



Explainable AI: The Effect of Contradictory Decisions and Explanations on Users' Acceptance of AI Systems

Carolin Ebermann, Matthias Selisky & Stephan Weibelzahl

To cite this article: Carolin Ebermann, Matthias Selisky & Stephan Weibelzahl (2023) Explainable AI: The Effect of Contradictory Decisions and Explanations on Users' Acceptance of AI Systems, International Journal of Human-Computer Interaction, 39:9, 1807-1826, DOI: [10.1080/10447318.2022.2126812](https://doi.org/10.1080/10447318.2022.2126812)

To link to this article: <https://doi.org/10.1080/10447318.2022.2126812>



Published online: 14 Oct 2022.



Submit your article to this journal [↗](#)



Article views: 417



View related articles [↗](#)



View Crossmark data [↗](#)



Citing articles: 1 View citing articles [↗](#)



Explainable AI: The Effect of Contradictory Decisions and Explanations on Users' Acceptance of AI Systems

Carolyn Ebermann, Matthias Selisky, and Stephan Weibelzahl

Business Psychology, PFH Private University of Applied Sciences, Göttingen, Germany

ABSTRACT

Providing explanations of an artificial intelligence (AI) system has been suggested as a means to increase users' acceptance during the decision-making process. However, little research has been done to examine the psychological mechanism of how these explanations cause a positive or negative reaction in the user. To address this gap, we investigate the effect on user acceptance if decisions and the associated provided explanations contradict between an AI system and the user. An interdisciplinary research model was derived and validated by an experiment with 78 participants. Findings suggest that in decision situations with cognitive misfit users experience negative mood significantly more often and have a negative evaluation of the AI system's support. Therefore, the following article provides further guidance regarding new interdisciplinary approaches for dealing with human-AI interaction during the decision-making process and sheds some light on how explainable AI can increase users' acceptance of such systems.

1. Introduction

Artificial intelligence (AI) is “the ability of a machine to perform cognitive functions that we associate with human minds, such as perceiving, reasoning, learning, interacting with the environment, problem solving, decision-making, and even demonstrating creativity” (Rai et al., 2019, p. iii). In the near future, deploying AI will be a necessary condition for many companies for maintaining competitiveness (Maedche et al., 2019). AI is perceived by leaders in business, government, academia, and non-government organizations as having the potential for immense benefits (World Economic Forum, 2016). For instance, AI facilitates automating business processes or decision-making through extensive data analysis (Maedche et al., 2019). Therefore, for a majority of companies, the introduction of AI is inevitable and imminent (Ransbotham et al., 2017).

While some stakeholders, e.g., the management levels in organizations and early adopters, often exhibit a positive attitude towards the introduction of AI, an increasing number of individuals are opposed to or even hostile towards AI usage. Some of them are afraid of getting replaced by AI and losing their job, others critique the configuration of the AI (Dietvorst et al., 2018; Janssen et al., 2019; Maedche et al., 2019; Sundar, 2020; Wang & Wang, 2022). When using AI for decision-making a lack of acceptance may arise from an incomplete understanding of AI's underlying operations and the inner-workings (Maedche et al., 2019). In contrast to traditional systems, AI systems have a much higher complexity and show a more dynamic behavior. Often they are constructed as black boxes, i.e., AI systems are

frequently not transparent in regard to which assumptions and decisions the underlying algorithms are making (Pentland et al., 2019). For instance, a self-driving car will consider thousands of parameters when deciding to accelerate or stop. It is difficult if not impossible to explain to the user how the system made the decision.

In this regard, *explainable AI* has been proposed as key to deciphering AI's black box. In reference to Ehsan and Riedl (2020), we define explainable AI as “artificial intelligence and machine learning techniques that can provide human-understandable justification for their output behavior” (p. 1). Besides making algorithm operations explainable, explainable AI approaches have been explored to provide human-interpretable decision processes. Previous studies about explainable AI suggest that explanations given by an AI decision-making system increase user acceptance, satisfaction, experience, and AI adoption (Guidotti et al., 2019;; Selvaraju et al., 2017; Tjoa & Guan, 2021). However, in which cases explanations given by an AI affects users' acceptance remains unclear (Ehsan & Riedl, 2020). To address this gap the present study empirically analyses the following research question: *How do decisions and associated explanations provided by an AI influence user acceptance, in particular when user and system contradict each other?*

To investigate this research question, we conducted an experiment with $N=78$ participants using the Wizard of Oz method. We propose a research model derived from existing theories from both information system research, i.e., cognitive fit theory (Vessey, 1991) and the explanation evaluation framework (Gunning & Aha, 2019), as well as from psychology,

i.e., the cognitive dissonance theory of Festinger (1957). This interdisciplinary research model provides valuable assumptions about how individuals become incapacitated or unruly in the decision-making process in the instance of inappropriate decisions and explanations.

The remainder of this paper is structured as follows. The section “Theoretical background” explains the general idea as well as the current situation of human-AI interaction during the decision-making process. Then we present the state of the art of human-AI interaction research on explainable AI within the decision-making process. We introduce the research model as well as the hypotheses. The next section describes the methodological approach followed by a presentation and discussion of the findings. The article closes by addressing the limitations of the investigation, offering an outlook for further research, and highlighting the study’s practical and theoretical implications.

2. Theoretical background

2.1. AI and decision-making

AI can take a large variety of shapes and forms. The present study focuses on a particular type, namely the application of AI in the decision making process. Decisions, on the most basic level, are “choices from among several options” (Klein, 2015). Decision-making is ingrained in human everyday life. People make choices all the time, consciously or unconsciously. In a few cases an ideal or *normative* choice can be identified, while in other cases, the optimal choice cannot be determined, and decision makers need to rely on preferences as a basis for their decisions (Slovic et al., 1977).

Using information systems for decision making like recommender/recommendation systems has been one of the most important applications in the history of information systems (Bouakkaz et al., 2017; Choi et al., 2016; Chung, 2014; Dugdale, 1996; Kim et al., 2010). However, these *traditional* information systems differ from AI systems because AI has the capability to learn from experience and is therefore able to reproduce aspects of human reasoning. Decision-making in AI systems is based on algorithms like rule-based inference (Araújo & Pestana, 2017; Kao et al., 2017; Rekik et al., 2018), semantic linguistic analysis (Ahmad & Laroche, 2017; Ogiela & Ogiela, 2014), bayesian networks (Ramírez-Noriega et al., 2017; Zhao et al., 2008), similarity measures (Bouakkaz et al., 2017; Ragini et al., 2018; Tseng et al., 2017), neural networks (Frias-Martinez et al., 2006; Yaqoob et al., 2016), frame-based representation (Dugdale, 1996), or genetic algorithms (Lebib et al., 2017). The application of these algorithms has been proven to enhance decision making in various contexts (Batra & Antony, 2001; Pisz & Łapunińska, 2017).

In the decision-making process AI can entirely replace human decision makers or the two can work together in a collaborative approach (Edwards et al., 2000). The ideal of such collaboration, often called human-AI symbiosis, was first articulated by Licklider (1960) as a relationship through which the strengths of one compensate for the limitations of the other. When following this collaborative approach, the

question arises, what are the strengths and limitations of humans and AI compared to each other? Table 1 summarizes the literature on the various superior skills AI and humans have over each other, respectively. There is strong consensus for many skills including creativity and social skills, listed by many different authors as a human strength (McAfee & Brynjolfsson, 2016; Cook & Heshmat, 2019; Harari, 2016; Jarrahi, 2018; Kolbjørnsrud et al., 2016; Rushkoff, 2019; Wilson & Daugherty, 2018). The strengths of AI are predominantly described as its quantitative abilities like performing calculations, simulations, classifications, and optimizations at speed and scale (Agrawal et al., 2016, 2017; McAfee & Brynjolfsson, 2016; Chui et al., 2018; Cook & Heshmat, 2019; Guszczka et al., 2017; Harari, 2016; Jarrahi, 2018; Kolbjørnsrud et al., 2016; Rushkoff, 2019; Wilson & Daugherty, 2018). In the decision-making process, this quantitative analytical power can be exploited for generating fresh ideas through probability and data-driven statistical inference approaches or for identifying relationships among factors in vast data sets (Jarrahi, 2018). Also, by creating simulations at low cost, AI excels in the skill of making predictions, i.e., the ability to take information you have and generate information you do not have (Agrawal et al., 2016; Kolbjørnsrud et al., 2016).

For some skills, such as *judgement*, the consensus is less clear-cut. Although judgement is described as a human strength by some authors (Agrawal et al., 2016; McAfee & Brynjolfsson, 2016; Chui et al., 2018; Cook & Heshmat, 2019; Guszczka et al., 2017; Kolbjørnsrud et al., 2016; Rushkoff, 2019), others argue that it is also a strength of AI, since skills like *classification* (Kolbjørnsrud et al., 2016) and *making recommendations* (Chui et al., 2018) clearly also involve some form of judgement. Jarrahi (2018) gives a more precise description about which sub-skills of judgment are expressed better by humans and AI, respectively:

- To mitigate uncertainty, humans can use their superior intuition to make swift decisions while AI might help by providing access to real-time information (e.g., anomaly detection).
- To reduce complexity, humans are superior at deciding where to seek and gather data and choosing among options with equal data support, while AI can quickly collect, curate, process, and analyze data.
- In the face of equivocality, humans have the skill to negotiate, build consensus, and rally support, while AI has the ability to perform sentiment analysis at scale to represent diverse interpretations.

In sum, guided by human questions, creativity and problem framing, AI can help humans sift through vast volumes of data to uncover patterns and generate options for solutions, which in turn guides humans to reach a conclusion. Lastly, humans can use their social skills to lead, negotiate, build consensus and rally support to ensure proper implementation of the decided solution.

van den Bosch and Bronkhorst (2018) assume three levels of interactivity in order to receive a satisfactory

Table 1. Overview of the specific strengths and skills of AI compared to humans and vice versa.

AI	Humans	Source
Quantitative capabilities (speed, scalability)	Creativity	Wilson & Daugherty, 2018
Quantitative capabilities (computation, analytical capabilities, collection, curation, processing, and analysis of data, scalability)	Social skills (leadership, teamwork) Creativity (intuition, holistic vision, imagination, sensitivity, rumination) Social skills (emotional intelligence, negotiate, build consensus, and rally support) Judgment (making swift intuitive decisions in the face of the unknown; deciding where to seek and gather data. choosing among options with equal data support)	Jarrahi, 2018
Quantitative capabilities (repetitive work, calculations, attention to detail)	Creativity Social Skills (collaboration, empathy) Judgment	Cook & Heshmat, 2019
Quantitative capabilities (classification, continuous estimation, clustering, all other optimization, anomaly detection, ranking, recommendations, data generation)	Judgment (application of context, curating data sets)	Chui et al., 2018
Quantitative capabilities (analytical tasks, simulations, data processing, anomaly detection, scalability)	Creativity (idea development, creative thinking and experimentation) Social Skills (networking, coaching, collaboration, empathy) Judgment (drafting strategy, data interpretation, application of context and history, problem framing, reaching conclusions and taking actions)	Kolbjørnsrud et al., 2016
Quantitative capabilities (predictive analytics)	Judgment	Agrawal et al., 2016, 2017
Quantitative capabilities (data processing, scalability, making predictions, recommendations)	Judgment (conceptual understanding and commonsense reasoning needed to evaluate novel situations)	Guszcza et al., 2017
Quantitative capabilities (pattern recognition within a predefined frame)	Creativity (ideation, innovation) Social Skills (complex communication) Judgment (large-frame pattern recognition)	McAfee & Brynjolfsson 2016
Quantitative capabilities (pattern recognition, analyzing large quantities of data, quickly gathering and updating datasets, scalability)	Social Skills (describing emotional experiences, cooperating in a flexible way with countless numbers of strangers)	Harari, 2016
Quantitative capabilities (speed, efficiency, multitasking)	Creativity (improvisation) Judgment (inferring things based on context, dealing with ambiguity)	Rushkoff, 2019

collaboration between AI and humans. On the first level, the interaction between humans and AI is unidirectional. The AI is a black box and does not provide insights into what assumptions and decisions the underlying algorithms are making (Pentland et al., 2019). Thus, AI's behavior seems sometimes unpredictable to the user and established usability guidelines of user interface design are disregarded (Nielsen & Molich, 1990, as cited in Amershi et al., 2019). On the second level there is a bi-directional interaction between the AI and human and the user receives explanations on demand, that enable a better understanding of the AI. This requires on one hand, that the user asks the AI for an explanation, and on the other hand that the AI has the capability to understand the purpose of the user's demand and is able to develop explanations fitting the user's purpose. Finally, the third level of collaboration is characterized by complete bi-directional interaction. On this level, the AI is also able to ask the user for explanations and provide information voluntarily. Therefore, the AI must analyze such states and generate a comprehensible line of argumentation for the users.

Using the framework of van den Bosch and Bronkhorst (2018) it becomes clear that today most human-AI interactions are unidirectional and thus run on the first level. Therefore, the AI is still a black box and no explanations about the decisions are provided (Gunning & Aha, 2019).

AI systems that provide an explanation focus mostly on the AI (decision) models and algorithms (Janssen, 2019; Tjoa & Guan, 2021). Hence, the explanations remain very abstract and at times far from the reality of the user. Explanations may thus be perceived as being of little help (Maedche et al., 2019; Rai, 2020; Sundar, 2020).

2.2. Human-AI interaction research on explainable AI within the decision-making process

Many researchers assume positive effects of explainable AI on user satisfaction, experience, and acceptance (Gedikli et al., 2014; Zhang & Sundar, 2019). Furthermore, users are more likely to follow the system's advice during the decision-making process when comprehensible explanations were provided (Arnold et al., 2006; Giboney et al., 2015; Ye & Johnson, 1995). However, several scholars assume that if the user perceives the explanation of an AI to be ineffective, it can result in a negative reaction in the user (e.g., reduced user satisfaction and trust), negative performance (e.g., wrong decisions), or reactance against the AI's decision and usage (Gunning & Aha, 2019; Rai, 2020). Besides this research strand, there is also a "large push" from European legislation towards "explainability" and explainable AI. European laws demand that AI decisions have to be explainable to humans (Wachter et al., 2017).

