Please quote as: Schmitt, A.; Wambsganss, T.; Söllner, M.; Janson, A. (2021): Towards a Trust Reliance Paradox? Exploring the Gap Between Perceived Trust in and Reliance on Algorithmic Advice. International Conference on Information Systems (ICIS).

# Towards a Trust Reliance Paradox? Exploring the Gap Between Perceived Trust in and Reliance on Algorithmic Advice

*Completed Research Paper*

**Anuschka Schmitt**
University of St.Gallen,
St.Gallen, Switzerland
anuschka.schmitt@unisg.ch

**Thiemo Wambsganss**
University of St.Gallen,
St.Gallen, Switzerland
Carnegie Mellon University,
Pittsburgh, USA
thiemo.wambsganss@unisg.ch

**Matthias Söllner**
University of Kassel,
Kassel, Germany
soellner@uni-kassel.de

**Andreas Janson**
University of St.Gallen,
St.Gallen, Switzerland
andreas.janson@unisg.ch

## Abstract

*Beyond AI-based systems' potential to augment decision-making, reduce organizational resources, and counter human biases, unintended consequences of such systems have been largely neglected so far. Researchers are undecided on whether erroneous advice acts as an impediment to system use or is blindly relied upon. As part of an experimental study, we turn towards the impact of incorrect system advice and how to design for failure-prone AI. In an experiment with 156 subjects we find that, although incorrect algorithmic advice is trusted less, users adapt their answers to a system's incorrect recommendations. While transparency on a system's accuracy levels fosters trust and reliance in the context of incorrect advice, an opposite effect is found for users exposed to correct advice. Our findings point towards a paradoxical gap between stated trust and actual behavior. Furthermore, transparency mechanisms should be deployed with caution as their effectiveness is intertwined with system performance.*

**Keywords:** Decision-making, artificial intelligence, trust, advice taking, transparency

## Introduction

Due to the nature of their statistical architecture, Artificial Intelligence (AI)-based Information System (IS) are suffering from opaqueness, limited robustness and reliability (Jordan and Mitchell 2015; Rahwan et al. 2019). State-of-the-art IS, such as conversational interfaces (CI) or decision support systems (DSS), offer outcomes that (1) are often not fully predictable nor explainable and, (2) most critically, bring along a certain probability of inaccuracy. For example, a skin cancer detection model with a classification accuracy of 98% might lead to 2% of all users receiving erroneous diagnoses. Against the backdrop of these challenges, trust plays an important role in understanding the adoption and use of AI-based IS. Their prevalence across a variety of high-impact use cases such as credit scoring (O'Neil 2016), predictive policing (Waardenburg et al. 2018), large-scale adaptive education (Wambsganss, Niklaus, et al. 2021) or cancer detection (Jussupow et al. 2021) illustrates the capabilities of AI-based IS to augment human decision-making and to even overcome human biases (Faltings et al. 2014; Rahwan et al. 2019). At the same time, lack of trust towards such systems' recommendations can impede successful adoption and deployment (Söllner et al. 2016).

Existing research has named algorithmic error as a key factor for explaining decreased trust in algorithmic advice (Dzindolet et al. 2002; Hoff and Bashir 2015). The desire for perfect forecasts has been cited as a prominent reason for reluctance to rely on imperfect algorithms, which may be superior to human advice, though (Dawes 1979). Yet, Liel and Zalmanson (2020) found that people adhered to erroneous AI, suggesting that such recommendations have strong persuasive power regardless of their correctness. However, little is known about how erroneous recommendations influence human behavior or perception (i.e., the 2% wrongly predicted outcomes).

Most empirical work has black boxed the nature of the AI-based system and neglected the correctness of its output when investigating related user outcomes (Haibe-Kains et al. 2020). Scholars have mostly discussed whether algorithmic advice is over- or underutilized in comparison to human advice (Logg et al. 2019). In fact, current literature falls short on three perspectives: First, previous studies have integrated the manipulation of source of advice (human versus machine) within experiment surveys following a Wizard-of-Oz approach, thereby not depicting a natural, contemporary human-computer interaction (Liel and Zalmanson 2020). The question arises of whether stating that the source of advice stems from an algorithm suffices as an instantiation of algorithmic advice. Second, studies comparing algorithmic advice with human advice often neglect under which circumstances algorithmic advice is to be trusted and relied upon beyond preference over human advice. Third, current research barely distinguishes between notions of trust and related, behavioral outcomes. At the same time, empirical studies in human-computer interaction and IS often define and use typologies of trust interchangeably (Yin et al. 2019), ultimately leaving unaddressed how perceptions of a system's functioning and behavioral response relate to each other.

Overall, researchers have called for "[…] a broader consideration of the behavior [… and the…] unintended consequences of AI" (Rahwan et al. 2019, p. 477). More specifically, in the domain of IS, incorrect system advice has been assessed as highly problematic, yet insufficiently addressed until now (Jussupow et al. 2021). In this study, we aim to arrive at a more nuanced consideration of both the antecedents (the underlying technological system and the correctness of advice) and the related consequences of AI-based advice (perceptual and behavioral outcomes). If error in algorithmic advice leads to a mis- or disuse of AI-based systems, the question arises of how to design for interaction with failure-prone IS. So far, research examining how specific design features alleviate potential over- or underreliance on algorithmic advice is rather scarce (Berger et al. 2021; Yeomans et al. 2019). We aim to address this gap and contribute to research around reliance on AI-based advice by answering the following research question (RQ):

***RQ:*** *What is the effect of incorrect algorithmic advice on user trust and reliance and how does transparency around an AI-based system's accuracy levels alter the effects of such incorrect advice?*

To answer our research question, we conducted a 2 x 2 between-subject experiment to test whether correctness of recommendation (correct vs. incorrect) and transparency (statement on accuracy of AI present vs. absent) result in higher levels of trust in the information and reliance on AI-based advice. The manipulations have been embedded in four instantiations of a text-based conversational agent (CA) to provide the user with a familiar and recognizable AI-based IS. In the context of a vignette study, users participated in an assessment center as part of which they had to complete a reading comprehension task in interaction with the CA. We found that participants being exposed to incorrect advice trusted provided information less, yet more strongly relied on the recommended answers. Our study suggests that adding transparency around a system's accuracy levels does not necessarily increase trust in the system's advice and depends on the correctness of advice.

Our research contributes to the understanding of whether users trust AI-based advice and whether such an acceptance is influenced by the system's accurateness in the specific context of a CA. We provide a more nuanced understanding of perceptual and behavioral outcomes by differentiating between trust and reliance towards the AI-based advice. By shedding light on these phenomena and empirically testing proposed strategies for AI trustworthiness, we hope to contribute to the existing research body around algorithmic aversion and appreciation. We first provide an overview of the conceptual background of this paper and developed hypotheses. The subsequent section describes the chosen methodology, including the experimental design manipulations and measurements, as well as the analysis of our constructs. Afterwards, we present and discuss the results, followed by the conclusion of this paper. The last section provides the contribution and limitations of this study, as well as an outlook on future research.

