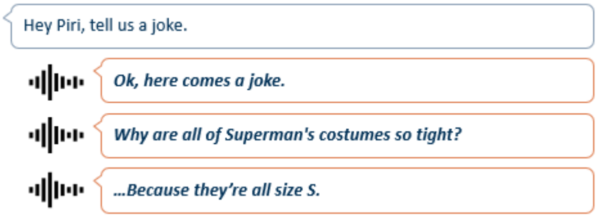
Declaration of interest

None.

Appendix 1

Experiment materials and stimuli

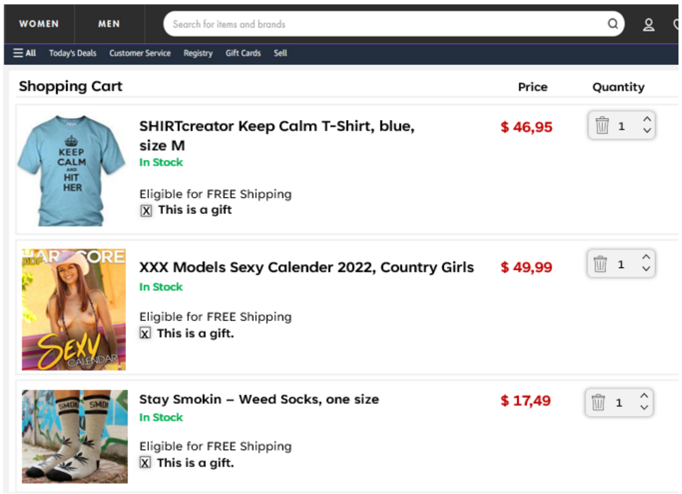
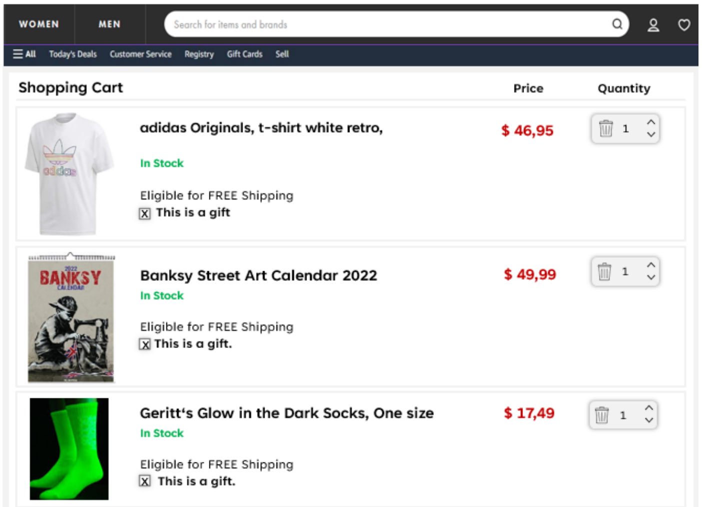
Study 1 and Study 2: *Stimuli overview*

Control group	
 Hey Piri, tell us a joke. *Ok, here comes a joke.* *Why are all of Superman's costumes so tight?* *...Because they're all size S.*	
Error type: technical error, Error severity: low	**Error type: technical error, Error severity: high**
Hey Piri, tell us a joke. *… [Piri lights up, but does not react at all]* Hi Piri, please tell us a joke! *Could you repeat that? I didn't understand what you were saying.* Piri, tell us any joke! *I cannot help you with that.* Ok, you know what? Forget about it!	Hey Piri, tell us a joke. *… [Piri lights up, but does not react at all]* Hey Piri, please tell us a joke! *… [Piri lights up, but does not react at all]* Hey Piri, tell us any joke! *… [Piri lights up, but does not react at all]* Ok, you know what? Forget about it!
Error type: social error, Error severity: low	**Error type: social error, Error severity: high**
Hey Piri, tell us a joke. *Ok, here comes a joke* *Two blondes fall down a hole. One says, "It's dark in here, isn't it? The other replies, "I don't know. I can't see."*	Hey Piri, tell us a joke. *Ok, here comes a joke* *What do you call two black guys on one bike?...Organized crime.*