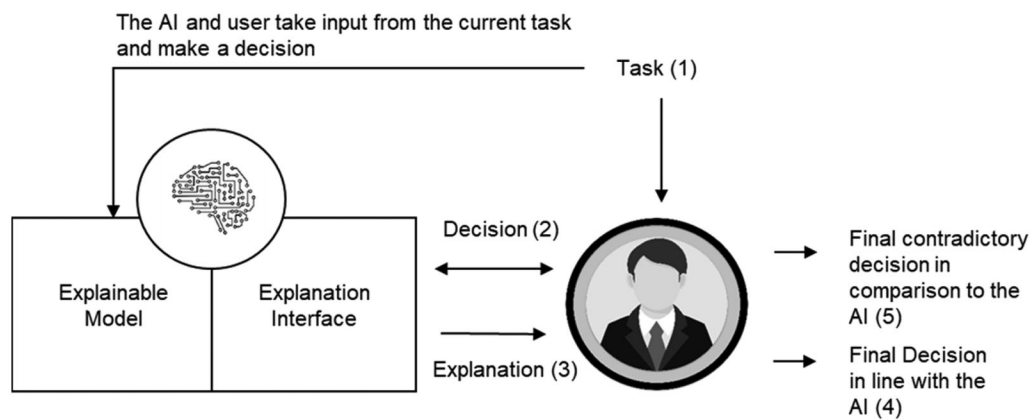


Figure 1. Explanation Evaluation Framework (EEF) adapted from Gunning and Aha (2019).

The effect of AI explanations on user acceptance is a nascent area of research (Vinson et al., 2019). Previous studies about explainable AI are often of conceptual nature, e.g., summarize assumptions about how design can support good human-AI interaction (Amershi et al., 2019) or give a literature overview of explainable AI or human-AI interaction (Adadi & Berrada, 2018; Hoffman et al., 2018; Pieters, 2011; Tintarev & Masthoff, 2012). However, there are only a few empirical studies that evaluate the effect of AI's explanations on real participants (Ehsan & Riedl, 2020; Keane et al., 2021; Tjoa & Guan, 2021). One example is Selvaraju et al. (2017) who found that their approach—a Grad-CAM visualization—can increase users' trust, and help users identify biases in datasets. Throughout the discussion on explainable AI a number of different concerns have frequently been mixed up. Gunning and Aha (2019) provide a helpful model that differentiates between an *explainable model* and *explanations* given to the user (see Figure 1). An *explainable model* can be understood as “a property of machine-learned models that dictates the degree to which a human user—AI expert or non-expert user—can come to conclusions about the performance of the model given specific inputs” (Ehsan & Riedl, 2020; pp. 1, 2). Hohman et al. (2019) call this part of explainability *interpretability* and propose visualizations of AI models to increase the explainability of the AI. In contrast to an *explainable model*, an *explanation* does not describe in detail how the AI (decision) models and algorithms work, but offers human-understandable justifications for the AI output behavior for practitioners and users (Ehsan & Riedl, 2020). Explanations significantly enhance users' attitudes and overall satisfaction with a system (Kizilcec, 2016) and adherence to the system's advice, as well as users' decision-making effectiveness (Rzepka & Berger, 2018). The predominant focus on *explainable models* (interpretability) is very unsatisfactory for practitioners, researchers, as well as users because “when we talk about an explanation for a decision, we generally mean the need for reasons or justifications for that particular outcome, rather than a description of the inner workings or the logic of reasoning behind the decision-making process in general” (Adadi & Berrada, 2018, p. 52142).

Only a few previous studies have applied psychological or social science theories (Adadi & Berrada, 2018; Lipton, 2018;

Shin & Park, 2019). One example is Lee (2018) who shows that workers' perceptions of decision fairness, trustworthiness, and emotional response in algorithmic managers compared to human managers depend significantly on task characteristics. The study indicates that algorithmic decisions are viewed by humans as less fair, less trustworthy, and more likely to elicit negative emotions for tasks that people think require uniquely human skills (Lee, 2018). Byrne (2019) discusses the use of counterfactuals in explanation of AI based on psychological findings. The limited number of studies from a psychological or social science perspective seems problematic because explanations are characterized as a form of communication or social interaction with many psychological, cognitive, and philosophical components (Adadi & Berrada, 2018). Furthermore, researchers of human-AI interaction and social sciences have been working in isolation (Abdul et al., 2018). However, the combination of both seems expedient because researchers of human-AI interaction can apply methods, which enable them to develop a transparent system with context-aware explanations. Consequently, the AI can react to environmental modification, e.g., user profile or setting, and thus, a collaborative human-AI system can arise, where both parties are equally adaptive team members (Adadi & Berrada, 2018).

In summary, there are some studies, which confirm the positive effect of explainable AI. However, most research about human-AI interaction and explainable AI is theoretical in nature, does not investigate the effects of explainable AI on humans empirically, focuses on interpretability, and does not apply theories from social science. To address these gaps, this study focuses on the *explanation* as opposed to the *interpretability*. Therefore, in this article, the term *explainable AI* means that the AI provides the information to increase the user's understanding of the reason for the decision instead of information about how the decision has been processed (Giboney et al., 2015; Mueller et al., 2019). We conducted an experiment on what the effect is on user acceptance if decisions and provided explanations contradict between an AI and the user. We focus on a cognitive psychological perspective without evaluating other factors and suggest that cognitive misfit can lead to reduced acceptance. To test the effect of cognitive misfit in this study, firstly, a

research model was developed based on the *Explanation Evaluation Framework* (EEF; Gunning & Aha, 2019), the *cognitive fit theory* (CFT; Vessey, 1991), and the *cognitive dissonance theory* (CDT; Festinger, 1957). While the EEF explains the interaction between AI and the user with the assumptions of ideal conditions during the decision-making process, the CFT and CDT describe mechanisms and consequences of cognitive misfit.

In reference to the CFT (Vessey, 1991) which was applied in various studies on *traditional* information systems (Dennis & Carte, 1998; Huang et al., 2006) cognitive misfit between the AI and the user can occur due to different problem representations of a task. In reference to Giboney et al. (2015) the problem representation of the user or AI during the decision-making process is defined as a person's or AI's knowledge or understanding of the problem and how to attempt to obtain a solution to the problem of the task. In the case of cognitive fit between the user's problem representation and the problem representation of the system, the user reaches a final decision faster and in a more exact manner (Vessey, 1991). The user can directly internalize the information and apply it to solving the problem of the task (Vessey, 1991).

The reasons for a difference in problem representation and therefore cognitive misfit between the AI and the user could be localized in a mismatch in cognitive styles between the AI and humans (Agustina et al., 2017; Chan, 1996). "Cognitive styles represent the characteristic modes of functioning and predispositions by individuals in the perceptual and thinking behavior in the decision-making process" (Dalal & Kasper, 1994, p. 679). In line with the dominant skills

described above, the cognitive style of the AI can be described as analytic, while the human cognitive style is more intuitive (Dalal & Kasper, 1994; Maedche et al., 2019). The AI is likely to use predominantly analytical approaches by for example generating quantitative models of the situation (Jarrahi, 2018). In contrast to AI, humans often revert to heuristics and are prone to biases, like affect heuristic (Lerner et al., 2015) or loss aversion (Kahneman & Tversky, 1979), because of the limited information or time available, as well as due to complex situations that challenge the user's information-processing capacity (Kahneman & Tversky, 1979).

3. Research model and hypotheses

The Explanation Evaluation Framework (EEF; Gunning & Aha, 2019) provides a conceptual structure of the collaborative decision-making process using explainable AI. An adapted version is illustrated in Figure 1. During the decision-making process, the AI and the user take input from a current task (1) and both parties make a decision in parallel (2). During the AI-human interaction the AI generates an explanation based on its explainable model to justify its decision. This explanation is displayed by the explanation interface to the user (3). Thus, the user can make a final decision which may or may not be in line with the decision and explanation of the AI (4 vs. 5).

Using this framework, the Cognitive Fit Theory (CFT) provides an interesting explanation of user acceptance. The assumptions of the CFT can be easily combined with the EEF (see Figure 2). The AI takes input from the current task (1) in order to make the decision and explain it to the user. The AI's

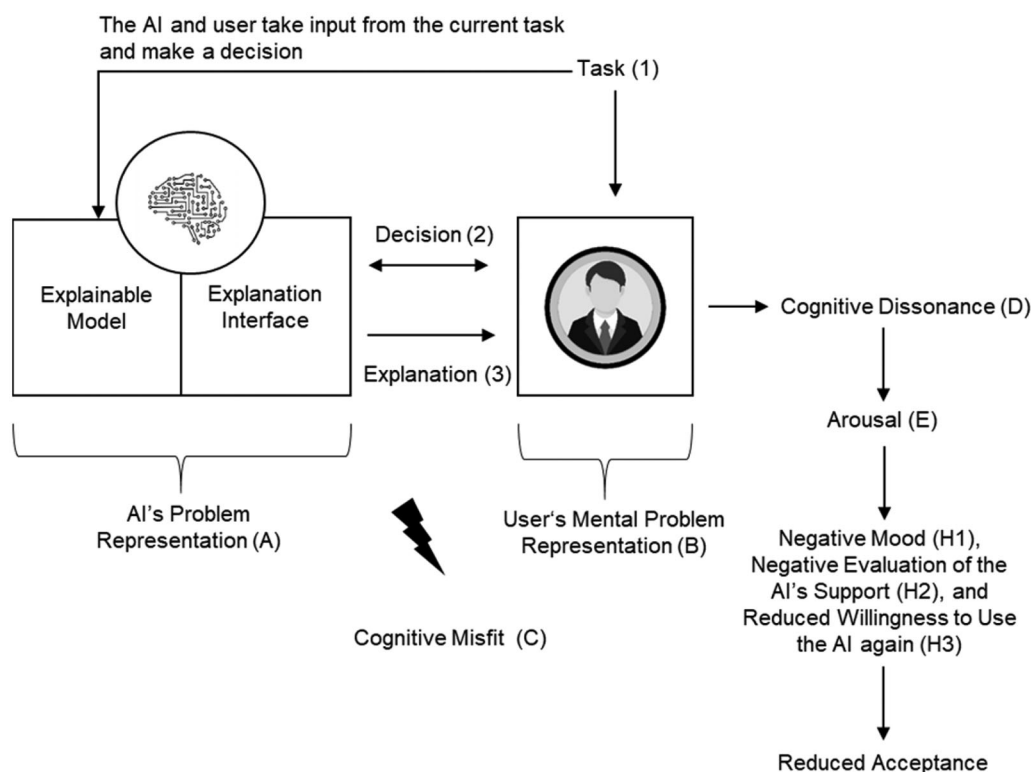


Figure 2. Research model with hypotheses in adaption to the EEF (Gunning & Aha, 2019), the CFT (Vessey, 1991), and CDT (Festinger, 1957).

explainable model signifies the AI's problem representation of the resulting decision and explanation (A). The explanation interface expresses this representation (A) by stating the decision (2) and the corresponding explanation (3) to the user. In parallel, the current task (1) also influences the user's problem representation of the decision and corresponding explanation in the user's mind (B). Therefore, cognitive misfit during the human-AI interaction (C) occurs when the mental problem representation of the user (B) is divergent from the AI's problem representation (A).

Cognitive Misfit Theory establishes that the problem representations of the AI and user may differ. Cognitive Dissonance Theory (CDT) by Festinger (1957) suggests that a difference in representations may result in mental discomfort. Festinger (1957) demonstrated that if an individual perceives inconsistency between his or her cognitions, a negative intrapersonal state, i.e. cognitive dissonance, occurs and causes arousal (Festinger, 1957; Festinger & Carlsmith, 1959). This phenomenon can be applied to our research model. In case of cognitive misfit, the user is confronted with two different cognitions due to AI's divergent decision and corresponding explanation. Therefore, cognitive dissonance (D) in the user can occur and cause arousal (E).

This article hypothesizes that this change in arousal could lead to a negative mood in the user as well as to negative evaluation of the AI's support, and thus, to reduced willingness of using the AI, by the user. Several studies found that the arousal associated with cognitive dissonance is followed by an aversive state related to negative emotions (Craig et al., 1996; Critchley, 2005; Elliot & Devine, 1994; Martinie et al., 2013). For example, Matz and Wood (2005) found in an experiment that participants of a group in disagreement felt more discomfort in comparison to participants of a group, who were in agreement. Furthermore, discussion about contradictory options within the groups influences the mood state of the participant negatively. In addition, a number of studies suggest that negative emotions and bad mood influence the acceptance of traditional as well as innovative technologies (Beaudry & Pinsonneault, 2010; Chea & Luo, 2007; Read et al., 2011). Therefore, this study postulates the following hypothesis:

Hypothesis 1 (H1): Cognitive misfit based on a difference between the decisions and corresponding explanations of the AI and the user results in a more negative mood of the user as compared to cognitive fit.

Beside negative mood empirical studies on the decision-making process using *traditional* information systems, e.g., expert or recommendation systems, found further negative consequences of *cognitive misfit* between the user and the system (Dalal & Kasper, 1994; Hwang, 1994; Lamberti & Wallace, 1990; Tsekouras & Li, 2015). For example, Giboney et al. (2015) found that in the case of cognitive misfit users evaluate the perceived quality of a system's support, as well as the usefulness, lower than in the case of cognitive fit. Therefore, this study postulates the following second hypothesis:

Hypothesis 2 (H2): Cognitive misfit based on a difference between the decisions and corresponding explanations of the

AI and the user results in a more negative evaluation of the AI's support as compared to cognitive fit.

The CDT states that individuals apply different strategies to reduce this cognitive dissonance (Festinger, 1957), e.g., modifying their cognitions or ignoring contradictory information. However, previous studies also found resistance to changing attitudes, opinions or decisions for many situations – especially in the case of increased negative affect (Devine et al., 1999; Festinger & Carlsmith, 1959; Pyszczynski et al., 1993). For example, the confirmation bias (Wason, 1960, 1968) postulated that users tend to ignore contradictory information and do not change their opinion, attitude or decision. These findings are in line with the study of Giboney et al. (2015) where the system had less influence on user's decision in the case of cognitive misfit. Based on these assumptions, it can be suggested that, in case of cognitive misfit, users disregard AI's different decision and corresponding explanation, do not change their decision and are less willing to use the AI again. Therefore, the following hypothesis is formulated:

Hypothesis 3 (H3): Cognitive misfit based on a difference between the decisions and corresponding explanations of the AI and the user results in a reduced willingness to use the AI again as compared to cognitive fit.

4. Method

To empirically test the research model and the suggested hypotheses, we designed an experiment (Bierhoff & Petermann, 2013) based on the Wizard of Oz method (Klemmer et al., 2000).