# Conceptual Background and Hypotheses

As part of this section, we provide an overview of related work on AI-based decision making, the notion of trust in this context, and transparency as a means to increase trustworthiness. On that basis, we develop hypotheses to be tested as part of our conceptual research model.

## *AI-Based Decision Making*

Fueled by continuous advancements in machine learning (ML), AI- based IS have led to a paradigm shift in the way information technology is designed and deployed (Knote et al. 2021). Encompassing various paradigms, such as algorithm, decision aid, expert systems, decision support systems, AI can be defined as "a system's ability to interpret external data correctly, to learn from such data, and to use those learnings to achieve specific goals and tasks through flexible adaption" (Haenlein and Kaplan 2019, p. 15). AI-based IS can thus make decisions without explicit human influence, thereby promising to reduce time, effort and costs associated with resource- and staff-intensive decision-making (Parasuraman and Riley 1997). At the same time, AI-advised human decision making can also allow users to take recommendations from an AI partner for solving a task (Zhang et al. 2020).While AI-based IS vary in their decision-making latitude, we understand AI-based IS and their role in decision making according to Baird and Maruping's (2021, p. 318) proposition of prescriptive agents that "act as substitutes for either behavior-based decision-making or outcome-based decision-making by prescribing or taking actions.". Despite AI-based IS' potential, such systems can introduce unintended and costly implications for user and business. Previously raised issues around uncertainty, accuracy and reliability can impede IS use and become apparent especially once ML models are deployed in real-world domains (Baird and Maruping 2021; D'Amour et al. 2020; Köchling et al. 2020).

Research is incongruent on if and when machine-made judgement is accepted and relied upon. Two dominant research streams, namely algorithmic aversion and algorithmic appreciation, propose two conflicting conclusions regarding the reliance on algorithmic advice. While algorithmic aversion follows the notion that human decision makers rather rely on human advice despite algorithmic advice being more accurate (Dietvorst et al. 2015), algorithmic appreciation claims that users rather rely on advice stated to come from an algorithmic source, as compared to a human (Logg et al. 2019). According to algorithmic aversion, humans can be unwilling to accept advice from AI-based systems, even if such ML-based models are proven to outperform humans in complex decision tasks (Elkins and Derrick 2013; Kuncel et al. 2013) or simple judgement tasks (Castelo et al. 2019). The existing research body has identified numerous reasons for decreased trust in algorithmic advice, including the desire for perfect prediction (Dawes 1979), human confidence in own reasoning (Whitecotton 1996), the difficulty of the decision task (Castelo et al. 2019), as well as the error rate of the algorithmic advice. This study specifically addresses the notion of erroneous advice. Based on the assumption that humans expect algorithms to work perfectly, coined as the perfection schema, errors are perceived as particularly negative because they are unexpected. Existing studies have suggested that even if algorithmic advice is erroneous only occasionally, humans overestimate the perceived error rate (Dzindolet et al. 2002; Hoff and Bashir 2015). In a similar vein, Dietvorst et al. (2015) find that perceptions of trust in the advice already decrease when humans see an AI-based system make small mistakes.

While most studies support the notion of algorithmic aversion, a more nascent research stream observed an (exaggerated) appreciation of algorithmic advice. Algorithmic appreciation has been found in time critical situations (Liel and Zalmanson 2020; Robinette et al. 2017) and has been explained by humans attributing more objectivity and rationality to algorithmic advice as compared to human judgement (Dijkstra et al. 2017). Previous research has pointed towards several design incentives to enhance the acceptance of AI-based advice, such as allowing users to modify or participate in the algorithmic process or to increase transparency around the algorithm. Yeomans et al. (2019), for instance, have found that users are more likely to rely on algorithmic advice when they are able to understand how the AI-based systems works. Interestingly, studies have also shown that humans erroneously rely on algorithms, i.e., when trust placed in the algorithmic advice actually exceeds an AI-based system's real capabilities (Lee and See 2004). More so, (Liel and Zalmanson 2020) found that gig platform workers conformed with erroneous recommendations which were stated to stem from an algorithmic source, even when the correct task outcome was relatively certain to judge without any advice.

Reliance on algorithmic advice has been studied in medical, economic, and legal contexts, as well as for subjective decisions (i.e., recommending a joke). These contexts are predominantly defined by high uncertainty regarding the decision outcome, user-specific influences, including domain expertise, or potential confounding factors such as delegation of decision responsibility. At the same time, little is known about how algorithmic advice performs in other contexts where costs of system error are high for the user, yet task uncertainty is limited. Liel and Zalmanson's (2020) paper was one of the few which investigated the effect of erroneous AI on advice reliance in the context of a simple judgement task where algorithmic mistakes were quite apparent to the user. Whereas previous research has investigated the preference and reliance between algorithmic and human advice, little is known about when algorithmic advice is over- or underutilized.

### *Trust in AI-Based Systems*

When turning towards the lack of predictability, control and complexity associated with the interaction with AI-based systems, one is quick to arrive at the necessity of trust in such an interaction. According to established trust theories, trust can be described as a human reaction to reduce complexity although an undesirable outcome is possible (Luhmann 1979). In an AI-mediated decision making context, hence, users presume a favorable "behavior" of the IS despite the uncertainty of the IS providing erroneous recommendations (Mayer et al. 1995). Recognizing the existence of algorithmic error as a point of departure, trust plays a crucial role in the (dis)use of algorithms (Lee and See 2004). If we recognize the existence of erroneous suggestions, including its unintended and costly implications (Köchling et al. 2020), trust can be viewed as a function of system performance and users' understanding thereof.

Literature generally offers mixed results not only regarding the preferences for source of advice, yet also regarding the implications of erroneous algorithmic advice. Whereas research on algorithmic aversion predicts a general distrust in AI-based systems, we suggest that perceptions of trust in and reliance on advice will be influenced by the correctness of the algorithmic advice. Recent studies have demonstrated that users conform with incorrect algorithmic advice (Lee and See 2004; Liel and Zalmanson 2020), which can be ascribed to users failing to recognize the erroneous recommendation (Endsley 2017; Goddard et al. 2012; Parasuraman and Manzey 2010). At the same time, the predominant literature stream on algorithmic aversion with erroneous recommendations shows that minor indications of erroneous algorithmic advice already leads to a decrease in trust (Dietvorst et al. 2015; Dzindolet et al. 2002). Considering a decision task defined by an objectively measurable outcome, we expect participants to be able to detect incorrect advice. Accordingly, we hypothesize:

***H1a:*** *Incorrect compared to correct algorithmic advice leads to lower levels of trust in the information provided by the CA.*

Interestingly, literature on algorithmic aversion and appreciation uses the term trust interchangeably with related, yet to be distinguished concepts, such as acceptance of, reliance on or conformity to advice. Yin et al. (2019), for instance, measure both perceptual, self-reported levels of trust, as well as conformity to algorithmic advice as a quantifiable, behavioral outcome. In a similar vein, Dietvorst et al. (2015) consider trusting beliefs as a mediator for behavioral outcomes. Robinette et al. (2017) consider trust as a binary, self-reported outcome when exploring the impact of robot performance in a time critical decision-making context. When discussing behavioral outcomes, the "complicity of IS artifacts in goal achievement (or failure)" (Baird and Maruping 2021, p. 316) and thus properties of the AI-based IS are oftentimes neglected. We therefore distinguish between trust in provided information as a perceptual outcome and reliance on advice as an additional, behavioral key outcome variable, following Yin et al.'s (2019) reasoning. More so, the conceptual framework by Lee and See (2004) offers a theoretical foundation to conceptually differentiate between trust as an attitude and distinct behavioral outcomes, such as reliance. We thus state in a subsequent hypothesis that:

***H1b:*** *Incorrect compared to correct algorithmic advice leads to decreased reliance on algorithmic advice provided by the CA.*

### *Transparency around AI-Based System Performance*

While performance and accuracy of AI-based IS vary from system to system and are dependent on numerous factors (i.e., quantity and quality, as well as labelling of data), the nature of ML-based systems

per default introduce outcomes that are not fully predictable nor explainable, and, most critically, bring along a certain probability of inaccuracy. Improving accuracy rates and model performance in practice can be achieved by building high quality datasets, however, the improvement of accuracy rates is a trade-off between resources and performance, with users still going to be exposed to erroneous recommendations in the near future (Rahwan et al., 2021). If experiencing an AI-based IS providing potentially erroneous information negatively affects perceptions of trust and reliance on advice, the question arises of how trust can be restored in the case of an erring AI-based system.

Different mechanisms have been proposed to overcome algorithmic aversion and thus increase trust in AI-based systems. Yeomans et al. (2019) gather evidence on how explanations on the underlying workings of an algorithm decreases algorithmic aversion. In a similar vein, Berger et al. (2021) find that demonstrating an AI-based system's ability to learn also acts as a countermeasure for algorithmic aversion. They argue that an increased understanding of how systems work leads users to assign increased capabilities to the AI-based IS, and thus trust it more heavily. More so, Berger et al. (2021) suggest exploring transparency as a further moderator to consider when studying the impact of erroneous algorithmic advice. Beyond their papers, however, measures to enhance trust in algorithmic advice have barely been explored.

One possible solution to increase trustworthiness is the revelation of the system's accuracy to the user. As part of the Ethics Guidelines for Trustworthy AI proposed by the European Commission, transparency has been mentioned as one of the key dimensions to establish trustworthiness around AI-based systems. More specifically, the Ethics Guidelines propose that "[…] the AI system's capabilities and limitations should be communicated to AI practitioners or end-users in a manner appropriate to the use case at hand. This could encompass communication of the AI system's level of accuracy." (European Commission 2019). Returning to Lee and See's (2004) conceptual model on trust and reliance, information on the IS' performance is considered as a relevant factor influencing trust and related behavior. Based on the reviewed studies and the notion of performance information allowing to potentially reduce algorithmic aversion, we thus propose that:

***H2a:*** *Transparency on a CA's stated accuracy levels will alleviate the stated effect of incorrect algorithmic advice on trust in the information provided by the CA.*

In a similar vein, if algorithmic output contradicts human experiences and intuitions, increased transparency and interpretable explanations gain in importance to avert resistance and aversion towards algorithmic advice. Subsequently, we propose that:

***H2b:*** *Transparency on a CA's stated accuracy levels will induce users receiving incorrect algorithmic advice conform more heavily to the advice.*

Overall, we suggest that the correctness of algorithmic advice has an effect on user trust in and reliance on the provided information (H1). In addition, making the stated accuracy levels of the AI-based system transparent to the user moderates the effect of advice correctness on user perceptions and behavior, thereby resolving algorithmic aversion in erroneous advice (H2). Our research model is illustrated in Figure 1.

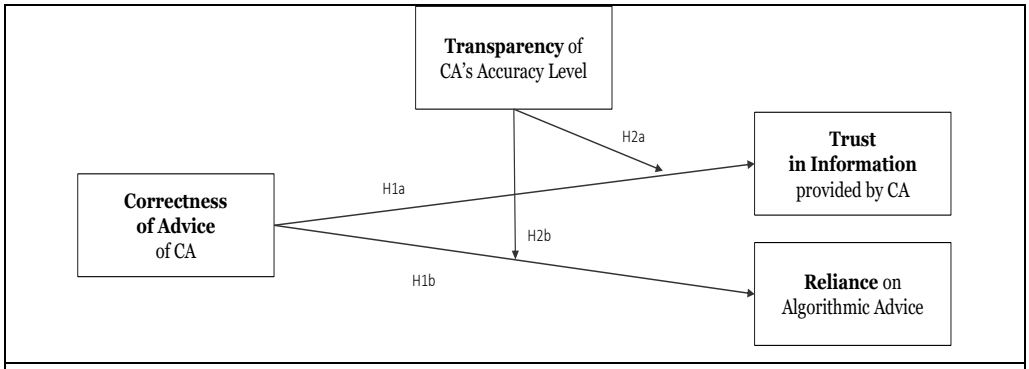


**Figure 1. Conceptual Research Model**

# Research Methodology

To test our conceptual model and related hypotheses, we conducted an online experiment to investigate (1) the *effect of incorrect algorithmic advice on related user perceptions and behavior and (2) if transparency on an AI-based system's stated accuracy levels alter the effects of such an incorrect advic*e. We relied on a 2 (correct vs incorrect advice) x 2 (no transparency cue vs transparency cue) between-subject design resulting in one control group (CG) and three treatment groups (TG). The participants were randomly assigned to one of the four groups. For the AI-based system, we implemented and manipulated an intent-based CA built with the *Dialogflow* framework. The web experiment was facilitated by the behavioral lab of our university.

## *Experimental Design and Procedure*

To test our hypotheses in a natural setting, we chose to design a virtual assessment center scenario where participants were asked to conduct a reading comprehension task. The type of the task fitted well with the selected participant pool which encompassed late-stage graduate students. Students were provided a text on a conversation between two teachers arguing about student violence at high schools (Flender et al. 1999). Based on this argumentation text, the participants' task was to complete four multiple choice questions testing the comprehension of the individual arguments raised in the text. We chose this task since reading comprehension tasks are often deployed in early-stage job application assessments and commonly used in human-computer interaction experiments (Fox et al. 2009; Wambsganss et al. 2020; Wambsganss, Kung, et al. 2021). At the same time, the chosen task exhibits a risk-present context where costs of system error are high for the user. In that sense, providing false answers to the multiple-choice questions can lead to a worsened performance in the job assessment. We are convinced that this type of application and task is representative for a broader class of AI-based applications where 1) the user has to make use of existing or recently gained knowledge and can thus assess the information provided by the AI to certain extent, and 2) other learning and job-related settings and tasks where the consideration of the AI-based advice has a direct and personal impact for the user. Examples include other tasks in education settings where AI-advised human decision making is directly tied to student's evaluation (i.e., grading) or training tasks in job-related settings where AI-advised human decision making is tied to an employee's assessment (i.e., being hired or promoted).