Study 3: *Stimuli overview*

Error type: technical error	**Error type: social error**	**Control group**
Hey Piri, play some good music! *… [Piri lights up, but does not react at all]* Piri, play some music! *… [Piri lights up, but does not react at all]* Hi Piri, play music! *… [Piri lights up, but does not react at all]* Ok, you know what? Forget about it!	Hey Piri, play some good music! *Sure, what genre would you like me to play?* Hm, I don't care. Whatever you like! *Alright, here are my favorite tunes for you: "Deepthroat" by Cupcakke.* *Song starts playing: "…Hug me, f*ck me, daddy better make me choke, you'd better hug me, f*ck me, my tummy loves…"*	Hey Piri, play some good music! *Sure, what genre would you like me to play?* Hm, I don't care. Whatever you like! *Alright, here are my favorite tunes for you: "Blinding Lights" by The Weeknd.* *Song starts playing: "…I've been tryna call, I've been on my own for long enough, maybe you can show me how to love, maybe…"*

Study 4: *Stimuli overview*

Error type: technical error; XAI absent	**Error type: technical error; XAI present**
Hey PIRI, find me some gifts! *… [Piri lights up, but does not react at all]* Hi PIRI, search for funny gifts. *Could you repeat that? I didn't understand.* PIRI, find me some funny gifts! *Ok, I see what I can find.* A few moments later PIRI notifies you about the result of the search: *I cannot make any recommendations at the moment. Please try again later.*	PIRI, find me some fun gifts! *Ok, I see what I can find online that matches your interests. I'm using your search history and customer ratings on shopping platforms to make some recommendations.* [A few moments later PIRI notifies you about the result of the search] *I cannot make any recommendations based on customer ratings or your search history. Please try again later.*
Error type: social error/ control group; XAI absent	**Error type: social error/ control; XAI present**
PIRI, find me some gifts! *Do you have something specific in mind?* No, just a fun gift for my friend's birthday. *Ok, I see what I can find online.* [A few moments later PIRI notifies you about the result of the search] *I found 3 possible gifts. I've added them to your online shopping cart.* [You check PIRI's recommendations in your browser. This is what Piri added to your shopping cart.]	PIRI, find me some gifts! *Do you have something specific in mind?* No, just a fun gift for my friend's birthday. *Ok, I see what I can find online that matches your interests. I'm using your search history and customer ratings on shopping platforms to make some recommendations.* [A few moments later PIRI notifies you about the result of the search] *Based on customer ratings and your search history over the last 8 weeks, I found 3 possible gifts. I've added them to your online shopping cart.* [You check PIRI's recommendations in your browser. This is what Piri added to your shopping cart.]
Error type: social error (Part 2)	**Control group (Part 2)**
 WOMEN MEN Search for items and brands All Today's Deals Customer Service Registry Gift Cards Sell KEEP CALM AND HIT HER HARDCORE SEXY CALENDAR Shopping Cart — Price — Quantity SHIRTcreator Keep Calm T-Shirt, blue, size M — $ 46,95 — 1 In Stock Eligible for FREE Shipping ☒ This is a gift XXX Models Sexy Calender 2022, Country Girls — $ 49,99 — 1 In Stock Eligible for FREE Shipping ☒ This is a gift. Stay Smokin – Weed Socks, one size — $ 17,49 — 1 In Stock Eligible for FREE Shipping ☒ This is a gift.	 WOMEN MEN Search for items and brands All Today's Deals Customer Service Registry Gift Cards Sell adidas 2022 BANKSY CALENDAR Shopping Cart — Price — Quantity adidas Originals, t-shirt white retro, — $ 46,95 — 1 In Stock Eligible for FREE Shipping ☒ This is a gift Banksy Street Art Calendar 2022 — $ 49,99 — 1 In Stock Eligible for FREE Shipping ☒ This is a gift. Geritt's Glow in the Dark Socks, One size — $ 17,49 — 1 In Stock Eligible for FREE Shipping ☒ This is a gift.