4.1. Research design

During the experiment participants had to interact with a chatbot three times to reach a decision, i.e., A or B (see the screenshot in the Appendix). While they assumed the chatbot was intelligent, the answers were in fact given by the investigator based on a predefined script (Klemmer et al., 2000). The Wizard of Oz approach can be used to evaluate an early-stage interactive AI prototype by the users (Kerly & Bull, 2006). Yang et al. (2020, p. 3) state that “this approach enables HCI professionals to rapidly explore many design possibilities and probe user behaviors.” We applied a chatbot as a case in point of an AI system because many people are familiar with the usage of chatbots in their daily life (Brandtzaeg & Følstad, 2017).

What kind of decisions were made? The selected decisions refer to situations, which participants may encounter in their everyday private life. For instance, participants were asked to decide whether they would buy a car at a certain price or not, given a set of financing conditions. To trigger different cognitive styles between the chatbot and user, each situation required participants to provide an intuitive and

Table 2. Decision situations of the experiment in reference to their level of uncertainty and addressed cognitive biases

Cognitive bias	Uncertainty		
	Low	Medium	High
Loss aversion	Salary negotiations	Flight offer	Traffic jam
Loss aversion and temporal discounting	Private pension insurance	Job offer	Investment start-up
Loss aversion and temporal discounting and affect heuristic	Car purchase	Medication	Medical surgery

quick decision (approximately 1–2 min per decision), enticing them to apply targeted cognitive biases and heuristics. Each situation was designed to trigger one or more out of three cognitive biases and heuristics: loss aversion (Kahneman & Tversky, 1979), temporal discounting (Doyle, 2012), and affect heuristic (Finucane et al., 2000). We conducted a pre-test with 15 persons to check the different decision situations. We evaluated whether the participants understand the situations and whether the situations could actually happen in their daily life. Furthermore, we analysed whether the participants could make a decision based on the given information and in the given time. Furthermore, we investigated whether the cognitive biases and heuristics influenced the decision of the participants in the suggested direction.

Prior investigations for “traditional” human-computer interaction demonstrated an effect of task conditions, such as complexity, ambiguity, and uncertainty, on user acceptance as well as on performance (Lamberti & Wallace, 1990; Nissen & Sengupta, 2006; Roth et al., 1987). Therefore, to control for this effect, the nine situations also varied in regard to the levels of uncertainty (low, medium, high) (see Table 2). The levels of uncertainty were defined in reference to Courtney et al. (1997). In their study, Courtney et al. (1997) introduced different levels of uncertainty and identified strategies how to deal with decision-making in each respective uncertainty level. Level 1 (“low uncertainty”) describes a situation where future states can be predicted easily. At level 2 (“medium uncertainty”), existing information is precise enough to describe the future as one of a few discrete alternative scenarios. The precise outcome cannot be predicted, although analysis may help establish probabilities. At level 3 (“high uncertainty”) probabilities are not well bounded and therefore no natural discrete scenarios can be defined. Here only a range or continuum of potential futures can be identified, defined by a limited number of key variables.

A detailed description of the final nine situations (three levels of uncertainty x three levels of bias strength), time allowed for each decision, applied heuristics, and the set of possible explanations is shown in Appendix A.

Participants made decisions for all nine situations, but they were supported by the chatbot in only three of these. During the three interaction sequences with the chatbot we manipulated systematically whether the interaction would result in cognitive fit or misfit for the user. The “wizard” (meaning the experimenter) behind the scenes followed a script on when to agree (cognitive fit condition) or disagree (cognitive misfit condition) with the user. The script was designed in such a way, that the chatbot agreed with the user on one decision and disagreed on the remaining two

decisions in random order. For each situation we predefined three possible explanations. In the cognitive fit condition the “wizard” chose the explanation closest to the user’s explanation. In the cognitive misfit condition, the “wizard” chose the explanation that differed the most from the user’s explanation. While the explanations were chosen by the experimenter, participants were under the assumption that they were interacting with an AI.

To measure the effects of cognitive misfit on user’s acceptance we presented different questions during the interaction with the chatbot. The acceptance is operationalized by the three different components, i.e. the affective (mood), the cognitive (evaluation of AI’s support), and the conative (willingness to use the AI) dimensions, in order to reach a valid measurement (Asiegbu et al., 2012; Hasan, 2010; Jain, 2014; Makanyeza, 2014). The cognitive dimension represents the knowledge, beliefs, and opinions about the target behavior, while expressions of feelings, mood, or emotions toward the target behavior are assigned to the affective dimension (Asiegbu et al., 2012). Finally, the conative dimension can be described as mental processes that activate target behavior and thus reflect behavioral tendencies (Asiegbu et al., 2012). Previous studies also included these three different dimensions in their research model in order to predict acceptance towards “traditional” information systems or AI (e.g., Beaudry & Pinsonneault, 2010; Chuttur, 2009; Sharp, 2006; Tang & Hsiao, 2016). Therefore, the definition of acceptance is in line with existing studies.

4.2. Procedure

Before the experiment, participants received an email, explaining the purpose, duration, and procedure of the study. Participants were told that the purpose of the study was to improve a chatbot which was designed to support individuals during their daily life in decision-making. Due to the restrictions of the COVID-19 pandemic, the experiment had to take place in the participants’ home environment and not in a lab setting. At the beginning of the experiment, the investigator sent a link to the online survey and chatbot to the participant by email. Directly afterwards the participants were called by telephone to describe how to set up the survey and chatbot in two different tabs of the internet browser, so that both were visible at the same time.

On completion of this call, the participants started with the online survey. Firstly, the experiment was introduced (e.g., procedure, data declaration). Furthermore, the purpose of the chatbot was explained to participants again. We informed them about the characteristics of the chatbot to ensure that the participants perceived the chatbot as intelligent. For example, the participants received the information

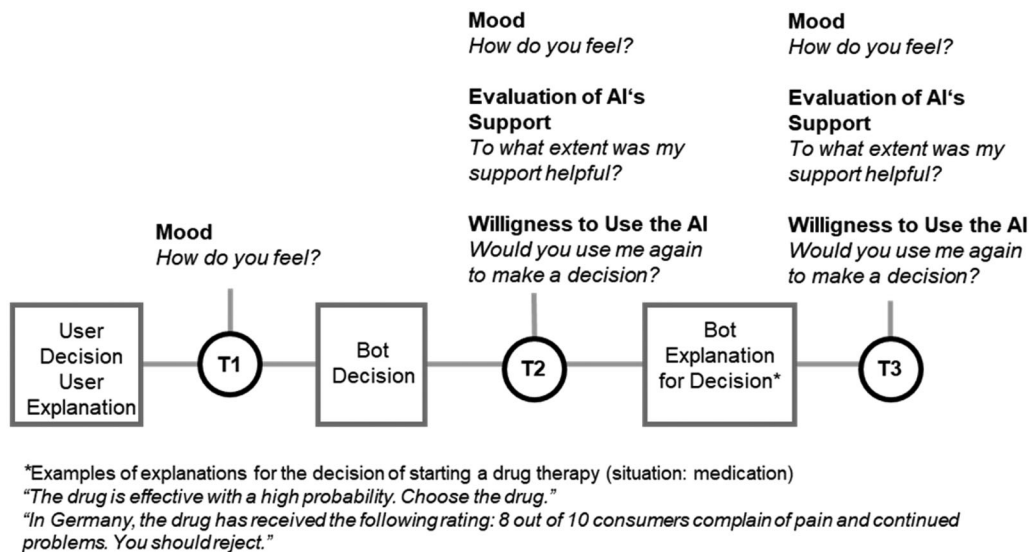


Figure 3. Sequence of interaction for each decision situation.

that the chatbot had access to vast amounts of information on the internet and was able to process this information to make a reasoned decision and formulate a suitable explanation in a very short time. In the next step, nine decision situations were presented one by one. At the end of each situation the participant had to make a decision by choosing option A or option B in a short given timeframe (see Appendix A for a detailed description of the decision situations and timeframes). After three of the nine situations, the participants chatted with the bot. The selection of the three situations was randomized. Lastly, the participants received questions about their gender, age, and knowledge of AI.

During the interaction with the chatbot, the participants were firstly asked about their choice (A or B) and their reason for the decision (see Figure 3). Furthermore, they had to rank their current mood using a 7-point Likert scale (T1). In the second step the chatbot gave its decision without an explanation. In the case of a differing decision, the AI asked the participant whether he or she would like to change her or his decision. Additionally, the chatbot asked the participants about their current mood, the evaluation of the AI's support, and the willingness to use the AI again. To answer these questions the chatbot offered again a 7-point Likert scale to the participants (T2). In the third step, the AI gave an explanation for its decision (see Figure 3 for an example). Finally, the participants were asked to rank their current mood, the evaluation of AI's support, and the willingness to use the AI again (T3).

4.3. Participants

In May and June 2020, participants were recruited from different regions of Germany to attend a one and a half hour-experiment. We published a post on different social media platforms and invited participants to the experiment. There were no specific requirements for participation other than providing an email address and having access to the internet during the experiment. Both survey and chatbot can be

opened with a conventional web browser. Participants did not receive any incentives for their participation. Before the experiment commenced, the participants did not have prior experience with a chatbot helping them to make decisions. The recruitment process attempted to enlist participants from a range of age groups. The study draws from a sample of $N=78$ participants ranging from 18 to 80 years (mean: 32.75 years) of which 57% were female. More than one-third were students (38%). Another third were employees in full- and part-time jobs (30.4%). The remaining third came from other backgrounds including pupils (3.8%), being out of job (6.3%), employees in training (3.8%), or self-employed workers.

4.4. Analysis

Data were analyzed using IBM SPSS Version 22.0. Firstly, the structure and distribution of the data was explored with the aid of descriptive statistical approaches to identify outliers and failed data records. Secondly, the requirements of the test procedure, i.e., normal distribution and homogeneity of variance, of the applied statistical approaches were checked including the Kolmogorov-Smirnov-Test (normal distribution; Rosenthal, 1968), Levene-Test (homogeneity of variance; Glass, 1966) and the Mauchly-Test (test of sphericity; Howell, 2002).

Firstly, we tested hypothesis H1 "Cognitive misfit based on a difference between the decisions and corresponding explanations of the AI and the user results in a more negative mood in the user as compared to cognitive fit." We analyzed the independent, nominal-scaled variables "cognitive fit/misfit" and the dependent, ordinal-scaled variable "mood" before and after the decision. As mentioned above, the variable "mood" was measured three times during the conversation with the chatbot: at the beginning (T1), after the decision of the AI (T2), and after the explanation from the AI (T3). Repeated measurement ensured that a change in mood could be measured depending on the decision and

explanation of the AI in order to evaluate the research model, i.e., the effect of cognitive misfit and cognitive dissonance on the mood of the user.

Based on this study design H1 cannot be rejected, if two conditions are fulfilled: Firstly, the mean of the variable “mood” has to be significantly ($p < 0.05$) lower in decision situations with cognitive misfit (second and third time point of measurement; T2 and T3) in comparison to decision situations with cognitive fit. The Mann-Whitney-Test (MacFarland & Yates, 2016) was applied to check this first assumption. The second condition refers to the change of mood over time. In decision situations with cognitive misfit, the mean of the variable “mood” has to deteriorate significantly ($p < 0.05$) over time. The Friedman-Test (MacFarland & Yates, 2016) was conducted to check this second assumption. To test both assumptions two non-parametric approaches were chosen as the requirements of the equivalent parametric approaches, like the repeated measures analyses of variance, were not met.

Testing hypotheses H2 “Cognitive misfit based on a difference between the decisions and corresponding explanations of the AI and the user results in a more negative evaluation of the AI’s support as compared to cognitive fit.” and H3 “Cognitive misfit based on a difference between the decisions and corresponding explanations of the AI and the user results in a reduced willingness to use the AI again as compared to cognitive fit.” is also based on the independent, nominal-scaled variable “cognitive fit/misfit” as mentioned above and the dependent, ordinal-scaled variables “evaluation of AI’s support” (H2) and “willingness to use the AI” (H3). As mentioned above, both variables “evaluation of AI’s support” and “willingness to use the AI” were measured two times during the conversation with the chatbot: after the decision of the AI (T2) and after the explanation of the AI (T3). These two times of measurement ensure that the changes in the variables “evaluation of AI’s support” and “willingness to use the AI” can be measured depending on the decision and explanation of the AI in order to evaluate the research model, i.e., the effect of cognitive misfit and cognitive dissonance on the evaluation of AI’s support (H2), and willingness to use the AI (H3).

Based on this study design H2 and H3 cannot be rejected, if two conditions are fulfilled: Firstly, the means of the variables “evaluation of AI’s support” (H2) resp. “willingness to use the AI” (H3) have to be significantly ($p < 0.05$) lower in decision situations with cognitive misfit in comparison to decision situations with cognitive fit. The Mann-Whitney-Test (MacFarland & Yates, 2016) was applied to check this first assumption. Secondly, in decision situations with cognitive misfit the means of the variables “evaluation of AI’s support” (H2) resp. “willingness to use the AI” (H3) have to deteriorate significantly ($p < 0.05$) over time. The Wilcoxon-Signed-Rank-Test (Woolson, 2007) was conducted to check this second assumption. To test both assumptions two non-parametric approaches were chosen as the requirements of equivalent parametric approaches, like the repeated measures analyses of variance, were not met.

5. Results

Initially we checked that all users started from the same level of “mood” at T1. Before the AI reveals its decision, the “mood” should be independent of whether a cognitive misfit or a cognitive fit situation is going to follow and thus no cognitive dissonance occurs. Indeed, a Mann-Whitney-Test reveals no significant difference in the value of the variable “mood” between participants in decision situations with cognitive misfit ($M_{T1} = 6.03$) at measurement T1 and participants in decision situations with cognitive fit ($M_{T1} = 6.12$), $U_{T1} = -0.797$, $p = 0.426$, $1 - \beta(d = 0.5) = 0.93$. In other words, a subsequent difference in mood must be caused by the AI’s action and cannot be explained by different starting points across conditions.

In contrast, at the second and third measurement (T2, T3), where the decision and explanation of the AI is known and cognitive dissonance might have occurred, the value of the variable “mood” should be significantly lower in situations with cognitive misfit in comparison to cognitive fit. In line with our expectation, a Mann-Whitney-Test indicates for the second and third time points of measurement (T2 and T3) where the decision and explanation of the AI is known and cognitive dissonance might occur, that the values of the variable “mood” are significantly lower in the case of participants in decision situations with cognitive misfit ($M_{T2} = 5.89$; $M_{T3} = 5.54$) than in the case of participants in decision situations with cognitive fit ($M_{T2} = 6.37$, $M_{T3} = 5.95$), $U_{T2} = -4.076$, $p < 0.001$; $U_{T3} = -2.646$, $p = 0.004$.