To imitate a real-life situation, we asked participants to conduct the survey on their laptop or a comparable larger screen in a quiet environment, similar as they would conduct a virtual assessment center for a real job. The students were compensated with credits as part of a university course. The design of the CA as well as the manipulations will be further described in the section below. On average, the experiment took 15 to 20 minutes. It consisted of three main phases: *1) A Pre-Test phase, 2) an Experiment phase* and a 3) *Post-Test phase*. The *Pre-* and *Post-phases* were consistent for all participants. In the *experiment phase*, we manipulated the advice the CA gave to users regarding the comprehension questions (correct vs incorrect advice), as well as the transparency cue communicating the CA's accuracy levels (transparency cue present vs transparency cue absent).

**1) Pre-Test:** In the first step, participants were informed about the reading comprehension task, asked to answer ten pre-survey questions collecting control items, and introduced to the experimental section. Participants were asked to imagine being in a job application process and being invited to a virtual assessment center of a company they have applied for. The participants received an explanation of key argumentation instruments and examples of each instrument. After this information, a training question on argumentation was implemented to ensure that participants could become acquainted with the topic of argumentation and studied the presented text attentively before engaging in the interaction with the AI-based system. An additional link compromising the reading comprehension text, an explanation and examples of argumentation theory, as well as all comprehension questions to be answered, was provided to allow participants to come back to the text at any time.

**2) Experiment:** In the experiment phase, we asked participants to imagine being in a job application process and being invited to a virtual assessment of a company they have applied for. As part of the case, participants had to complete a reading comprehension task. The task is part of the construct of passive argumentative competency following the design of Flender et al.'s (1999) argumentation task as a proven

construct to measure argumentative competencies. Participants were asked to read a discussion of two teachers concerning the topic "Does TV make students aggressive?".

As part of the task, students had to answer certain questions about the argumentation strategy of the involved parties in the discussion. After two control questions to ensure the past understanding of the task, we asked the participant to interact with our CA. Participants were asked to answer four multiple choice questions with the help of a CA which would guide them and provide suggestions to the individual questions. The CA announced the sequence of questions and asked users for their initial answer for each multiple-choice question. Subsequently, the agent provided its recommended answer to the respective multiple-choice question and allowed participants to revise their initial answer. After users communicated their final answer, the CA moved on to the next question. This was repeated for all four multiple choice questions. To ensure a conscious interaction with the CA and increase critical judgement, we set a mandatory minimum time of three minutes to spend on the four questions (Liel and Zalmanson 2020). A countdown indicated the remaining time. Participants were able to continue with the experiment after the countdown was finished. The control group used a CA providing correct advice without a transparency cue. TG1 used a CA providing incorrect advice without a transparency cue, whereas TG2 (correct advice) and TG3 (incorrect advice) used a CA that incorporated a statement on the accuracy level of the CA. We did not provide any further introduction to any of the tools.

**3) Post-Test:** In the last step, participants completed a post-experiment questionnaire. In total, we asked 30 items to assess participants' perception regarding the advice and the conversational agent itself, and to control for manipulation. Moreover, we asked three qualitative questions: "What did you particularly like about the use of the chatbot?", "What else could be improved?" and "Do you have any other ideas?". Finally, we captured the demographics. Final information informed participants about the actual purpose of the study and provided them with contact details of the researching team.

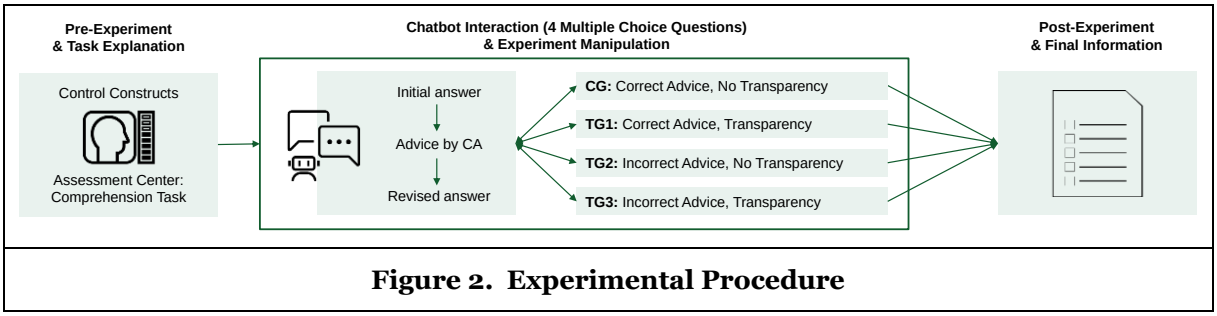


**Figure 2. Experimental Procedure**

## *Design and Manipulation of an AI-Based CA for Task Advice*

Participants were randomly assigned to a CA providing correct versus incorrect advice on the assessment center task. In the correct recommendation condition, participants were suggested the correct answer to a respective multiple-choice question after submitting their initial, own answer. Accordingly, participants in the erroneous advice condition were suggested incorrect answers to the multiple-choice questions. Next to the correctness of the advice (correct vs incorrect advice), we manipulated the presence of a transparency cue on the CA's accuracy levels (no transparency statement cue vs transparency statement cue), ultimately resulting in four artifacts.

Interaction flow and communication style were inspired by existing agent-based interactions and dialogues found in similar research (Zierau et al. 2020). To avoid any confounding elements, for instance caused by the level of anthropomorphism or formality of language, the style and appearance of the text-based CA was kept as simple and static as possible. For the humanic interaction cues, we chose to not embed our CA in an embodied personification to ensure that perception measures are not influenced, e.g., by the choice of the avatar. Moreover, we kept the interaction style transactional and thus, language simple and design functional. Our aim was not to design a humanic CA interaction, but rather a functional conversational AI-based IS, to investigate our hypotheses without potential confounds due to the interaction design.