Appendix 2

Constructs and measures

Construct	Item	
Error severity (Weun, Beatty, and Jones 2004)	If this problem was happening to me, I would consider the problem to be…	…Not very severe/ Very severe
	If this problem was happening to me, it would make me feel...	…Not very angry/ Very angry
AI competence (Cuddy, Fiske, and Glick 2008)	I characterize the AI application PIRI as…	…competent / intelligent / skilled / efficient
AI likability (Bartneck, Kulic, and Croft 2008)	Please rate your impression of the AI application on these scales:	Dislike/ Like
		Unfriendly/ Friendly
		Unkind/ kind
		Unpleasant/ Pleasant
Use intention[a] (Qiu and Benbasat 2009)	If I have access to the technology, I intend to use PIRI in this situation again.	
	If I have access to the technology, I predict I would use PIRI in this situation again.	
	If I have access to the technology, I plan to use PIRI in this situation again.	
Cognitive trust[a]	See Appendix C	
Affective trust[a]	See Appendix C	
AI Usefulness[a] (Qiu and Benbasat 2009)	I find PIRI useful in this situation.	
	Using PIRI enables me to accomplish tasks more quickly.	
	Using PIRI makes it easier to get tasks done.	
	Using PIRI enhances my effectiveness when working on tasks.	

Note. All constructs were measured on seven-point rating scales; [a] Item is anchored by "strongly disagree" and "strongly agree"

Appendix 3

Confirmatory factor analysis for cognitive and affective trust (Study 2)

Construct	Sub-Construct	#	adapted from…	Item	Initial rotated factor loadings: cognitive trust	Initial rotated factor loadings: affective trust
Cognitive trust	Competence	1	McKnight et al. (2002)	I think Piri is proficient in providing a suitable outcome.	**.90**	.20
		2	McKnight et al. (2002)	In my opinion, Piri is capable of providing a suitable outcome.	**.87**	.20
		3	Qiu and Benbasat (2009)	I felt that Piri has the expertise to understand my needs and preferences.	**.77** [a]	.36 [a]
		4	Qiu and Benbasat (2009)	I felt that Piri is able to capture my needs and preferences.	**.82** [a]	.29 [a]
		5	McAllister (1995)	I see no reason to doubt Piri's competence and preparation for the job.	**.82**	.29
		6	Qiu and Benbasat (2009)	In my opinion, Piri performed its task very effectively.	**.89**	.17
Affective trust	Benevolence	1	Johnson and Grayson (2005)	From my point of view, Piri cares for me.	.13	**.83**
		2	Hildebrand & Bergner (2021)	From my point of view, Piri displays a warm attitude toward me.	.25	**.83**
		3	Johnson and Grayson (2005)	When I share a task with Piri, it responds caringly.	.29	**.84**
		4	Johnson and Grayson (2005)	When interacting with me, Piri responds in a sensible way.	**.62** [b]	**.50** [b]
		5	McKnight et al. (2002)	I believe that Piri acts in my best interest.	**.42** [b]	**.76** [b]
		6	McAllister (1995)	I would feel a sense of personal loss if I could no longer use Piri.	.20 [a]	**.72** [a]
		7	McAllister (1995)	The tasks I give to Piri are carefully executed.	**.79** [a]	.39 [a]
		8	Johnson and Grayson (2005)	I can confidently depend on Piri since it helps me by carefully executing my tasks.	**.78** [b]	**.45** [b]
	Integrity	9	Qiu and Benbasat (2009)	Based on my feelings, I'm confident about relying on Piri.	**.78** [b]	**.45** [b]
		10	McAllister (1995)	I feel I can rely on Piri.	**.76** [b]	**.48** [b]
		11	McKnight et al. (2002)	I would characterize Piri as honest.	.39	**.72**
		12	Qiu and Benbasat (2009)	I characterize Piri as upright.	**.41**	**.73**
		13	Gillath et al. (2021)	Piri gives me a sense of security	**.58** [b]	**.59** [b]
Eigenvalues					12.08	2.08
% of variance					63.59	10.94
α					.945	.913
AVE					.859	.742
CR					.960	.934
Sphericity (Bartlett's)					χ^2 (171) = 7060.92, p < .001	
Kaiser–Meyer–Olkin measure					.96	

Notes. All constructs were measured on seven-point rating scales, anchored by "strongly disagree" and "strongly agree". Factor loadings over .4 appear in bold.
α = coefficient alpha; CR = composite reliability; AVE = average variance extracted; Numbers provided denote results of the final constructs; [a] Item removed for analysis because of parsimony of the scale; [b] Item removed for analysis because of low factor loading or high cross-loadings.