Furthermore, in the case of decision situations with cognitive misfit, participants’ value on the variable “mood” should deteriorate over time, i.e., both after the decision (second time point of measurement T2) and secondly, after the explanation (third time point of measurement T3). The Friedman-Test shows that this is indeed the case ($\chi^2(2, N = 105) = 19.130$, $p < 0.001$). Both the announcement of a contrary decision as well as the associated explanation have a negative impact on “mood.”

Therefore, the first hypothesis is supported by the results. The findings suggest that in the case of cognitive misfit resulting from different decisions and corresponding explanations of the AI and the user, the user’s mood is significantly more negative in comparison to cases of cognitive fit. Figure 4 illustrates the course of the mean values across the three time points of measurement T1, T2, and T3 for decision situations with cognitive misfit (solid line) and cognitive fit (dashed line).

Hypothesis H2 examines the change in “evaluation of AI’s support” when cognitive misfit occurs at T2 and T3. We expected that participants who experience cognitive misfit will score lower. A Mann-Whitney-Test reveals for both time points of measurement (T2 and T3) lower values of the variable “evaluation of AI’s support” for participants in decision situations with cognitive misfit ($M_{T2} = 3.64$; $M_{T3} = 4.36$) in comparison to participants in decision situations with cognitive fit ($M_{T2} = 4.05$; $M_{T3} = 4.83$, $U_{T2} = -1.815$, $p = 0.035$; $U_{T3} = -2.162$, $p = 0.016$). This suggests that cognitive misfit reduces the subjective support. However, in contrast to our expectations, the evaluation improved from

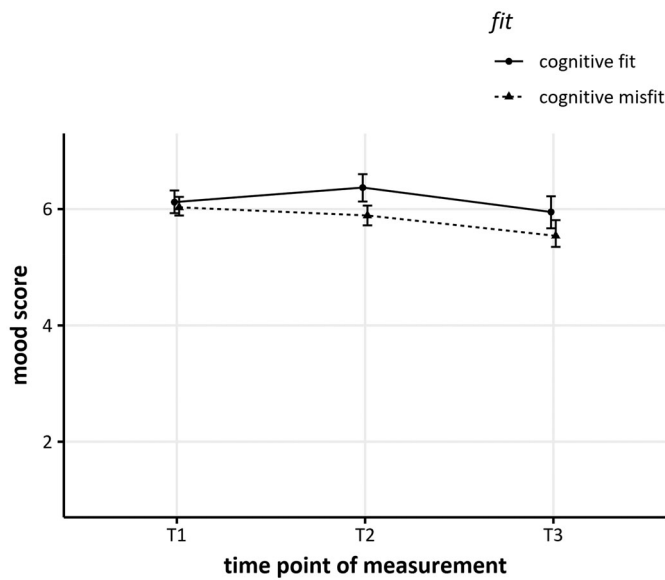


Figure 4. Results of Hypothesis H1: participants' mean "mood" across time points of measurement.

T2 to T3 in the case of decision situations with cognitive misfit. A Wilcoxon-Signed-Rank-Test (Woolson, 2007) reveals that participants' value on the variable "evaluation of AI's support" was higher at the third time point of measurement (T3) than on the second time point of measurement (T2), $Z = -4.035$, $p < 0.001$. Therefore, giving an explanation for a contrary decision improved the subjective support despite of the cognitive misfit.

Accordingly, these results are partially consistent with hypothesis H2. The "evaluation of AI's support" is indeed lower when cognitive misfit occurs. However, giving an explanation for the unexpected decision can re-establish subjective support. Figure 5 illustrates the mean values over the two time points of measurement T2 and T3 for decision situations with cognitive misfit (solid line) and cognitive fit (dashed line).

Hypothesis H3 examines the "willingness to use the AI" in future. In parallel to H2, we expected lower scores when cognitive misfit occurs. A Mann-Whitney-Test reveals for the second and third time points of measurement (T2 and T3) no significant difference in regard to "willingness to use the AI" between participants in decision situations with cognitive misfit ($M_{T2} = 4.06$; $M_{T3} = 4.18$) and participants in decision situations with cognitive fit ($M_{T2} = 4.35$; $M_{T3} = 4.47$), $U_{T2} = -1.376$, $p = 0.169$; $U_{T3} = -1.478$, $p = 0.139$, $1-\beta(d=0.5) = 0.93$. Furthermore, participants' willingness to use the AI again did not change from T2 to T3. A Wilcoxon-Signed-Rank-Test shows in the case of decision situations with cognitive misfit that participants' value on the variable "willingness to use the AI" was not significantly higher on the second time point of measurement (T2) compared to the third time point of measurement (T3), $Z = -1.658$, $p = 0.097$, $1-\beta(d=0.5) = 0.99$. Again, these results are not consistent with hypothesis H3 because during both time points of measurement (T2 and T3) the user knows the decision or explanation of the AI and cognitive misfit and cognitive dissonance should already have occurred for participants in decision situations with

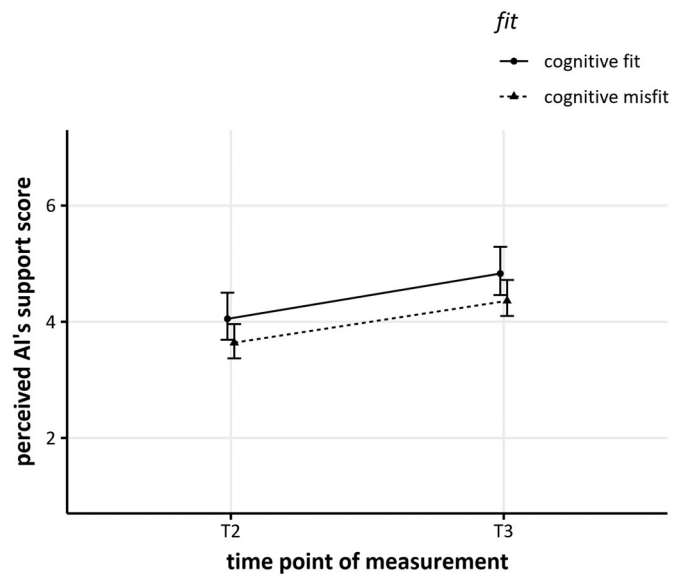


Figure 5. Results of Hypothesis H2: participants' mean "evaluation of AI's support" across time points of measurement.

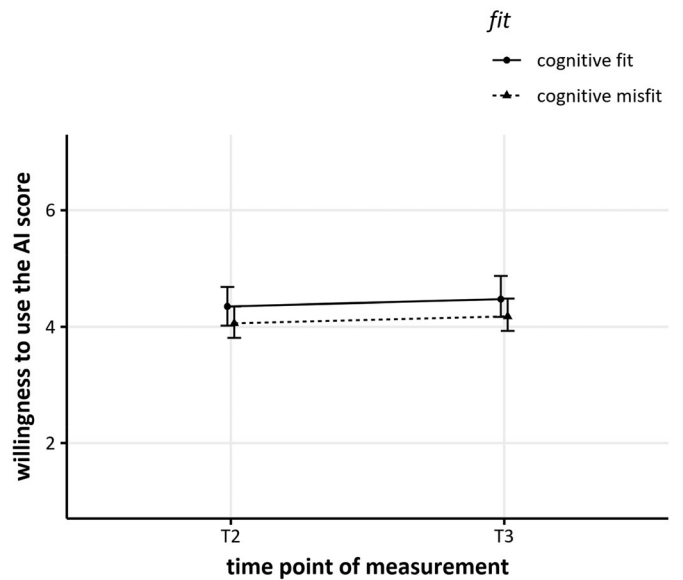


Figure 6. Results of Hypothesis H3: participants' mean "willingness to use the AI" across time points of measurement.

cognitive misfit. Therefore, the third hypothesis can be falsified. In the case of cognitive misfit resulting from different decisions and corresponding explanations of the AI and the user, users do not have significantly less willingness to use the AI again in comparison to cases where cognitive fit occurs. Figure 6 illustrates the mean values over the two time points of measurement T2 and T3 for decision situations with cognitive misfit (solid line) and cognitive fit (dashed line).

6. Discussion

6.1. Discussion of results

The main aim of this research was to investigate the effect on user acceptance where decisions and the associated

provided explanations contradict between an AI and the user. Based on a research model which was developed with the aid of established theories, i.e. cognitive fit theory (Vessey, 1991), the explanation evaluation framework (Gunning & Aha, 2019), and cognitive dissonance theory (Festinger, 1957), this research hypothesized that cognitive misfit in the user leads to less acceptance of an individual towards AI during the decision-making process. The research model assumes that cognitive dissonance in the user occurs when there is cognitive misfit between AI's and user's decision as well as the corresponding explanation. This cognitive dissonance has a negative influence on the user's mood, the user's evaluation of AI's support as well as the user's willingness to use the AI and thus, decreases the acceptance towards the AI.

An experiment with 78 participants was conducted using the Wizard of Oz method. The findings of the experiment showed that only the user's mood and the user's evaluation of AI's support were significantly influenced by a cognitive misfit. In decision situations with cognitive misfit, users are significantly more often in a negative mood and tend significantly more to a negative evaluation of AI's support in comparison to decision situations where cognitive fit exists.

In addition, in decision situations with cognitive misfit, giving a contradictory explanation after the decision further intensifies this negative trend. In contrast, the evaluation of AI's support is further enhanced after the contradictory explanation. This result indicates that contradictory explanations have different effects on the affective and cognitive dimensions of acceptance. While the user's mood can worsen after a contradictory explanation, the user's evaluation of AI's support can improve although a contradictory explanation has been given.

In general, the results of this study are in accordance with hypotheses H1 and H2 as well as other acceptance research about traditional information systems or AI which shows e.g., that explanations that fit users' cognitive style are perceived as being of higher quality (Giboney et al., 2015; Rzepka & Berger, 2018) or underlining the importance of user's emotional response on the acceptance (Beaudry & Pinsonneault, 2010; Chea & Luo, 2007; Lee, 2018; Shin, 2021).

One reason for the different effect of contradictory explanations on the affective and cognitive components of acceptance may be provided by past studies applying the dissonance theory. Previous studies have often confirmed the effect of cognitive dissonance on emotion or emotional states but not on cognitive opinions (Elliot & Devine, 1994; Harmon-Jones, 2000; Rhodewalt & Comer, 1979; Zanna & Cooper, 1974). Therefore, the cognitive dissonance – triggered by the contradictory decision and explanation – might influence only the mood and not the opinion of the user i.e. the evaluation of AI's support. However, previous studies indicate that negative emotion and aversive states have a significant influence on the evaluation of information systems (Beaudry & Pinsonneault, 2010; Chea & Luo, 2007; Read et al., 2011). Therefore, the results indicate that the user's negative mood influences the evaluation of AI's support negatively and the cognitive misfit and cognitive dissonance

do not have a direct effect on the evaluation of AI's support as suggested in the hypothesis H2.

Furthermore, the absence of the expected effect of cognitive misfit on the conative component of acceptance could potentially be explained by factors outside the scope of this study, which also influence the willingness to use the AI. Previous studies and research models analyzing the acceptance of information systems or AI indicate that e.g., perceived usefulness, ease of use, effort expectancy, performance expectancy, or curiosity in technology predict the willingness to use (also described as behavioral intention) (Im et al., 2011; Rauniar et al., 2014; Sohn & Kwon, 2020; Williams et al., 2015). Therefore, it could be assumed that the willingness to use the AI again is not affected by cognitive misfit and cognitive dissonance as suggested in the hypothesis H3 and other factors as mentioned above are more relevant. Alternatively, our measurement of intended use may not be sensitive enough to capture the long-term effects of cognitive dissonance immediately after each interaction sequence. Based on these findings, we reworked the postulated research model as illustrated in Figure 7.

6.2. Limitations and future research

This study has several limitations which future work should address. The research model was evaluated by using one special type of AI, i.e., a chatbot, which was used three times in an online experiment by the participants. We argue, that chatbots incorporate a number of characteristics that qualify this type of system as an interesting representative of AI systems. Chatbots support many aspects of AI's anthropomorphism which facilitates users' adoption (Araujo, 2018; Rietz et al., 2019; Sheehan et al., 2020). The form of interaction resembles chats between humans and uses natural language. A second limitation is the fact that the decision situations only address fictive problems from the context of personal life although AI is also commonly applied in decision situations in business environments (Rai, 2020). Additionally, studies on "traditional" information systems suggest that the experience with the system has a significant influence on the evaluation of a traditional decision support system (Tsekouras & Li, 2015). Our study observed user acceptance during a short period of interaction only. Participants indicated their willingness to use the system in future, but the research design did not allow to observe long-term behavior. Nevertheless, studies about user acceptance demonstrated that user acceptance in an early stage of interaction predicts continuance usage of a system (Bhattacharjee, 2001). The transferability of the results to decision situations in a business environment for long-term usage should be investigated in future studies. Furthermore, other types of AI for decision support need to be investigated in regard to user acceptance as well.