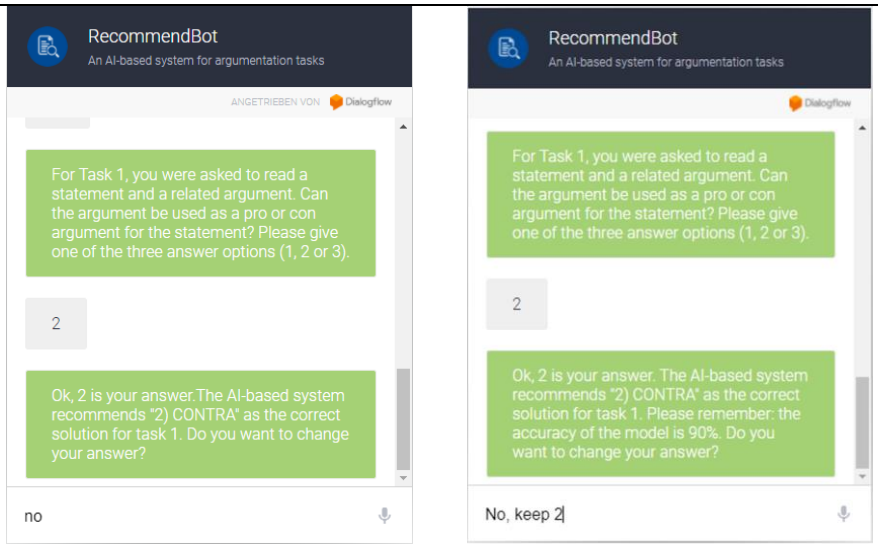


**Figure 3. Correct CA Advice with and without Transparency Cue**

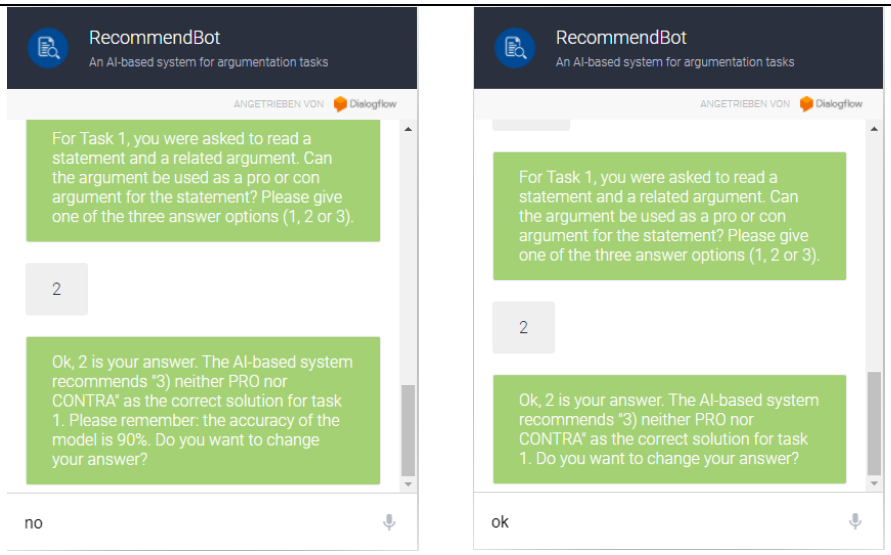


**Figure 4. Incorrect CA Advice with and without Transparency Cue**

To understand the influence of transparency, we additionally tested the presence of performance information on the AI system's level of accuracy as a moderating effect of advice correctness. For participants in the transparency treatments who received either erroneous or accurate recommendations, the CA stated its accuracy level several times throughout the interaction. As part of the first point of interaction, the CA disclosed its level of accuracy and the related implication of potentially incorrect recommendations. In addition, the CA reminded the user of its accuracy level after each multiple-choice answer suggestion. Participants were randomly assigned to one of the four versions of the interaction. The content and sequence of survey questions, as well as experiment task-specific questions, were held constant across all four conditions.

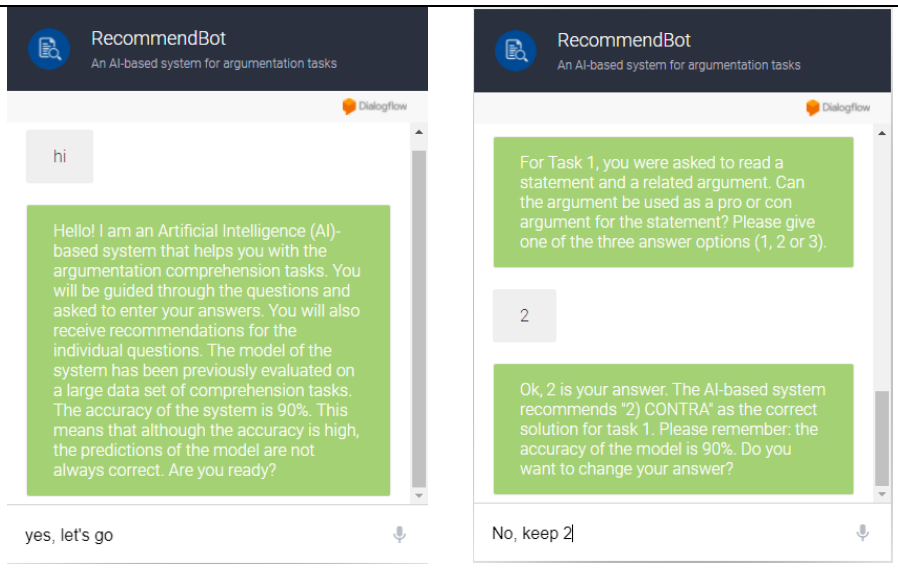


**Figure 5. CA Transparency Manipulations**

## *Measures*

The measurement of our behavioral dependent variable was part of the multiple-choice reading comprehension in our experiment. Previous studies on algorithmic advice have relied upon several different instruments to measure the influence of advice given on the user's decision making. Since our experiment task encompassed a binary outcome possibility, we followed Liel and Zalmanson (2020, p. 4), who also investigated the effect of erroneous recommendations, and their measure of conformity, which they defined as "the act of choosing the incorrect answer recommended by the algorithm.". This is congruent with alternative operationalizations of reliance used in related empirical research including Zhang et al.'s (2020) switch percentage measure. Hence, reliance was measured as the proportion of participants who adjusted their final answers according to the CA's suggestions at least once. Reliance on advice therefore captured if participants "followed" the AI-based advice and was quantified as 1 if a participant's initial prediction disagreed with the advice and the participant's final prediction agreed with the AI-based advice. We also measured the frequency of adjusted answers according to the advice for each treatment group and whether participants switched their answers regardless of the CA's recommendation. Another key measurement is self-reported trust in information provided by the CA. We measured trust in information with three items (scale adapted from McKnight et al., 2020, sample item: "To me, the chatbot is generally accurate in providing recommendations"; 7-point scale, from 1: "Strongly Disagree" to 7: "Strongly Agree", $\alpha_{TrustInformation}$= .91).

Besides our two key dependent variables, we measured multiple control and demographic variables in both the pre- and post-experimental questionnaire, including participants' trusting disposition (Gefen and Straub 2004), personal innovativeness (Agarwal and Prasad 1998), cognitive rigidity (Lewicki and Bunker 2012), Big Five personality constructs (Rammstedt and John 2007), felt risk of technology (scale adapted from McKnight et al. 2020, sample item: "How would you characterize the decision of whether to use the chatbot as a decision aid"; 7-point Likert scale , from 1: "Strongly Disagree" to 7: "Strongly Agree", $\alpha_{FeltRisk}$= .85) and algorithmic familiarity (scale adapted from Johnson and Russo 1984).

## *Data Collection and Cleaning*

To check for validity and reliability, five people who were uninvolved in the research completed the questionnaire in a pre-test. One of those pre-testers was a Native English speaker who also proofread the questionnaire. Recommendations and preliminary results from this pre-test (i.e., shorten minimum completion time for multiple choice questions) were incorporated before distributing the survey on a large scale online.

A total of 198 participants were recruited from a university course at a higher-level education facility and randomly assigned to the two-by-two cell between subject experiment. Attentive participation was incentivized by making the content of the experiment part of the students' final course assessment. We removed those subjects who failed the attention check or did not complete the interaction with the AI-based agent by providing complete answers to all four multiple choice questions. We further removed participants who exhibited outlier characteristics regarding, i.e., completion time, leaving us with a final sample set of 156 subjects. A potential boundary represents the survey distribution within the course of master students who most probably exhibit higher literary and familiarity with AI-based systems as compared to the general public. Additional analyses on the control and demographic variables comparing the experiment groups confirm participants' random assignment to the different experimental conditions. Specifically, there are no significant differences in trusting disposition, personal innovativeness, or cognitive rigidity among the four treatments (all $p > .1$). In addition, no differences were found regarding the demographic variables age and gender (all $p > .1$).