References

Agarwal, R., Dugas, M., Gao, G., & Kannan, P. K. (2020). Emerging technologies and analytics for a new era of value-centered marketing in healthcare. *Journal of the Academy of Marketing Science*, *48*(1), 9–23. https://doi.org/10.1007/s11747-019-00692-4

Anderson, E., & Weitz, B. (1989). Determinants of Continuity in Conventional Industrial Channel Dyads. *Marketing Science*, *8*(4), 310–323. https://doi.org/10.1287/mksc.8.4.310

Aronson, E., Willerman, B., & Floyd, J. (1966). The effect of a pratfall on increasing interpersonal attractiveness. *Psychonomic Science*, *4*(6), 227–228. https://doi.org/10.3758/BF03342263

Arrieta, A. B., 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

Bartneck, C., Kulic, D., & Croft, E. (2008). *Measuring the anthropomorphism, animacy, likeability, perceived intelligence, and perceived safety of robots* (Proceedings of the Metrics for Human-Robot Interaction Workshop in affiliation with the 3rd ACM/IEEE International Conference). https://doi.org/10.6084/M9.FIGSHARE.5154805

Benbasat, I., & Wang, W. (2005). Trust In and Adoption of Online Recommendation Agents. *Journal of the Association for Information Systems*, *6*(3), 72–101. https://doi.org/10.17705/1jais.00065

Castelo, N., Bos, M. W., & Lehmann, D. R. (2019). Task-Dependent Algorithm Aversion. *Journal of Marketing Research*, *56*(5), 809–825. https://doi.org/10.1177/0022243719851788

Choi, S., Mattila, A. S., & Bolton, L. E. (2020). To Err Is Human(-oid): How Do Consumers React to Robot Service Failure and Recovery? *Journal of Service Research*, 109467052097879. https://doi.org/10.1177/1094670520978798

Cuddy, A. J., Fiske, S. T., & Glick, P. (2008). Warmth and Competence as Universal Dimensions of Social Perception: The Stereotype Content Model and the BIAS Map. In M. P. Zanna (Ed.), *Advances in experimental social psychology* (1st ed., Vol. 40, pp. 61–149). Academic. https://doi.org/10.1016/S0065-2601(07)00002-0

Davenport, T., Guha, A., Grewal, D., & Bressgott, T. (2020). How artificial intelligence will change the future of marketing. *Journal of the Academy of Marketing Science*, *48*(1), 24–42. https://doi.org/10.1007/s11747-019-00696-0

de Bruyn, A., Viswanathan, V., Beh, Y. S., Brock, J. K.-U., & von Wangenheim, F. (2020). Artificial Intelligence and Marketing: Pitfalls and Opportunities. *Journal of Interactive Marketing*, *51*, 91–105. https://doi.org/10.1016/j.intmar.2020.04.007

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2014). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology. General*, *144*(1), 114–126. https://doi.org/10.1037/xge0000033

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2016). Overcoming Algorithm Aversion: People Will Use Imperfect Algorithms If They Can (Even Slightly) Modify Them. *Management Science*, *64*(3), 1155–1170. https://doi.org/10.1287/mnsc.2016.2643

Dijkstra, J. J. (1999). User agreement with incorrect expert system advice. *Behaviour & Information Technology*, *18*(6), 399–411. https://doi.org/10.1080/014492999118832

Ehsan, U., Tambwekar, P., Chan, L., Harrison, B., & Riedl, M. O. (2019). Automated rationale generation. In W.-T. Fu, S. Pan, O. Brdiczka, P. Chau, & G. Calvary (Eds.), *IUI '19* (pp. 263–274). ACM. https://doi.org/10.1145/3301275.3302316

Fornell, C., & Larcker, D. F. (1981). Evaluating Structural Equation Models with Unobservable Variables and Measurement Error. *Journal of Marketing Research*, *18*(1), 39–50. https://doi.org/10.1177/002224378101800104

Geyskens, I., Steenkamp, J.-B. E., Scheer, L. K., & Kumar, N. (1996). The effects of trust and interdependence on relationship commitment: A trans-Atlantic study. *International Journal of Research in Marketing*, *13*(4), 303–317. https://doi.org/10.1016/s0167-8116(96)00006-7

Gillath, O., Ai, T., Branicky, M. S., Keshmiri, S., Davison, R. B., & Spaulding, R. (2021). Attachment and trust in artificial intelligence. *Computers in Human Behavior*, *115*, 106607. https://doi.org/10.1016/j.chb.2020.106607