The cognitive misfit and resulting cognitive dissonance are two theoretical cognitive psychological constructs to explain a user's reduced acceptance towards AI. The mechanism of cognitive misfit and resulting cognitive dissonance is purely hypothetical and the connection to acceptance

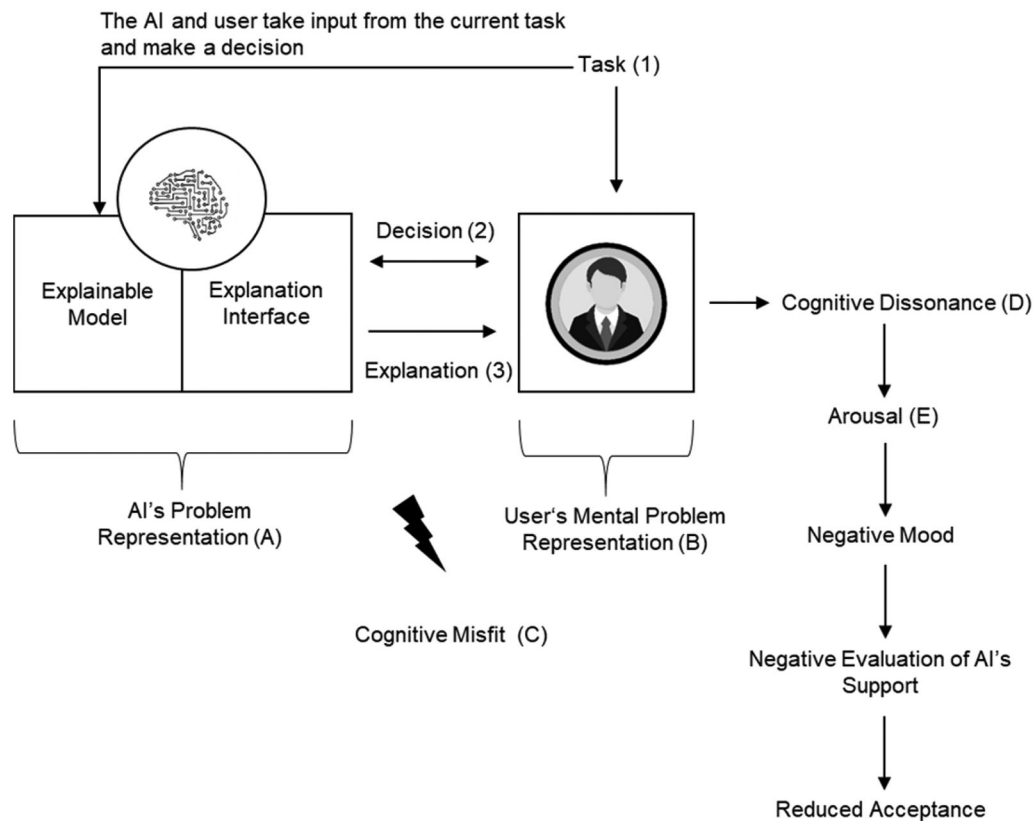


Figure 7. Modified research model.

cannot be measured directly although several studies suggest that the measurement of a participant's emotional response is a suitable means of measuring cognitive dissonance (Elliot & Devine, 1994). Nevertheless, measuring users' mood, users' evaluation of AI's support as well as users' willingness to use the AI may still provide an incomplete picture of user system acceptance. Other psychological or cognitive constructs resulting from the relationship between the AI and user or the explanation itself are likely to be relevant, such as trust, intimacy, comfort, transferability, information, causality, and explicitness, to influence the collaboration between AI and the user or the acceptance towards explainable AI (Ehsan & Riedl, 2020; Lipton, 2018; Möller et al., 2018; Rosenfeld & Richardson, 2019). For example, Samek et al. (2017) and Holzinger et al. (2019) underline the importance of users' existing knowledge and beliefs to achieve a causal understanding of AI's explanation. However, most of these studies are theoretical in nature or applied in online surveys instead of physical measurement or observations of real long-term usage to assess a user's acceptance towards "traditional" information systems or AI. Therefore, future studies need to use other study designs and add more variables to the research model as well as apply more objective measurements, like physical measurements, or long-term usage behavior in order to evaluate the user's acceptance towards AI.

Another critical aspect of this study is that the effect of a contradictory decision and explanation on user's acceptance towards AI was confounded. Either the decision and the

explanation differ from the user's decision or they do not, but we did not implement all $2 \times 2 = 4$ pairings. Accordingly, the separate effect of decision and explanation and potential interaction effects remain unclear. However, in reference to the CFT and CDT it can be assumed that in this situation cognitive misfit and cognitive dissonance also arise and the user's mood becomes negative. Therefore, future studies need to analyze empirically the effect of cognitive misfit and cognitive dissonance on user's acceptance towards AI with a more complex study design.

Rzepka and Berger (2018) presented a framework for the interaction of AI and humans in order to illustrate the research gaps of human-AI interaction. Drawing on this framework both the task or context and the AI and user have influence on the interaction between AI and humans. While this and previous studies focused primarily on the user and AI itself as well as its interaction, the task or context remains largely unstudied. The effects of task conditions, such as complexity, ambiguity, and uncertainty, on decision-making performance are well established (Lamberti & Wallace, 1990; Nissen & Sengupta, 2006; Roth et al., 1987). Our study design addresses this fact by varying the level of uncertainty and the potential cognitive biases in a systematic way. However, we did not analyze the effects of these factors on the user's acceptance in order to reduce the complexity of this exploratory study. Therefore, based on this study data a new research approach could close the mentioned gap and analyze whether and how the different levels of uncertainty impact the acceptance in reference to cognitive misfit and cognitive fit.

6.3. Theoretical and practical implication

The results impose many practical and theoretical implications. Firstly, explanations do not necessarily lead to higher user acceptance of AI. In reference to other studies and practical stakeholders the different effects of contradictory decisions and corresponding explanations on the affective and cognitive components of acceptance are surprising. Many previous studies on explainable AI have emphasized the positive aspects of explanation and suggested explanations as a one size fits all solution in order to increase the acceptance of AI. However, the findings of this study indicate that if the AI has a contradictory decision and corresponding explanation in contrast to the user, participants want to reduce cognitive misfit and cognitive dissonance. This, in turn will result in user's negative mood and unsatisfactory evaluation of AI's support. The detrimental effect of negative mood on various aspects of working as well as private life, such as employees' daily task performance and service sabotage (Chi et al., 2015), prosocial organizational behavior and fairness (George, 1991), team processes and team performance (Jordan et al., 2006), sleep quality, physical activity and mental health (Ingram et al., 2020) or altruistic and helpful behavior (Capra, 2004), have been confirmed in a variety of studies. Moreover, in the context of system evaluation, Laumer et al. (2012) found that a negative assessment of a new financial information system leads to different negative consequences, including decreased organizational commitment as well as lower job satisfaction and an increased turnover intention with a higher number of sick days.

How can a system intervene when cognitive misfit is detected in order to avoid those consequences? One possible suggestion would be to allow a conversation between the AI and the user after the decision and corresponding explanation to discuss their reasoning. Thus, on one hand, both parties can learn from each other why they made this decision, and the AI can learn human-like argumentation and explanations. The AI can then apply these reasons to justify its decisions in the future and is no longer limited to "rational" argumentation. On the other hand, based on the conversation both parties can express their appreciation, understanding, as well as their freedom of action to each other. In reference to the reactance theory (Miron & Brehm, 2006) this interaction is indispensable for an efficient collaboration since both low perceived appreciation and threats to freedom have negative consequences, e.g., negative emotion (Nesterkin, 2013) and decreased motivation (Pavey & Sparks, 2009; Putri & Hovav, 2014).

Another possible suggestion would be a personalized explanation. Besides cognitive styles describing how users think about a task, different user goals like maximizing money, connecting with people, or saving time might be used to personalize explanations to achieve cognitive fit. Scheutz (2017) provides a comprehensive overview of shared mental models in human-agent teams including how users think about equipment, tasks, team interaction and teammates, which might all be suitable aspects to construct cognitive style. In sum, to create explanations that fit the cognitive style of the user and may therefore increase

acceptance, a comprehensive user model would be required representing the user's needs, goals preferences and the world at large (Fischer, 2001).

From a theoretical point of view, the study addresses many research gaps as mentioned above: the majority of research on human-AI interaction and explainable AI is theoretical in nature, does not investigate empirically the specific effect of explainable AI with humans, focuses on interpretability, and does not apply theories from social science. To address these gaps, this study concentrates on *explanations* instead of *interpretability*. Furthermore, the article develops a new research model combining theories from information system research as well as psychology and evaluates this model empirically in an experiment with participants.

Although some studies confirm the positive effect of explainable AI, it remains unclear in which cases explanations given by an AI affect the users' acceptance (Ehsan & Riedl, 2020). The study at hand showed whether and under which conditions explainable AI can increase acceptance of AI during the decision process. Furthermore, while current studies apply established acceptance theories, like the technology acceptance model (TAM), the theory of planned behavior (TPB), the unified theory of acceptance and use of technology (UTAUT), as well as the value-based adoption model (VAM) in order to predict the acceptance towards AI (Sohn & Kwon, 2020), the present study focuses on the effect of cognitive misfit and cognitive dissonance. Moreover, in contrast to these previous studies using online surveys (Lee, 2018) or simulation experiments (Shin, 2019; Shin & Biocca, 2018), this study conducted an online experiment applying the Wizard of Oz method and the use of a chatbot. These experimental conditions enabled the direct observation of participants' emotional and behavioral responses during the human-AI interaction.

7. Conclusion

We shine some light on the effect on user acceptance when decisions and the associated provided explanations contradict between an AI and the user. Our self-developed research model suggests that cognitive misfit arising from a difference between the decisions and corresponding explanations of the AI and the user results in a more negative mood in the user, a more negative users' evaluation of the AI's support, as well as a reduced willingness to use the AI. With the aid of an experiment we found, in line with our hypotheses, that participants in decision situations with cognitive misfit are significantly more often in a negative mood and tend significantly more to a negative evaluation of AI's support in comparison to decision situations where cognitive fit exists. However, we also discovered two unexpected results. Firstly, in decision situations with cognitive misfit the evaluation of AI's support is impacted negatively by a contradictory decision but reestablished after giving the corresponding explanation. Secondly, we were not able to demonstrate an effect of cognitive misfit on the users' willingness to use the AI in the future. The findings are based on decision situations from a private context and with

a special type of AI, i.e., a chatbot. To validate our research model and results, further studies need to apply other types of AI in business decision situations.

Disclosure statement

No potential conflict of interest was reported by the author(s).

References

- Abdul, A., Vermeulen, J., Wang, D., Lim, B. Y., & Kankanhalli, M. (2018). *Trends and trajectories for explainable, accountable and intelligible systems: An HCI research agenda* [Paper presentation]. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems.
- Adadi, A., & Berrada, M. (2018). Peeking inside the black-box: A survey on explainable artificial intelligence (XAI). *IEEE Access*, 6, 52138–52160. <https://doi.org/10.1109/ACCESS.2018.2870052>
- Agrawal, A., Gans, J., Goldfarb, A. (2016, October 7). *Managing the machines*. Data Science Association. Retrieved October 16, 2020, from <http://www.datascienceassn.org/sites/default/files/Managing%20the%20Machines%20-%20AI%20is%20Making%20Prediction%20Cheap,%20Posing%20New%20Challenges%20for%20Managers.pdf>
- Agrawal, A., Gans, J., Goldfarb, A. (2017). *What to expect from artificial intelligence*. MIT Sloan Management Review. Retrieved October 16, 2020, from <https://static1.squarespace.com/static/578cf5ace58c62ac649ec9ce/t/589a5c99440243b575aaedaa/1486511270947/What+to+Expect+From+Artificial+Intelligence.pdf>
- Agustina, L., Meyliana, M., & Tin, S. T. S. (2017). Assessing accounting students' performance in "cognitive misfit" condition. *Journal of Business & Retail Management Research*, 11(4), 131–139. <https://doi.org/10.24052/JBRMR/V11IS04/AASPICMC>
- Ahmad, S. N., & Laroche, M. (2017). Analyzing electronic word of mouth: A social commerce construct. *International Journal of Information Management*, 37(3), 202–213. <https://doi.org/10.1016/j.ijinfomgt.2016.08.004>
- Amershi, S., Weld, D., Vorvoreanu, M., Fourney, A., Nushi, B., Collisson, P., Suh, J., Iqbal, S., Bennett, P., Inkpen, K., Teevan, J., Kikin-Gil, R., & Horvitz, E. (2019). *Guidelines for human-AI interaction* [Paper presentation]. Proceedings of the Conference on Human Factors in Computing Systems. <https://doi.org/10.1145/3290605.3300233>
- Araújo, J., & Pestana, G. (2017). A framework for social well-being and skills management at the workplace. *International Journal of Information Management*, 37(6), 718–725. <https://doi.org/10.1016/j.ijinfomgt.2017.07.009>
- Araujo, T. (2018). Living up to the chatbot hype: The influence of anthropomorphic design cues and communicative agency framing on conversational agent and company perceptions. *Computers in Human Behavior*, 85, 183–189. <https://doi.org/10.1016/j.chb.2018.03.051>
- Arnold, V., Clark, N., Collier, P. A., Leech, S. A., & Sutton, S. G. (2006). The differential use and effect of knowledge-based system explanations in novice and expert judgment decisions. *MIS Quarterly*, 30(1), 79–97. <https://doi.org/10.2307/25148718>
- Asiegbu, I. F., Powei, D. M., & Iruka, C. H. (2012). Consumer attitude: Some reflections on its concept, trilogy, relationship with consumer behavior, and marketing implications. *European Journal of Business and Management*, 4(13), 38–50.
- Batra, D., & Antony, S. R. (2001). Consulting support during conceptual database design in the presence of redundancy in requirements specifications: An empirical study. *International Journal of Human-Computer Studies*, 54(1), 25–51. <https://doi.org/10.1006/ijhc.2000.0406>
- Beaudry, A., & Pinsonneault, A. (2010). The other side of acceptance: Studying the direct and indirect effects of emotions on information technology use. *MIS Quarterly*, 34(4), 689–710. <https://doi.org/10.2307/25750701>
- Bhattacharjee, A. (2001). Understanding information systems continuance: An expectation-confirmation model. *MIS Quarterly*, 25(3), 351–370. <https://doi.org/10.2307/3250921>
- Bierhoff, H. W., & Petermann, F. (2013). *Forschungsmethoden der Psychologie*. Hogrefe.
- Bouakkaz, M., Quinten, Y., Loudcher, S., & Strekalova, Y. (2017). Textual aggregation approaches in OLAP context: A survey. *International Journal of Information Management*, 37(6), 684–692. <https://doi.org/10.1016/j.ijinfomgt.2017.06.005>
- Brandtzaeg, P. B., Følstad, A. (2017). *Why people use chatbots* [Paper presentation]. Proceedings of the International Conference on Internet Science.
- Brynjolfsson, E., & McAfee, A. (2016). *The second machine age: Work, progress, and prosperity in a time of brilliant technologies*. WW Norton & Company.
- Byrne, R. M. (2019). Counterfactuals in explainable artificial intelligence (XAI): Evidence from human reasoning [Paper presentation]. Proceedings of the Twenty-Eighth International Joint Conference on Artificial Intelligence.
- Capra, M. C. (2004). Mood-driven behavior in strategic interactions. *American Economic Review*, 94(2), 367–372. <https://doi.org/10.1257/0002828041301885>
- Chan, D. (1996). Cognitive misfit of problem-solving style at work: A facet of person-organization fit. *Organizational Behavior and Human Decision Processes*, 68(3), 194–207. <https://doi.org/10.1006/obhd.1996.0099>
- Chea, S., Luo, M. M. (2007). *Cognition, emotion, satisfaction, and post-adoption behaviors of e-service customers* [Paper presentation]. Proceedings of the Hawaii International Conference on System Sciences.
- Chi, N. W., Chang, H. T., & Huang, H. L. (2015). Can personality traits and daily positive mood buffer the harmful effects of daily negative mood on task performance and service sabotage? A self-control perspective. *Organizational Behavior and Human Decision Processes*, 131, 1–15. <https://doi.org/10.1016/j.obhdp.2015.07.005>
- Choi, I. Y., Oh, M. G., Kim, J. K., & Ryu, Y. U. (2016). Collaborative filtering with facial expressions for online video recommendation. *International Journal of Information Management*, 36(3), 397–402. <https://doi.org/10.1016/j.ijinfomgt.2016.01.005>
- Chui, M., Manyika, J., Miremadi, M., Henke, N., Chung, R., Nel, P., & Malhotra, S. (2018). *Notes from the AI frontier: Insights from hundreds of use cases*. McKinsey Global Institute. Retrieved October 16, 2020, from http://governance40.com/wp-content/uploads/2018/12/MGI_Notes-from-AI-Frontier-Discussion-paper.pdf
- Chung, W. (2014). BizPro: Extracting and categorizing business intelligence factors from textual news articles. *International Journal of Information Management*, 34(2), 272–284. <https://doi.org/10.1016/j.ijinfomgt.2014.01.001>
- Chuttur, M. Y. (2009). Overview of the technology acceptance model: Origins, developments and future directions. *Working Papers on Information Systems*, 9(37), 1–24. <http://sprouts.aisnet.org/9-37>
- Cook, M., Heshmat, S. (2019). *The symbiosis of humans and machines: Planning for our AI-augmented future*. Cognizant. Retrieved October 11, 2020, from <https://www.cognizant.com/whitepapers/planning-for-our-ai-augmented-future-codex4744.pdf>
- Courtney, H., Kirkland, J., & Viguerie, P. (1997). Strategy under uncertainty. *Harvard Business Review*, 75(6), 67–79.
- Craig, A. D., Reiman, E. M., Evans, A. C., & Bushnell, M. C. (1996). Functional imaging of an illusion of pain. *Nature*, 384(6606), 258–260.
- Critchley, H. D. (2005). Neural mechanisms of autonomic, affective and cognitive integration. *The Journal of Comparative Neurology*, 493(1), 154–166. <https://doi.org/10.1002/cne.20749>
- Dalal, N. P., & Kasper, G. M. (1994). The design of joint cognitive systems: The effect of cognitive coupling on performance. *International Journal of Human-Computer Studies*, 40(4), 677–702. <https://doi.org/10.1006/ijhc.1994.1031>
- Dennis, A. R., & Carte, T. A. (1998). Using geographical information systems for decision making: Extending cognitive fit theory to map-based presentations. *Information Systems Research*, 9(2), 194–203. <https://doi.org/10.1287/isre.9.2.194>