# Results

We used a manipulation check for the transparency treatment, asking to what extent participants agree with the following two statements: 1) "As part of the tasks, the chatbot revealed information about itself, namely why its recommendations might be flawed.", and, 2), "As part of the tasks, the chatbot revealed information about itself, namely an explanation of the legal guidelines that the chatbot must adhere to." (7-point Likert scale, from 1: "Does not apply at all" to 7: "Applies completely"). The results of the manipulation checks indicated that the transparency manipulation worked as intended: First, an ANOVA on the first statement revealed a significant manipulation effect ($F = 66.62$, $p < .001$) with participants from the two treatment groups receiving transparency statements (Accurate, Transparency; Erroneous, Transparency) exhibiting a significant higher confirmation of the first statement ($M_{Transparency} = 4.32$) than participants who were not exposed to the transparency-enhancing statements ($M_{NoTransparency} = 2.02$). Another ANOVA on the second statement strengthened this finding as no significant effect between the transparency present ($M_{Transparency} = 1.96$) versus transparency absent ($M_{NoTransparency} = 1.86$) treatments could be found. We tested our hypotheses by conducting a series of analyses in R Studio.

| **Group** | **N** | Trust in Information (Mean) | Reliance on Algorithmic Advice (%) | Correct Final Answer for each Multiple Choice Question (%) | | | | Age (in years) | Gender (male) (%) |
|---|---|---|---|---|---|---|---|---|---|
| | | | | MC 1 | MC 2 | MC 3 | MC 4 | | |
| **CG:** Correct Advice, No Transparency Statements | 40 | 5.23 | 15% | 100% | 97.5% | 97.5% | 97.5% | 24.34 | 77.5% |
| **TG1:** Correct Advice, Transparency Statements | 40 | 4.67 | 7.5% | 97.5% | 92.5% | 97.5% | 97.5% | 24.35 | 70% |
| **TG2:** Incorrect Advice, No Transparency Statements | 41 | 2.33 | 22% | 80.5% | 82.9% | 82.9% | 95.1% | 25.32 | 68.3% |
| **TG3:** Incorrect Advice, Transparency Statements | 35 | 3.06 | 31.4% | 68.6% | 80% | 85.7% | 97.1% | 24.34 | 62.9% |
| **Correctness Manipulation** | | *** $p < .001$ | ** $p < .01$ | *** $p < .001$ | *** $p < .001$ | *** $p < .001$ | ns | NA | |
| **Transparency Moderation** | | ** $p < .01$ | * $p < .05$ | ns | ns | ns | ns | | |

**Table 1. Means for Perceptual and Behavioral Outcome Variables across Four Groups**

For Hypothesis 1a – the effect of advice correctness on trust in the provided information – we conducted an ANOVA comparing participants in the two conditions receiving correct advice with participants receiving incorrect advice. The test revealed a significant main effect ($F(1, 156) = 143$, $p < .001$). In fact, participants who received incorrect recommendations perceived the information the CA provided as less trustworthy ($M_{Incorrect} = 2.78$) as compared to those who received correct recommendations to the multiple-choice answers ($M_{Correct} = 4.95$).

Hypothesis 1b predicted a negative impact of the correctness of CA on reliance on advice. A chi-square test of independence revealed a statistically significant association between the correctness of the AI-based advice and reliance on advice, $X^2$ (1, N = 156) = 4.89, $p < .01$. As such, H1b is not supported in that participants who received an incorrect advice by the CA significantly relied more heavily on those recommendations as compared to participants who received correct algorithmic advice.

The results of our moderation analysis (see Fig. 5) show that a significant interaction effect between advice correctness and transparency exists for trust in information ($F(3, 156) = 8.76$, $p < .001$). Pairwise t-tests reveal significant mean differences between all treatment groups. The pairwise comparison between the correct, no transparency ($M_{Correct,\ No\ Transparency} = 5.23$) and the erroneous, no transparency ($M_{Incorrect,\ No\ Transparency} = 2.54$) treatments ($p < .001$), as well as that between the correct, no transparency ($M_{Correct,\ No\ Transparency} = 5.23$) and the erroneous, transparency ($M_{Incorrect,\ Transparency} = 3.06$) treatment ($p < .001$) are significant. In a similar vein, the pairwise comparison between the correct, transparency ($M_{Correct,\ Transparency} = 4.67$) and the erroneous, no transparency ($M_{Incorrect,\ No\ Transparency} = 2.54$) treatments ($p < .001$), as well as that between the correct, transparency ($M_{Correct,\ Transparency} = 4.67$) and the incorrect, transparency ($M_{Incorrect,\ Transparency} = 3.06$) treatment ($p < .001$) are significant. Last, we found that participants receiving correct advice without any transparency statements reported significantly higher levels of trust in information as opposed to participants who also received correct advice yet including transparency statements ($p < .05$). At the same time, participants receiving incorrect advice without any transparency statements reported significantly lower levels of trust in information as opposed to participants who also received correct advice yet including transparency statements ($p < .05$). These results provide support for H2a.

Replicating the moderation effect of transparency on reliance on algorithmic advice, a chi-square test of independence shows that transparency moderates the effect of recommendation correctness on reliance on the recommended answers, $X^2$ (3, N = 156) = 7.71, $p < .05$. Participants receiving correct advice and transparency statements adjusted their answers according to the system's advice significantly less than expected. At the same time, participants receiving incorrect advice and transparency statements adjusted their final answers in line with the CA's advice significantly more than expected. While hypothesis 2b can be confirmed, the results for the accurate advice treatment groups exhibit unexpected results.