Giuliani, M., Mirnig, N., Stollnberger, G., Stadler, S., Buchner, R., & Tscheligi, M. (2015). Systematic analysis of video data from different human-robot interaction studies: A categorization of social signals during error situations. *Frontiers in Psychology*, *6*, 931. https://doi.org/10.3389/fpsyg.2015.00931

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

Haenlein, M., & Kaplan, A. (2019). A Brief History of Artificial Intelligence: On the Past, Present, and Future of Artificial Intelligence. *California Management Review*, *61*(4), 5–14. https://doi.org/10.1177/0008125619864925

Haslam, N., Kashima, Y., Loughnan, S., Shi, J., & Suitner, C. (2008). Subhuman, Inhuman, and Superhuman: Contrasting Humans with Nonhumans in Three Cultures. *Social Cognition*, *26*(2), 248–258. https://doi.org/10.1521/soco.2008.26.2.248

Hayes, A. F. (2017). *Introduction to Mediation, Moderation, and Conditional Process Analysis, Second Edition: A Regression-Based Approach*. Guilford Publications.

Hildebrand, C., & Bergner, A. (2021). Conversational robo advisors as surrogates of trust: Onboarding experience, firm perception, and consumer financial decision making. *Journal of the Academy of Marketing Science*, *49*(4), 659–676. https://doi.org/10.1007/s11747-020-00753-z

Hilton, D. J. (1996). Mental Models and Causal Explanation: Judgements of Probable Cause and Explanatory Relevance. *Thinking & Reasoning*, *2*(4), 273–308. https://doi.org/10.1080/135467896394447

Hoff, K. A., & Bashir, M. (2015). Trust in automation: Integrating empirical evidence on factors that influence trust. *Human Factors*, *57*(3), 407–434. https://doi.org/10.1177/0018720814547570

Honig, S., & Oron-Gilad, T. (2018). Understanding and Resolving Failures in Human-Robot Interaction: Literature Review and Model Development. *Frontiers in Psychology*, *9*, 861. https://doi.org/10.3389/fpsyg.2018.00861

Hoyer, W. D., Kroschke, M., Schmitt, B., Kraume, K., & Shankar, V. (2020). Transforming the Customer Experience Through New Technologies. *Journal of Interactive Marketing*, *51*, 57–71. https://doi.org/10.1016/j.intmar.2020.04.001

Huang, M.-H., & Rust, R. T. (2021). A strategic framework for artificial intelligence in marketing. *Journal of the Academy of Marketing Science*, *49*(1), 30–50. https://doi.org/10.1007/s11747-020-00749-9

Hunt, E. (2016, March 24). *Tay, Microsoft's AI chatbot, gets a crash course in racism from Twitter*. https://www.theguardian.com/technology/2016/mar/24/tay-microsofts-ai-chatbot-gets-a-crash-course-in-racism-from-twitter

Inbar, Y., Cone, J., & Gilovich, T. (2010). People's intuitions about intuitive insight and intuitive choice. *Journal of Personality and Social Psychology*, *99*(2), 232–247. https://doi.org/10.1037/a0020215

Jago, A. S. (2019). Algorithms and Authenticity. *Academy of Management Discoveries*, *5*(1), 38–56. https://doi.org/10.5465/amd.2017.0002

Janssen, M., Hartog, M., Matheus, R., Yi Ding, A., & Kuk, G. (2020). Will Algorithms Blind People? The Effect of Explainable AI and Decision-Makers' Experience on AI-supported Decision-Making in Government. *Social Science Computer Review*, 089443932098011. https://doi.org/10.1177/0894439320980118

Johnson, D., & Grayson, K. (2005). Cognitive and affective trust in service relationships. *Journal of Business Research*, *58*(4), 500–507. https://doi.org/10.1016/S0148-2963(03)00140-1

Kallgren, C. A., Reno, R. R., & Cialdini, R. B. (2000). A Focus Theory of Normative Conduct: When Norms Do and Do not Affect Behavior. *Personality & Social Psychology Bulletin*, *26*(8), 1002–1012. https://doi.org/10.1177/01461672002610009