- Devine, P. G., Tauer, J. M., Barron, K. E., Elliot, A. J., & Vance, K. M. (1999). Moving beyond attitude change in the study of dissonance-related processes. In E. Harmon-Jones & J. Mills (Eds.), *Science conference series. Cognitive dissonance: Progress on a pivotal theory in social psychology* (pp. 297–323). American Psychological Association.
- Dietvorst, B. J., Simmons, J. P., & Massey, C. (2018). 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>
- Doyle, J. R. (2012). Survey of time preference, delay discounting models. *Judgment and Decision Making*, 8(2), 116–135. <https://doi.org/10.2139/ssrn.1685861>
- Dugdale, J. (1996). A cooperative problem-solver for investment management. *International Journal of Information Management*, 16(2), 133–147. [https://doi.org/10.1016/0268-4012\(95\)00074-7](https://doi.org/10.1016/0268-4012(95)00074-7)
- Edwards, J. S., Duan, Y., & Robins, P. C. (2000). An analysis of expert systems for business decision making at different levels and in different roles. *European Journal of Information Systems*, 9(1), 36–46. <https://doi.org/10.1057/palgrave.ejis.3000344>
- Ehsan, U., & Riedl, M. O. (2020). Human-centered explainable AI: Towards a reflective sociotechnical approach. In C. Stephanidis, M. Kuros, H. Degen, & L. Reinerman-Jones (Eds.), *HCI international 2020 - Late breaking papers: Multimodality and intelligence. HCII 2020. Lecture notes in computer science* (pp. 449–466). Springer. https://doi.org/10.1007/978-3-030-60117-1_33
- Elliot, A. J., & Devine, P. G. (1994). On the motivational nature of cognitive dissonance: Dissonance as psychological discomfort. *Journal of Personality and Social Psychology*, 67(3), 382–394. <https://doi.org/10.1037/0022-3514.67.3.382>
- Festinger, L. (1957). *A theory of cognitive dissonance*. Stanford University Press.
- Festinger, L., & Carlsmith, J. M. (1959). Cognitive consequences of forced compliance. *The Journal of Abnormal and Social Psychology*, 58(2), 203–210. <https://doi.org/10.1037/h0041593>
- Finucane, M. L., Alhakami, A., Slovic, P., & Johnson, S. M. (2000). The affect heuristic in judgments of risks and benefits. *Journal of Behavioral Decision Making*, 13(1), 1–17. [https://doi.org/10.1002/\(SICI\)1099-0771\(200001/03\)13:1<1::AID-BDM333>3.0.CO;2-S](https://doi.org/10.1002/(SICI)1099-0771(200001/03)13:1<1::AID-BDM333>3.0.CO;2-S)
- Fischer, G. (2001). User modeling in human-computer interaction. *User Modeling and User-Adapted Interaction*, 11(1/2), 65–86. <https://doi.org/10.1023/A:1011145532042>
- Frias-Martinez, E., Magoulas, G., Chen, S., & Macredie, R. (2006). Automated user modeling for personalized digital libraries. *International Journal of Information Management*, 26(3), 234–248. <https://doi.org/10.1016/j.ijinfomgt.2006.02.006>
- Gedikli, F., Jannach, D., & Ge, M. (2014). How should I explain? A comparison of different explanation types for recommender systems. *International Journal of Human-Computer Studies*, 72(4), 367–382. <https://doi.org/10.1016/j.ijhcs.2013.12.007>
- George, J. M. (1991). State or trait: Effects of positive mood on prosocial behaviors at work. *Journal of Applied Psychology*, 76(2), 299–307. <https://doi.org/10.1037/0021-9010.76.2.299>
- Giboney, J. S., Brown, S. A., Lowry, P. B., & Nunamaker, J. F. Jr. (2015). User acceptance of knowledge-based system recommendations: Explanations, arguments, and fit. *Decision Support Systems*, 72, 1–10. <https://doi.org/10.1016/j.dss.2015.02.005>
- Glass, G. V. (1966). Testing homogeneity of variances. *American Educational Research Journal*, 3(3), 187–190. <https://doi.org/10.3102/00028312003003187>
- Guidotti, R., Monreale, A., Ruggieri, S., Turini, F., Giannotti, F., & Pedreschi, D. (2019). A survey of methods for explaining black box models. *ACM Computing Surveys*, 51(5), 1–42. <https://doi.org/10.1145/3236009>
- Gunning, D., & Aha, D. (2019). DARPA's explainable artificial intelligence (XAI) program. *AI Magazine*, 40(2), 44–58. <https://doi.org/10.1609/aimag.v40i2.2850>
- Guszcza, J., Lewis, H., Evans-Greenwood, P. (2017). *Cognitive collaboration: Why humans and computers think better together*. Deloitte University Press. Retrieved October 17, 2020, from <https://www2.deloitte.com/us/en/insights/deloitte-review/issue-20/augmented-intelligence-human-computer-collaboration.html>
- Harari, Y. N. (2016). *Homo Deus: A brief history of tomorrow*. Random House.
- Harmon-Jones, E. (2000). Cognitive dissonance and experienced negative affect: Evidence that dissonance increases experienced negative affect even in the absence of aversive consequences. *Personality and Social Psychology Bulletin*, 26(12), 1490–1501. <https://doi.org/10.1177/01461672002612004>
- Hasan, B. (2010). Exploring gender differences in online shopping attitude. *Computers in Human Behavior*, 26(4), 597–601. <https://doi.org/10.1016/j.chb.2009.12.012>
- Hoffman, R. R., Klein, G., & Mueller, S. T. (2018). Explaining explanation for “explainable AI”. *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, 62(1), 197–201. <https://doi.org/10.1177/1541931218621047>
- Hohman, F., Kahng, M., Pienta, R., & Chau, D. H. (2019). Visual analytics in deep learning: An interrogative survey for the next frontiers. *IEEE Transactions on Visualization and Computer Graphics*, 25(8), 2674–2693. <https://doi.org/10.1109/TVCG.2018.2843369>
- Holzinger, A., Langs, G., Denk, H., Zatloukal, K., & Mueller, H. (2019). Causability and explainability of artificial intelligence in medicine. *Data Mining and Knowledge Discovery*, 9(4), e1312. <https://doi.org/10.1002/widm.1312>
- Howell, D. (2002). *Statistical methods for psychology* (5th ed.). Duxbury Press.
- Huang, Z., Chen, H., Guo, F., Xu, J. J., Wu, S., & Chen, W. H. (2006). Expertise visualization: An implementation and study based on cognitive fit theory. *Decision Support Systems*, 42(3), 1539–1557. <https://doi.org/10.1016/j.dss.2006.01.006>
- Hwang, M. I. (1994). Decision making under time pressure: A model for information systems research. *Information & Management*, 27(4), 197–203. [https://doi.org/10.1016/0378-7206\(94\)90048-5](https://doi.org/10.1016/0378-7206(94)90048-5)
- Im, I., Hong, S., & Kang, M. S. (2011). An international comparison of technology adoption: Testing the UTAUT model. *Information & Management*, 48(1), 1–8. <https://doi.org/10.1016/j.im.2010.09.001>
- Ingram, J., Maciejewski, G., & Hand, C. J. (2020). Changes in diet, sleep, and physical activity are associated with differences in negative mood during COVID-19 lockdown. *Frontiers in Psychology*, 11, 2328. <https://doi.org/10.3389/fpsyg.2020.588604>
- Jain, V. (2014). 3D model of attitude. *International Journal of Advanced Research in Management and Social Sciences*, 3(3), 1–12.
- Janssen, C. P., Donker, S. F., Brumby, D. P., & Kun, A. L. (2019). History and future of human-automation interaction. *International Journal of Human-Computer Studies*, 131, 99–107. <https://doi.org/10.1016/j.ijhcs.2019.05.006>
- Jarrah, M. H. (2018). Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making. *Business Horizons*, 61(4), 577–586. <https://doi.org/10.1016/j.bushor.2018.03.007>
- Jordan, P. J., Lawrence, S. A., & Troth, A. C. (2006). The impact of negative mood on team performance. *Journal of Management & Organization*, 12(2), 131–145. <https://doi.org/10.1017/S1833367200004077>
- Kahneman, D., & Tversky, A. (1979). Prospect theory: An analysis of decision under risk. *Econometrica*, 47(2), 263–291. <https://doi.org/10.2307/1914185>
- Kao, J. H., Chan, T. C., Lai, F., Lin, B. C., Sun, W. Z., Chang, K. W., Leu, F. Y., & Lin, J. W. (2017). Spatial analysis and data mining techniques for identifying risk factors of out-of-hospital cardiac arrest. *International Journal of Information Management*, 37(1), 1528–1538. <https://doi.org/10.1016/j.ijinfomgt.2016.04.008>
- Keane, M. T., Kenny, E. M., Delaney, E., Smyth, B. (2021). *If only we had better counterfactual explanations: Five key deficits to rectify in the evaluation of counterfactual XAI techniques*. Retrieved October 12, 2020, from <https://arxiv.org/abs/2103.01035>
- Kerly, A., Bull, S. (2006). *The potential for chatbots in negotiated learner modelling: A wizard-of-oz study* [Paper presentation]. Proceedings of the International Conference on Intelligent Tutoring Systems.
- Kim, J. K., Kim, H. K., Oh, H. Y., & Ryu, Y. U. (2010). A group recommendation system for online communities. *International Journal*