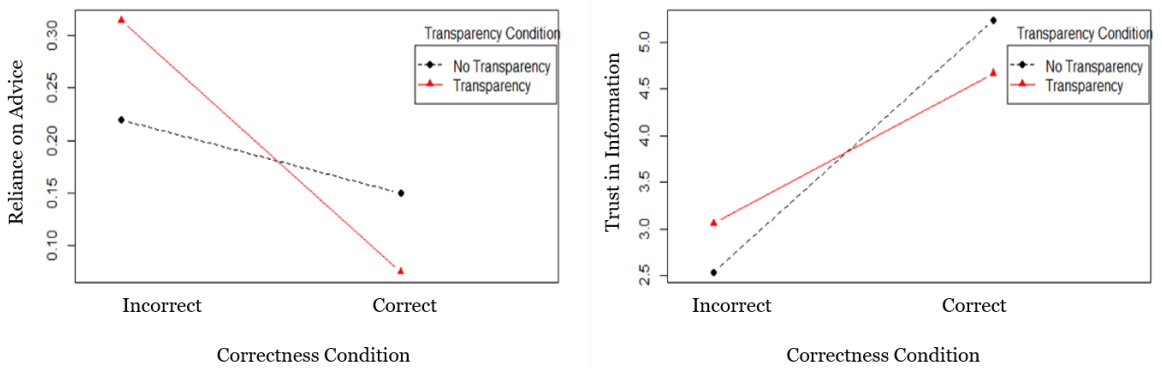


**Figure 6. Interaction Plots for Reliance on Advice and Trust in Information**

## Discussion

AI-supported decision making has been heavily discussed at the forefront of public attention and has drawn researchers from different fields alike to arrive at contrasting conclusions regarding the trust decision-makers place in algorithmic advice. As part of the current study, we aimed to further clarify under which conditions algorithmic advice is trusted and relied upon. Past research is incongruent on the influence of algorithmic judgement, as predominant literature streams on algorithmic aversion and algorithmic appreciation show. Furthermore, we sought to explore whether making an algorithm's accuracy level transparent to the user may influence perceptions and reliance on advice, given that transparency is oftentimes mentioned as a design mechanism to increase trust in IS yet can also increase cognitive load of the user. We studied these questions by simulating a reading comprehension task within a job assessment setting. Our experiment setting implied a limited set of possible answer options and algorithmic advice was embodied in a CA's recommendations. While this experiment design attempted to resemble a realistic scenario, these boundary conditions should be considered when comparing the current study's results with similar research. Earlier studies predominantly focused on forecasting and estimation task, as well as did not embody algorithmic advice as part of a technological artefact. Ultimately, we strived to arrive at a more nuanced understanding of how perceptions and behavior related to algorithmic advice differ depending on the correctness of the advice, as well as how transparency on the system's accuracy levels moderates such perceptions and behavior.

According to our results (see Table 2), users generally perceive incorrect algorithmic advice as less trustworthy than correct advice. Our findings thereby support the claim that performance of AI-based systems matters (Liel and Zalmanson 2020). At the same time, our analyses indicate that users receiving incorrect algorithmic advice more heavily rely on this advice. Our findings thus diverge from existing notions that humans resist statistical models as they trust more their own intuition (Highhouse 2008). Our behavioral results are in line with Liel and Zalmanson's (2020) research who find that algorithmic recommendations have strong persuasive power in leading users to adapt their behavior and judgement according to incorrect algorithmic advice. While our findings converge with the literature stream around algorithmic appreciation, stating that algorithmic advice is often trusted beyond the algorithm's actual capabilities (Logg et al. 2019), they point towards a novel, previously neglected pattern, namely the relation between perceived trust in the algorithmic advice and behavioral reliance on the recommended answers. Notions such as the personalization privacy paradox (Kehr et al. 2015; Li et al. 2017) have already discussed the gap between attitude and behavior in related domains and in IS. However, no such gap or paradox has been studied in the context of algorithmic acceptance or algorithmic aversion before.

| Hypothesis | Key Findings | Results |
|---|---|---|
| ***H1a:*** *Incorrect, as compared to correct, algorithmic advice leads to lower levels of trust in the information provided by the CA.* | Incorrect algorithmic advice leads to lower levels of trust in provided information. | *** $p < .001$ |
| ***H1b:*** *Incorrect, as compared to correct, algorithmic advice leads to decreased reliance on algorithmic advice provided by the CA.* | Users conform more heavily with algorithmic advice in their final answers when being exposed to incorrect advice. | ** $p < .01$ |
| ***H2a:*** *Making the CA's stated accuracy levels transparent to the user will alleviate the stated effect of incorrect advice on trust in the information provided by the CA.* | While making accuracy levels transparent increases trust for users who receive incorrect algorithmic advice, the opposite holds for users who receive correct algorithmic advice. | ** $p < .01$ |
| ***H2b:*** *Making the CA's accuracy levels transparent to the user will induce users receiving incorrect advice conform more heavily to the advice.* | While users receiving incorrect algorithmic advice conform more heavily with this advice when the CA's accuracy levels are transparent to the user, the opposite holds for users who receive correct algorithmic advice. | * $p < .05$ |

**Table 2. Review of Key Hypotheses**

In a similar vein, our findings regarding the transparency manipulation provide an indication of the complexity trust-enhancing mechanisms can introduce. Indeed, our results show that users appreciate transparency, yet only when algorithmic advice has been erroneous. Accordingly, making transparent a system's performance level to the user elicits even higher levels of reliance on advice and trust in advice and can thus compensate for a lack of trust introduced through erroneous recommendations. At the same time, transparency decreases trust in and reliance on correct advice. While these findings support the idea of that demonstrating transparency around an AI's performance levels is a promising countermeasure specifically for systems that suffer from inaccuracy, transparency does not act as a default design mechanism to foster trust. Making transparent an AI-based system's performance level, thus, can create costly behavioral biases in decision making (i.e., trusting and conforming to erroneous AI) and prevent effective system use (i.e., conforming to correct AI). We suggest that while participants in the incorrect advice conditions felt reassured by a stated confidence level of 90%, participants in the correct advice conditions might have started to consider the possibility for the AI-based system to err, which they potentially would not have done without the statement. In addition, extant research on trust in AI-based systems has largely explored how to enhance trust, i.e., through a system's social cues (Feine et al. 2019).As previous examples illustrate, however, when viewing effective system use as a function of system performance, an increase in trust is not always desirable.

Last, comparing final correct answers across the four treatments and the four multiple choice questions (see Table 1), the significant differences between the correct versus incorrect advice groups strengthen our previous results. However, one can see for users who received incorrect algorithmic advice, the percentage of subjects which submits a correct final answer increases over the time of the four questions. The range of correct final answer submitted between the first and last multiple-choice question is particularly great for our treatment group 3, with the CA giving incorrect advice and providing transparency statements. This preliminary finding is in line with Berger et al.'s (2021) study which shows that the influence of incorrect algorithmic advice decreases with increased familiarity of the decision maker. As part of future research, the effect of algorithmic advice over time must be further explored. Little is known about the sequence and threshold of an erring AI-based system, i.e., the number of times a system can err before a user deviates from his or her reliance on the system.

## Implications

Our research holds important and novel implications for understanding decision makers' trust in and reliance on AI-based systems in the context of erroneous advice, thereby contributing to researchers' call for a consideration of the unintended consequences of AI-based IS (Rahwan et al. 2019). While model interpretability has been investigated to improve performance of human decision-making and perceived trustworthiness of the underlying ML model (Alufaisan et al. 2020; Doshi-Velez and Kim 2017; Shin 2021), researchers in the field of IS have been urging to further empirically study the disclosure of ML's inner workings to users and how such interpretations can meet ultimate users' demands (Bauer et al. 2021). The findings of the current study point towards a more nuanced picture on the paradox between user perception and related behavior, as well as on the contradicting effect transparency additionally has on perception and behavior. We contribute to existing research in the fields of IS, Human-Computer Interaction and psychology on the current understanding of how humans perceive the use of (erroneous) machines as decision aids. From a design science perspective, we empirically evaluate a specific design feature, namely an instantiation of transparency for trustworthy AI.