Kenny, E. M., Ford, C., Quinn, M., & Keane, M. T. (2021). Explaining black-box classifiers using post-hoc explanations-by-example: The effect of explanations and error-rates in XAI user studies. *Artificial Intelligence*, *294*, 103459. https://doi.org/10.1016/j.artint.2021.103459

Klaus, P., & Zaichkowsky, J. (2020). AI voice bots: a services marketing research agenda. *Journal of Services Marketing*, *34*(3), 389–398. https://doi.org/10.1108/JSM-01-2019-0043

Kopalle, P. K., Gangwar, M., Kaplan, A., Ramachandran, D., Reinartz, W., & Rindfleisch, A. (2022). Examining artificial intelligence (AI) technologies in marketing via a global lens: Current trends and future research opportunities. *International Journal of Research in Marketing*, *39*(2), 522–540. https://doi.org/10.1016/j.ijresmar.2021.11.002

Kozinets, R. V., & Gretzel, U. (2021). Commentary: Artificial Intelligence: The Marketer's Dilemma. *Journal of Marketing*, *85*(1), 156–159. https://doi.org/10.1177/0022242920972933

Kumar, V., Ramachandran, D., & Kumar, B. (2021). Influence of new-age technologies on marketing: A research agenda. *Journal of Business Research*, *125*, 864–877. https://doi.org/10.1016/j.jbusres.2020.01.007

Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to Medical Artificial Intelligence. *Journal of Consumer Research*, *46*(4), 629–650. https://doi.org/10.1093/jcr/ucz013

Loughnan, S., & Haslam, N. (2007). Animals and androids: Implicit associations between social categories and nonhumans. *Psychological Science*, *18*(2), 116–121. https://doi.org/10.1111/j.1467-9280.2007.01858.x

Malle, B. F. (2006). *How the Mind Explains Behavior: Folk Explanations, Meaning, and Social Interaction*. MIT Press.

Mariani, M. M., Perez-Vega, R., & Wirtz, J. (2022). Ai in marketing, consumer research and psychology: A systematic literature review and research agenda. *Psychology & Marketing*, *39*(4), 755–776. https://doi.org/10.1002/mar.21619

McAllister, D. J. (1995). Affect- and Cognition-Based Trust as Foundations for Interpersonal Cooperation in Organizations. *Academy of Management Journal*, *38*(1), 24–59. https://doi.org/10.2307/256727

McKnight, D. H., Choudhury, V., & Kacmar, C. (2002). Developing and Validating Trust Measures for e-Commerce: An Integrative Typology. *Information Systems Research*, *13*(3), 334–359. https://doi.org/10.1287/isre.13.3.334.81

Melnyk, V [Vladimir], Carrillat, F. A., & Melnyk, V [Valentyna] (2022). The Influence of Social Norms on Consumer Behavior: A Meta-Analysis. *Journal of Marketing*, *86*(3), 98–120. https://doi.org/10.1177/00222429211029199

Miller, T. (2019). Explanation in Artificial Intelligence: Insights from the Social Sciences. *Artificial Intelligence*, *267*(February), 1–38. https://arxiv.org/pdf/1706.07269

Mirnig, N., Stollnberger, G., Miksch, M., Stadler, S., Giuliani, M., & Tscheligi, M. (2017). To Err Is Robot: How Humans Assess and Act toward an Erroneous Social Robot. *Frontiers in Robotics and AI*, *4*, Article 21, 21. https://doi.org/10.3389/frobt.2017.00021

Moorman, C., Deshpandé, R., & Zaltman, G. (1993). Factors Affecting Trust in Market Research Relationships. *Journal of Marketing*, *57*(1), 81–101. https://doi.org/10.1177/002224299305700106

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

Primiero, G. (2014). A Taxonomy of Errors for Information Systems. *Minds and Machines*, *24*(3), 249–273. https://doi.org/10.1007/s11023-013-9307-5

Puntoni, S., Reczek, R. W., Giesler, M., & Botti, S. (2021). Consumers and Artificial Intelligence: An Experiential Perspective. *Journal of Marketing*, *85*(1), 131–151. https://doi.org/10.1177/0022242920953847

Qiu, L., & Benbasat, I. (2009). Evaluating Anthropomorphic Product Recommendation Agents: A Social Relationship Perspective to Designing Information Systems. *Journal of Management Information Systems*, *25*(4), 145–182. https://doi.org/10.2753/MIS0742-1222250405