- of *Information Management*, 30(3), 212–219. <https://doi.org/10.1016/j.ijinfomgt.2009.09.006>
- Kizilcec, R. F. (2016). *How much information? Effects of transparency on trust in an algorithmic interface* [Paper presentation]. Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems.
- Klein, G. (2015). A naturalistic decision making perspective on studying intuitive decision making. *Journal of Applied Research in Memory and Cognition*, 4(3), 164–168. <https://doi.org/10.1016/j.jar-mac.2015.07.001>
- Klemmer, S. R., Sinha, A. K., Chen, J., Landay, J. A., Aboobaker, N., & Wang, A. (2000). *Suede: A wizard of Oz prototyping tool for speech user interfaces* [Paper presentation]. Proceedings of the 13th Annual ACM Symposium on User Interface Software and Technology. <https://doi.org/10.1145/354401.354406>
- Kolbjørnsrud, V., Amico, R., Thomas, R. J. (2016). *The promise of artificial intelligence: Redefining management in the workforce of the future*. Accenture Institute for High Performance Business. Retrieved October 2, 2020, from https://www.accenture.com/_acnmedia/PDF-19/AI_in_Management_Report.PDF
- Lamberti, D. M., & Wallace, W. A. (1990). Intelligent interface design: An empirical assessment of knowledge presentation in expert systems. *MIS Quarterly*, 14(3), 279–311. <https://doi.org/10.2307/248891>
- Laumer, S., Maier, C., Weitzel, T., Eckhardt, A. (2012). *The implementation of large-scale information systems in small and medium-sized enterprises—A case study of work-and health-related consequences* [Paper presentation]. Proceedings of the Hawaii International Conference on System Sciences.
- Lebib, F. Z., Mellah, H., & Drias, H. (2017). Enhancing information source selection using a genetic algorithm and social tagging. *International Journal of Information Management*, 37(6), 741–749. <https://doi.org/10.1016/j.ijinfomgt.2017.07.011>
- Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, 5(1), 205395171875668. <https://doi.org/10.1177/2053951718756684>
- Lerner, J. S., Li, Y., Valdesolo, P., & Kassam, K. S. (2015). Emotion and decision making. *Annual Review of Psychology*, 66(1), 799–823. <https://doi.org/10.1146/annurev-psych-010213-115043>
- Licklider, J. C. (1960). Man-computer symbiosis. *IRE Transactions on Human Factors in Electronics, HFE-1*(1), 4–11. <https://doi.org/10.1109/THFE2.1960.4503259>
- Lipton, Z. C. (2018). The Mythos of Model Interpretability: In machine learning, the concept of interpretability is both important and slippery. *Communications of the ACM*, 61(10), 36–43. <https://doi.org/10.1145/3233231>
- MacFarland, T. W., & Yates, J. M. (2016). *Introduction to nonparametric statistics for the biological sciences* using R. Springer.
- Maedche, A., Legner, C., Benlian, A., Berger, B., Gimpel, H., Hess, T., Hinz, O., Morana, S., & Söllner, M. (2019). AI-based digital assistants. *Business & Information Systems Engineering*, 61(4), 535–544. <https://doi.org/10.1007/s12599-019-00600-8>
- Makanyeza, C. (2014). Measuring consumer attitude towards imported poultry meat products in a developing market: An assessment of reliability, validity and dimensionality of the tri-component attitude model. *Mediterranean Journal of Social Sciences*, 5(20), 874–881. <https://doi.org/10.5901/mjss.2014.v5n20p874>
- Martinie, M. A., Milland, L., & Olive, T. (2013). Some theoretical considerations on attitude, arousal and affect during cognitive dissonance. *Social and Personality Psychology Compass*, 7(9), 680–688. <https://doi.org/10.1111/spc3.12051>
- Martinie, M. A., Olive, T., Milland, L., Joule, R. V., & Capa, R. L. (2013). Evidence that dissonance arousal is initially undifferentiated and only later labeled as negative. *Journal of Experimental Social Psychology*, 49(4), 767–770. <https://doi.org/10.1016/j.jesp.2013.03.003>
- Matz, D. C., & Wood, W. (2005). Cognitive dissonance in groups: The consequences of disagreement. *Journal of Personality and Social Psychology*, 88(1), 22–37. <https://doi.org/10.1037/0022-3514.88.1.22>
- McAfee, A., & Brynjolfsson, E. (2016). Human work in the robotic future: Policy for the age of automation. *Foreign Affairs*, 95(4), 139–150.
- Miron, A. M., & Brehm, J. W. (2006). Reactance theory-40 years later. *Zeitschrift Für Sozialpsychologie*, 37(1), 9–18. <https://doi.org/10.1024/0044-3514.37.1.9>
- Möller, J., Trilling, D., Helberger, N., & van Es, B. (2018). Do not blame it on the algorithm. *Information, Communication & Society*, 21(7), 959–977. <https://doi.org/10.1080/1369118X.2018.1444076>
- Mueller, S. T., Hoffman, R. R., Clancey, W., Emrey, A., Klein, G. (2019). *Explanation in human-AI systems: A literature meta-review, synopsis of key ideas and publications, and bibliography for explainable AI*. Retrieved October 11, 2020, from <https://arxiv.org/abs/1902.01876>
- Nesterkin, D. A. (2013). Organizational change and psychological reactance. *Journal of Organizational Change Management*, 26(3), 573–594. <https://doi.org/10.1108/09534811311328588>
- Nissen, M. E., & Sengupta, K. (2006). Incorporating software agents into supply chains: Experimental investigation with a procurement task. *MIS Quarterly*, 30(1), 145–166. <https://doi.org/10.2307/25148721>
- Ogiela, L., & Ogiela, M. R. (2014). Cognitive systems for intelligent business information management in cognitive economy. *International Journal of Information Management*, 34(6), 751–760. <https://doi.org/10.1016/j.ijinfomgt.2014.08.001>
- Pavey, L., & Sparks, P. (2009). Reactance, autonomy and paths to persuasion: Examining perceptions of threats to freedom and informational value. *Motivation and Emotion*, 33(3), 277–290. <https://doi.org/10.1007/s11031-009-9137-1>
- Pentland, A., Daggett, M., Hurley, M. (2019, March). *Human-AI decision systems*. MIT Connection Science. Retrieved September 13, 2020, from <https://connection.mit.edu/sites/default/files/publication-pdfs/Human-AIDecisionSystems.pdf>
- Pieters, W. (2011). Explanation and trust: What to tell the user in security and AI? *Ethics and Information Technology*, 13(1), 53–64. <https://doi.org/10.1007/s10676-010-9253-3>
- Pisz, I., & Łapunka, I. (2017). Logistics project planning under conditions of risk and uncertainty. *Transport Economics and Logistics*, 68(1), 89–102. <https://doi.org/10.5604/01.3001.0010.5325>
- Putri, F. F., Hovav, A. (2014). *Employees compliance with BYOD security policy: Insights from reactance, organizational justice, and protection motivation theory* [Paper presentation]. Proceedings of the European Conference on Information System.
- Pyszczyński, T., Greenberg, J., Solomon, S., Sideris, J., & Stubing, M. J. (1993). Emotional expression and the reduction of motivated cognitive bias: Evidence from cognitive dissonance and distancing from victims' paradigms. *Journal of Personality and Social Psychology*, 64(2), 177–186. <https://doi.org/10.1037/0022-3514.64.2.177>
- Ragini, J. R., Anand, P. R., & Bhaskar, V. (2018). Big data analytics for disaster response and recovery through sentiment analysis. *International Journal of Information Management*, 42, 13–24. <https://doi.org/10.1016/j.ijinfomgt.2018.05.004>
- Rai, A., Constantinides, P., & Sarker, S. (2019). Editor's comments: Next-generation digital platforms: toward human-AI hybrids. *MIS Quarterly*, 43(1), iii–ix.
- 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>
- Ramírez-Noriega, A., Juárez-Ramírez, R., & Martínez-Ramírez, Y. (2017). Evaluation module based on Bayesian networks to Intelligent Tutoring Systems. *International Journal of Information Management*, 37(1), 1488–1498. <https://doi.org/10.1016/j.ijinfomgt.2016.05.007>
- Ransbotham, S., Kiron, D., Gerbert, P., & Reeves, M. (2017). Reshaping business with artificial intelligence: Closing the gap between ambition and action. *MIT Sloan Management Review*, 59(1), 1–17.
- Rauniar, R., Rawski, G., Yang, J., & Johnson, B. (2014). Technology acceptance model (TAM) and social media usage: An empirical study on Facebook. *Journal of Enterprise Information Management*, 27(1), 6–30. <https://doi.org/10.1108/JEIM-04-2012-0011>
- Read, W., Robertson, N., & McQuilken, L. (2011). A novel romance: The technology acceptance model with emotional attachment. *Australasian Marketing Journal*, 19(4), 223–229. <https://doi.org/10.1016/j.ausmj.2011.07.004>

- Rekik, R., Kallel, I., Casillas, J., & Alimi, A. M. (2018). Assessing web sites quality: A systematic literature review by text and association rules mining. *International Journal of Information Management*, 38(1), 201–216. <https://doi.org/10.1016/j.ijinfomgt.2017.06.007>
- Rhodewalt, F., & Comer, R. (1979). Induced-compliance attitude change: Once more with feeling. *Journal of Experimental Social Psychology*, 15(1), 35–47. [https://doi.org/10.1016/0022-1031\(79\)90016-7](https://doi.org/10.1016/0022-1031(79)90016-7)
- Rietz, T., Benke, I., Maedche, A. (2019). *The impact of anthropomorphic and functional chatbot design features in enterprise collaboration systems on user acceptance* [Paper presentation]. Proceedings of the 2019 Conference on Wirtschaftsinformatik.
- Rosenfeld, A., & Richardson, A. (2019). Explainability in human-agent systems. *Autonomous Agents and Multi-Agent Systems*, 33(6), 673–705. <https://doi.org/10.1007/s10458-019-09408-y>
- Rosenthal, R. (1968). An application of the Kolmogorov-Smirnov test for normality with estimated mean and variance. *Psychological Reports*, 22(2), 570. <https://doi.org/10.2466/pr0.1968.22.2.570>
- Roth, E. M., Bennett, K. B., & Woods, D. D. (1987). Human interaction with an “intelligent” machine. *International Journal of Man-Machine Studies*, 27(5–6), 479–525. [https://doi.org/10.1016/S0020-7373\(87\)80012-3](https://doi.org/10.1016/S0020-7373(87)80012-3)
- Rushkoff, D. (2019). *Team human*. Norton & Company.
- Rzepka, C., Berger, B. (2018). *User interaction with AI-enabled systems: A systematic review of IS research* [Paper presentation]. Proceedings of the International Conference of Information Systems.
- Samek, W., Binder, A., Montavon, G., Lapuschkin, S., & Müller, K. (2017). Evaluating the visualization of what a deep neural network has learned. *IEEE Transactions on Neural Networks and Learning Systems*, 28(11), 2660–2673. <https://doi.org/10.1109/TNNLS.2016.2599820>
- Scheutz, M., DeLoach, S. A., & Adams, J. A. (2017). A framework for developing and using shared mental models in human-agent teams. *Journal of Cognitive Engineering and Decision Making*, 11(3), 203–224. <https://doi.org/10.1177/1555343416682891>
- Selvaraju, R. R., Cogswell, M., Das, A., Vedantam, R., Parikh, D., Batra, D. (2017). *Grad-cam: Visual explanations from deep networks via gradient-based localization* [Paper presentation]. Proceedings of the IEEE International Conference on Computer Vision.
- Sharp, J. H. (2006). Development, extension, and application: A review of the technology acceptance model. *Director*, 7(9), 3–11.
- Sheehan, B., Jin, H. S., & Gottlieb, U. (2020). Customer service chatbots: Anthropomorphism and adoption. *Journal of Business Research*, 115, 14–24. <https://doi.org/10.1016/j.jbusres.2020.04.030>
- Shin, D., & Biocca, F. (2018). Exploring immersive experience in journalism what makes people empathize with and embody immersive journalism? *New Media & Society*, 20(8), 2800–2823. <https://doi.org/10.1177/1461444817733133>
- Shin, D. (2019). Toward fair, accountable, and transparent algorithms: Case studies on algorithm initiatives in Korea and China. *Javnost: The Public*, 26(3), 1–17. <https://doi.org/10.1080/13183222.2019.1589249>
- Shin, D., & Park, Y. J. (2019). Role of fairness, accountability, and transparency in algorithmic affordance. *Computers in Human Behavior*, 98, 277–284. <https://doi.org/10.1016/j.chb.2019.04.019>
- 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>
- Slovic, P., Fischhoff, B., & Lichtenstein, S. (1977). Behavioral decision theory. *Annual Review of Psychology*, 28(1), 1–39. <https://doi.org/10.1146/annurev.ps.28.020177.000245>
- Sohn, K., & Kwon, O. (2020). Technology acceptance theories and factors influencing artificial intelligence-based intelligent products. *Telematics and Informatics*, 47, 101324. <https://doi.org/10.1016/j.tele.2019.101324>
- Sundar, S. S. (2020). Rise of machine agency: A framework for studying the psychology of human-AI interaction (HAI). *Journal of Computer-Mediated Communication*, 25(1), 74–88. <https://doi.org/10.1093/jcmc/zmz026>
- Tang, K. Y., & Hsiao, C. H. (2016). The literature development of technology acceptance model. *International Journal of Conceptions on Management and Social Sciences*, 4(1), 1–4.
- Tintarev, N., & Masthoff, J. (2012). Evaluating the effectiveness of explanations for recommender systems. *User Modeling and User-Adapted Interaction*, 22(4–5), 399–439. <https://doi.org/10.1007/s11257-011-9117-5>
- Tjoa, E., & Guan, C. (2021). A survey on explainable artificial intelligence (XAI): Toward medical XAI. *IEEE Transactions on Neural Networks and Learning Systems*, 32(11), 4793–4813. <https://doi.org/10.1109/TNNLS.2020.3027314>
- Tsekouras, D., Li, T. (2015). *The dual role of perceived effort in personalized recommendations* [Paper presentation]. Proceedings of the European Conference on Information System.
- Tseng, S. S., Chen, H. C., Hu, L. L., & Lin, Y. T. (2017). CBR-based negotiation RBAC model for enhancing ubiquitous resources management. *International Journal of Information Management*, 37(1), 1539–1550. <https://doi.org/10.1016/j.ijinfomgt.2016.05.009>
- van den Bosch, K., & Bronkhorst, A. (2018). *Human-AI cooperation to benefit military decision making*. NATO. Retrieved January 3, 2021, from https://www.karelvandenbosch.nl/documents/2018_Bosch_et_al_NATO-IST160_Human-AI_Cooperation_in_Military_Decision_Making.pdf
- Vessey, I. (1991). Cognitive fit: A theory-based analysis of the graphs versus tables literature. *Decision Sciences*, 22(2), 219–240. <https://doi.org/10.1111/j.1540-5915.1991.tb00344.x>
- Vinson, N. G., Molyneux, H., & Martin, J. D. (2019). Explanations in artificial intelligence decision making: A user acceptance perspective. In P. Isaias & K. Blashki (Eds.), *Handbook of research on human-computer interfaces and new modes of interactivity* (pp. 96–117). IGI Global.
- Wachter, S., Mittelstadt, B., & Floridi, L. (2017). Why a right to explanation of automated decision-making does not exist in the general data protection regulation. *International Data Privacy Law*, 7(2), 76–99. <https://doi.org/10.1093/idpl/ix005>
- Wang, Y. Y., & Wang, Y. S. (2022). Development and validation of an artificial intelligence anxiety scale: An initial application in predicting motivated learning behavior. *Interactive Learning Environments*, 30(4), 619–634. <https://doi.org/10.1080/10494820.2019.1674887>
- Wason, P. C. (1960). On the failure to eliminate hypotheses in a conceptual task. *Quarterly Journal of Experimental Psychology*, 12(3), 129–140. <https://doi.org/10.1080/17470216008416717>
- Wason, P. C. (1968). On the failure to eliminate hypotheses: A second look. In P. C. Wason & P. N. Johnson-Laird (Eds.), *Thinking and reasoning* (pp. 165–174). Penguin.
- Williams, M. D., Rana, N. P., & Dwivedi, Y. K. (2015). The unified theory of acceptance and use of technology (UTAUT): A literature review. *Journal of Enterprise Information Management*, 28(3), 443–488. <https://doi.org/10.1108/JEIM-09-2014-0088>
- Wilson, H. J., & Daugherty, P. R. (2018). Collaborative intelligence: Humans and AI are joining forces. *Harvard Business Review*, 96(4), 114–123.
- Woolson, R. F. (2007). Wilcoxon signed-rank test. *Wiley Encyclopedia of Clinical Trials*, 1–3. <https://doi.org/10.1002/9780471462422.eoct979>
- World Economic Forum. (2016). *The global risks report 2016* (11th ed.). World Economic Forum. Retrieved March 2, 2021, from <https://www.weforum.org/reports/the-global-risks-report-2016>
- Yang, Q., Steinfeld, A., Rosé, C., & Zimmerman, J. (2020). *Re-examining whether, why, and how human-ai interaction is uniquely difficult to design* [Paper presentation]. Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems. <https://doi.org/10.1145/3313831.3376301>
- Yaqoob, I., Hashem, I. A. T., Gani, A., Mokhtar, S., Ahmed, E., Anuar, N. B., & Vasilakos, A. V. (2016). Big data: From beginning to future. *International Journal of Information Management*, 36(6), 1231–1247. <https://doi.org/10.1016/j.ijinfomgt.2016.07.009>
- Ye, L. R., & Johnson, P. E. (1995). The impact of explanation facilities on user acceptance of expert systems advice. *MIS Quarterly*, 19(2), 157–172. <https://doi.org/10.2307/249686>
- Zanna, M. P., & Cooper, J. (1974). Dissonance and the pill: An attribution approach to studying the arousal properties of dissonance. *Journal of Personality and Social Psychology*, 29(5), 703–709. <https://doi.org/10.1037/h0036651>
- Zhang, B., & Sundar, S. S. (2019). Proactive vs. reactive personalization: Can customization of privacy enhance user experience?