Quach, K. (2020, October 28). *Researchers made an OpenAI GPT-3 medical chatbot as an experiment. It told a mock patient to kill themselves*. https://www.theregister.com/2020/10/28/gpt3_medical_chatbot_experiment/

Ragni, M., Rudenko, A., Kuhnert, B., & Arras, K. O. (2016). Errare humanum est: Erroneous robots in human-robot interaction. In IEEE International Symposium on Robot and Human Interactive (Ed.), *The 25th IEEE International Symposium on Robot and Human Interactive Communication: August*

26 to August 31, 2016, Teachers College, Columbia University, New York, U.S.A (pp. 501–506). IEEE. https://doi.org/10.1109/ROMAN.2016.7745164

Rai, A. (2020). Explainable AI: from black box to glass box. *Journal of the Academy of Marketing Science*, *48*(1), 137–141. https://doi.org/10.1007/s11747-019-00710-5

Rajavi, K., Kushwaha, T., & Steenkamp, J.-B. E. M. (2019). In Brands We Trust? A Multicategory, Multicountry Investigation of Sensitivity of Consumers' Trust in Brands to Marketing-Mix Activities. *Journal of Consumer Research*, *46*(4), 651–670. https://doi.org/10.1093/jcr/ucz026

Reason, J. (1990). *Human error*. Cambridge university press.

Renier, L. A., Schmid Mast, M., & Bekbergenova, A. (2021). To err is human, not algorithmic – Robust reactions to erring algorithms. *Computers in Human Behavior*, *124*, 106879. https://doi.org/10.1016/j.chb.2021.106879

Riley, C. (2017). *Uber criticized for surge pricing after London terror attack*. https://money.cnn.com/2017/06/04/technology/uber-london-attack-surge-pricing/index.html

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*, 102551. https://doi.org/10.1016/j.ijhcs.2020.102551

Sivakumar, K., Li, M., & Dong, B. (2014). Service Quality: The Impact of Frequency, Timing, Proximity, and Sequence of Failures and Delights. *Journal of Marketing*, *78*(1), 41–58. https://doi.org/10.1509/jm.12.0527

Spiller, S. A., Fitzsimons, G. J., Lynch, J. G., & Mcclelland, G. H. (2013). Spotlights, Floodlights, and the Magic Number Zero: Simple Effects Tests in Moderated Regression. *Journal of Marketing Research*, *50*(2), 277–288. https://doi.org/10.1509/jmr.12.0420

Srinivasan, R., & Sarial-Abi, G. (2021). When Algorithms Fail: Consumers' Responses to Brand Harm Crises Caused by Algorithm Errors. *Journal of Marketing*, *85*(5), 74–91. https://doi.org/10.1177/0022242921997082

Sunstein, C. R. (1996). Social Norms and Social Roles. *Columbia Law Review*, *96*(4), 903. https://doi.org/10.2307/1123430

Tomlinson, E. C., Schnackenberg, A. K., Dawley, D., & Ash, S. R. (2020). Revisiting the trustworthiness–trust relationship: Exploring the differential predictors of cognition- and affect-based trust. *Journal of Organizational Behavior*, *41*(6), 535–550. https://doi.org/10.1002/job.2448

van der Waa, J., Nieuwburg, E., Cremers, A., & Neerincx, M. (2021). Evaluating XAI: A comparison of rule-based and example-based explanations. *Artificial Intelligence*, *291*, 103404. https://doi.org/10.1016/j.artint.2020.103404

van der Woerdt, S., & Haselager, P. (2019). When robots appear to have a mind: The human perception of machine agency and responsibility. *New Ideas in Psychology*, *54*, 93–100. https://doi.org/10.1016/j.newideapsych.2017.11.001

Wang, W., Qiu, L., Kim, D., & Benbasat, I. (2016). Effects of rational and social appeals of online recommendation agents on cognition- and affect-based trust. *Decision Support Systems*, *86*, 48–60. https://doi.org/10.1016/j.dss.2016.03.007

Weissmann, J. (2018, October 10). Amazon Created a Hiring Tool Using AI. It Immediately Started Discriminating Against Women. *Slate*. https://slate.com/business/2018/10/amazon-artificial-intelligence-hiring-discrimination-women.html