International Journal of Human-Computer Studies, 128, 86–99.
<https://doi.org/10.1016/j.ijhcs.2019.03.002>

Zhao, Y., Tang, L. C., Darlington, M. J., Austin, S. A., & Culley, S. J. (2008). High value information in engineering organisations. *International Journal of Information Management*, 28(4), 246–258.
<https://doi.org/10.1016/j.ijinfomgt.2007.09.007>

About the authors

Carolin Ebermann is a research associate at PFH Private University of Applied Sciences Göttingen, Germany. Before she was a User Experience Consultant at eresult GmbH. Carolin studied Psychology at

the TU Braunschweig. Afterwards, she completed her PhD in business informatics at the Georg-August-University of Göttingen.

Matthias Selisky is a UX Designer at generic.de software technologies AG. Before he was a research associate at PFH Private University of Applied Sciences Göttingen. Matthias earned a Master of Science in Neuropsychology from Oldenburg University, and a Bachelor of Science degree in Cognitive Science from Osnabrück University.

Stephan Weibelzahl is a professor of business psychology at PFH Private University of Applied Sciences Göttingen, Germany. He received a PhD from University of Trier, headed a research team at Fraunhofer IESE, Kaiserslautern and worked as lecturer at National College of Ireland, Dublin.

Appendix A. Decision situations

A detailed description of the situations, time for each decision, level of uncertainty, applied heuristics, and cognitive biases, as well as the set of possible explanations

Description of the decision situation	Options	Explanation (in agreement with decision A, B, or either)	Time allowed (in sec.) and level of uncertainty
Loss aversion			
Salary negotiations			
You work as an executive assistant and are in salary negotiations with your boss. Actually, this year you are due a salary increase of 1000 Euro. Your boss offers you to waive the 1000 Euro and to work 37 instead of 38 h per week. Consequently, instead of 51,000 Euro, you would earn 50,000 Euro net per year. Do you want to accept your boss's offer?#	A: 37 h, 50,000 Euro B: 38 h, 51,000 Euro	(A): Hourly rate is higher at 37 hours. Waive the salary increase. (B): Choose the 51,000 Euros, (Either): Business management is on average 30 (A)/35 Euro (B) per hour in Germany. So your hourly wage would be the average (A)/not the average (B).	40 low uncertainty
Flight offer			
You plan to go on vacation. You have saved 1000 Euro over 5 years. You look at possible destinations and find Singapore interesting. The return flight costs 999 Euro. The online provider advertises that if you buy now, you will get 250 Euro pocket money. Alternatively, you can get on a list that gives you 1000 Euro with a probability of 39%. How do you decide?	A: Book flight now and get 250 Euros extra B: Book flight now and get on the list	(A): Have the 250 Euros paid out, otherwise you may not get any money at all. (B): The expectation value is very high. Get on the list. (Either): Today, 89% of bookings have already chosen the list (B)/250 Euros (A).	60 medium uncertainty
Traffic jam			
You are driving on the highway. The radio tells you that there is a 10 km traffic jam ahead of you and that you will lose up to 30 min of time. You consider taking the next exit and taking secondary roads instead. The alternative route is 20 km longer than the highway. However, you do not know what the current traffic situation is like on the alternative route. How do you decide?	A: Stay on the highway B: Depart and drive overland	(A): You are faster on the highway. Choose option A. (B): Drive overland. You lose time on the highway. (Either): 85% (A)/25% (B) of people who stay in traffic jams on the highway make faster progress than on the secondary road.	50 high uncertainty
Loss aversion and temporal discounting			
Private pension insurance			
You are 20 years old. You want to invest some money from your 1500 Euro salary per month in a life insurance policy so that you can put money aside for after retirement. You receive an offer from your insurance company to receive 150 Euros per month less in salary from now until you retire at age 67. These 150 Euro are invested with an interest rate of 0.8%. If you decide against the model, the earliest you will be able to invest again is at age of 30 at an interest rate of 0.5%. How do you decide?	A: Take out the insurance now at 0.8%. B: Take out insurance at 30 at 0.5%	(A): Option A gives you a higher interest rate overall and you end up with higher total assets. Take out the insurance directly. (B): Take out insurance at 30 at 0.5% (Either): 70% of life insurance owners have an interest rate at 0.4% (A)/above 0.8% (B).	60 low uncertainty
A: Take out the insurance now at 0.8%. B: Take out insurance at 30 at 0.5%			
Job offer			
You are currently looking for a job and have gone through many interviews. Now you receive an	A: 42,000, 5 years B: 40,000, 8% every 12 months for 5 years	(A): Choose option A. This way you will directly receive 2000 euros more salary	60 medium uncertainty

(continued)

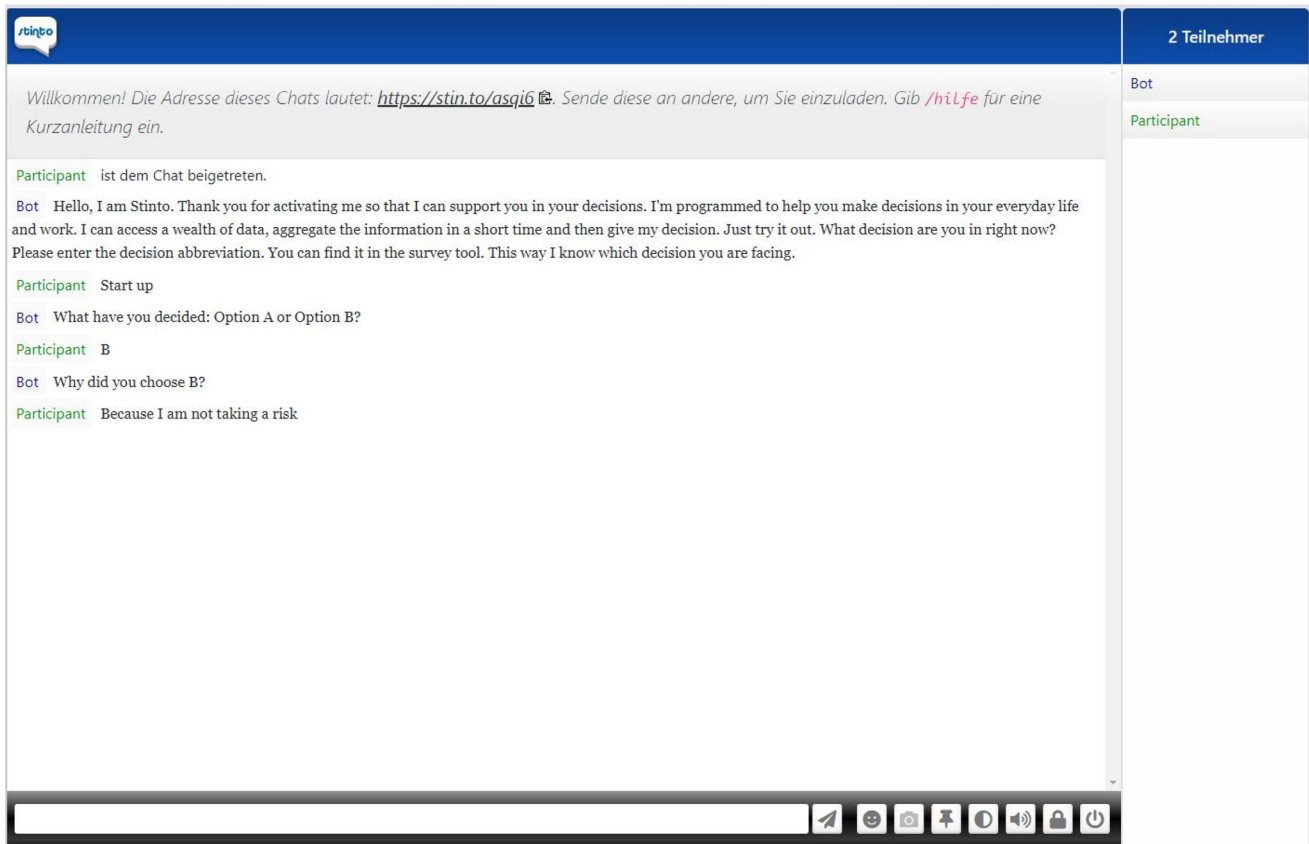
Appendix A. (Continued.)

Description of the decision situation	Options	Explanation (in agreement with decision A, B, or either)	Time allowed (in sec.) and level of uncertainty
acceptance letter from a company that you like. During the salary negotiation, the CEO offers you 42,000 Euro per year for the next 5 years (option A). You know that this means you will have 2200 Euro after taxes in your account per month. Alternatively, the managing director offers you 2000 Euro less in salary per year (option B). With Option B, you are promised that your pay increases 8% every 12 months for 5 years. How do you decide?		(B): In the long run, you get more pay in option B. Choose option B. (Either): Only 15% (A)/76% (B) of employees in Germany in their position get more than 40,000 euros.	
Investment start-up You are 40 years old. Your account contains 70,000 Euro that you have saved over the course of your life. You want to invest this money in ventures. You now have the opportunity to join a start-up with 50,000 Euro. You like the business idea. You would get 40% of the start-up's profit after 2 years. How do you decide?	A: Invest B: Not invest	(A): The industry is currently experiencing a strong increase in sales. Invest in the start-up. (B): Do not invest, as you could lose a lot of money if no profit is made. (Either): Only 3570 (B)/1 million (A) people in Germany have successfully invested their assets in startups.	30 high uncertainty
Loss aversion and temporal discounting and affect heuristic Car purchase You want to buy your dream car. To do so, however, you first have to sell your current Golf 5 to get more money. You know that you can bargain about 5000 Euro for your Golf 5. Now you find a car dealer who has your dream car on offer for 20,000 Euro. He would also buy your Golf 5 on the spot for 1000 Euro less than you had assumed, because he already has a buyer for it. You would be able to take the new car with you immediately, as he would offer you financing for the remaining amount of the new car. You would have to pay off 9% of the purchase price every month for 10 months. In the next few days, the offer would no longer be valid because the buyer for your current car would no longer be available. Do you go for the offer?	A: Reject offer B: Accept offer	(A): Do not buy, because this way you pay more for dream car and you get less money for your Golf. (B): You can take the dream car (Either): 80% (A)/20% (B) of private sellers of used cars get 850 Euro more than the car is worth.	90 low uncertainty
Medication You have an inflammation in your stomach area. There is a medication that you can take to heal the inflammation. Studies show that 30% of patients suffer chronic inflammation of the stomach area and have permanent pain during their lifetime, despite taking the drug. The drug also has a side effect, inflammation of the esophagus. 10% of patients will experience inflammation of the esophagus if they take the drug. Without the drug, patients recover completely 68% of the time. The doctor asks you if you want to be treated with medication?	A: With medication B: Without medication	(A): The drug is effective with a high probability. Choose the drug. (B): Don't take any drug as there is a chance of getting sick with drug too. (Either): In Germany, the drug has received the following rating: 8 (B)/1 (A) out of 10 consumers complain of pain and continued problems.	60 medium uncertainty
Medical surgery You are 25 years old and have had an accident. You arrive in hospital with several injuries. Your doctor tells you that you need a knee replacement. However, recovery is also possible without surgery. Surgery can only be done twice in a lifetime. If you have surgery now, you will still have pain afterwards with 30% certainty. If you continue to have pain after the first surgery, you might have another surgery while 70% of re-operated patients continue to have problems. How do you decide?	A: Surgery B: Not Surgery	(A): It is very unlikely that you will have knee problems again. Have an operation. (B): Don't have surgery because there are chances of more problems after surgery. (Either): According to the statistics in Germany, only 2 (A)/7 (B) out of 10 people suffer complications or consequential damages during the operations.	60 high uncertainty

Note: Which explanation was chosen by the wizard depended on the experimental condition (cognitive fit or cognitive misfit), the option selected by the participant (A or B), and the explanation given by the participant. In the cognitive misfit condition: The decision of the AI must contradict the participant's decision (A or B). Furthermore, the participant's explanation must refer to a different cognitive style. Therefore, the wizard chose the explanation that differed the most from the user's explanation. In the cognitive condition: The decision of the AI must be in line with the participant's decision (A or B). The explanation of the AI has to be similar to the participants' explanation. Therefore, the wizard chose the explanation closest to the user's explanation.

Appendix B. Chatbot

Screenshot of the interaction between a participant and the chatbot (translated)



The screenshot displays a chatbot interface with a blue header bar. On the left, the 'stinto' logo is visible. On the right, a sidebar indicates '2 Teilnehmer' (2 participants), listing 'Bot' and 'Participant'. The main chat area contains the following text:

Willkommen! Die Adresse dieses Chats lautet: <https://stin.to/asgi6>. Sende diese an andere, um Sie einzuladen. Gib */hilfe* für eine Kurzanleitung ein.

Participant ist dem Chat beigetreten.

Bot Hello, I am Stinto. Thank you for activating me so that I can support you in your decisions. I'm programmed to help you make decisions in your everyday life and work. I can access a wealth of data, aggregate the information in a short time and then give my decision. Just try it out. What decision are you in right now? Please enter the decision abbreviation. You can find it in the survey tool. This way I know which decision you are facing.

Participant Start up

Bot What have you decided: Option A or Option B?

Participant B

Bot Why did you choose B?

Participant Because I am not taking a risk

At the bottom of the chat area, there is a text input field and a row of icons for various functions: a paper plane (send), a smiley face (emojis), a camera, a pushpin, a moon (night mode), a speaker (audio), a lock, and a power button.