We explored the notions of trust and reliance in the specific context of a reading comprehension task. Our results suggest that a general trust in and reliance on algorithmic advice do not exist but can be viewed as a function of system performance. Additionally, we demonstrate that transparency on the algorithm's accuracy levels strengthens the effect that erroneous algorithmic advice has on trusting in and relying on that advice. More so, transparency influences the impact of algorithmic advice in different ways, very much depending on the correctness of the advice. We revisit established notions of trust (Söllner et al. 2012) and Lee and See's (2004) conceptual model of the dynamic process on trust and its effect on reliance specifically. While we cannot demonstrate a mediating effect of trust on subsequent behavior, i.e., reliance, our findings point towards the importance of conceptually differentiating between trust as a perceptual and reliance as a behavioral outcome. In addition, our results contribute to Lee and See's (2004) notion of information on system performance as a crucial factor influencing trust in IS.

Instantiations of AI-based IS as used as part of our study are increasingly deployed in learning and support tasks such as providing suggestions in e-mail applications (Seabrook 2019). From a practitioner's point of view, a key concern is the aversion towards algorithmic advice, especially if such advice outperforms human judgement (Yeomans et al. 2019). Beyond attempting to improving accuracy and performance rates of deployed systems, practitioners should consider how advice correctness affects actual behavior, and thus, decision making. While previous research suggests to increase transparency in the algorithmic forecasting process to foster adoption and use of algorithmic advice (Yeomans et al. 2019), our results indicate that transparency-enhancing design features should be implemented with caution. The current research shows that transparency can raise counterproductive actions on the decision-maker side. In that sense, users might rely more heavily and easily on erroneous advice, while (unnecessarily) questioning correct advice.

## Limitations and Future Research

The presented research and findings should be interpreted with caution and examined in the light of several limitations. First, our focus on a reading comprehension task embedded in a job application scenario implies limited external validity and thus impedes a generalization of results to scenarios, where, i.e., decision makers are domain experts or costs of system error are higher than in our study context. The question arises of whether our findings on trust and the asymmetrical moderating impact of transparency hold for, i.e., medical diagnoses. Future research should consider system users' initial disposition to trust in a specific AI-based system and perceived risk associated with a specific task. In a similar vein, the design of the experiment task ensured that participants could more easily identify whether an advice is correct and thus assess the CA's competences. As a result, the overall decision task was of comparatively low uncertainty.

Our study explores AI-based judgement in the context of a specific instantiation of an IS. While the CA was kept static and simple, the question arises of how trust is further influenced by an anthropomorphisation of the system. Existing research has shown that users are more likely to rely on algorithmic advice if the technological artefact exhibits similar personality cues to those of the user (Nass and Lee 2001). If humans believe they have more in common with human- than with algorithm-based advice (Prahl and Van Swol 2017), a personified CA is expected to increase reliance on recommendations. A natively built and trained CA was explicitly avoided to control for the algorithmic advice provided. However, the question arises to what extent our results hold for other instantiations of AI-based advice. Building on the findings by Yin et al. (2019), current manipulations could be extended by considering varying accuracy levels.

Last, the chosen experiment context is related to a homogenous participant pool. While this was deemed suitable regarding the chosen task at hand, one could assume that a younger student population exhibits a greater understanding and habit of interacting with AI-based systems. The chosen context and task allowed us to study the reliance on algorithmic advice in a realistic scenario, yet additional experiments should be conducted to verify our findings with different types of tasks and domains, as well as a more heterogenous participant pool. Overall, our work contributes to the fragmented discussion around reliance on algorithmic advice and towards clarifying the ambiguity around trust and related behavioral outcomes in existing research. Our findings point towards the importance of further empirical investigation on the circumstances under which humans place trust in AI-based systems and their advice.

## Acknowledgements

We thank the Swiss National Science Foundation for funding parts of this research (100013_192718) and acknowledge funding from the Basic Research Fund (GFF) of the University of St. Gallen.

## References

Agarwal, R., and Prasad, J. 1998. "A Conceptual and Operational Definition of Personal Innovativeness in the Domain of Information Technology," *Information Systems Research* (9:2), pp. 204–215. (https://doi.org/10.1287/isre.9.2.204).

Alufaisan, Y., Marusich, L. R., Bakdash, J. Z., Zhou, Y., and Kantarcioglu, M. 2020. "Does Explainable Artificial Intelligence Improve Human Decision-Making?," *ArXiv*. (https://doi.org/10.31234/osf.io/d4r9t).

Baird, A., and Maruping, L. M. 2021. "The next Generation of Research on Is Use: A Theoretical Framework of Delegation to and from Agentic Is Artifacts," *MIS Quarterly: Management Information Systems* (45:1), pp. 315–341. (https://doi.org/10.25300/MISQ/2021/15882).

Bauer, K., Hinz, O., van der Aalst, W., and Weinhardt, C. 2021. "Expl(AI)n It to Me – Explainable AI and Information Systems Research," *Business and Information Systems Engineering* (63:2), Springer Fachmedien Wiesbaden, pp. 79–82. (https://doi.org/10.1007/s12599-021-00683-2).

Berger, B., Adam, M., Rühr, A., and Benlian, A. 2021. "Watch Me Improve—Algorithm Aversion and Demonstrating the Ability to Learn," *Business and Information Systems Engineering* (63:1), pp. 55–68. (https://doi.org/10.1007/s12599-020-00678-5).

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

Cathy O'Neil. 2016. *Weapons of Math Destruction: How Big Data Increases Inequality and Threatens Democracy.*, Crown Publishing Group, USA.

Commission, E. 2019. "Ethics Guidelines for Trustworthy AI." (https://digital-strategy.ec.europa.eu/en/library/ethics-guidelines-trustworthy-ai).

D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., Chen, C., Deaton, J., Eisenstein, J., Hoffman, M. D., Hormozdiari, F., Houlsby, N., Hou, S., Jerfel, G., Karthikesalingam, A., Lucic, M., Ma, Y., McLean, C., Mincu, D., Mitani, A., Montanari, A., Nado, Z., Natarajan, V., Nielson, C., Osborne, T. F., Raman, R., Ramasamy, K., Sayres, R., Schrouff, J., Seneviratne, M., Sequeira, S., Suresh, H., Veitch, V., Vladymyrov, M., Wang, X., Webster, K., Yadlowsky, S., Yun, T., Zhai, X., and Sculley, D. 2020. "Underspecification Presents Challenges for Credibility in Modern Machine Learning," *ArXiv*.

Dawes, R. M. 1979. "The Robust Beauty of Improper Linear Models in Decision Making," *American Psychologist* (34:7), pp. 571–582. (https://doi.org/10.1037/0003-066X.34.7.571).

Dietvorst, B. J., Simmons, J., and Massey, C. 2015. "Understanding Algorithm Aversion: Forecasters Erroneously Avoid Algorithms After Seeing Them Err," *Academy of Management Proceedings* (2014:1), pp. 12227–12227. (https://doi.org/10.5465/ambpp.2014.12227abstract).

Dijkstra, J. J., Liebrand, W. B. G., and Timminga, E. 2017. *Persuasiveness of Expert Systems Persuasiveness of Expert Systems*, (3001:January), pp. 155–163.

Doshi-Velez, F., and Kim, B. 2017. *Towards A Rigorous Science of Interpretable Machine Learning*, (Ml), pp. 1–13. (http://arxiv.org/abs/1702.08608).

Dzindolet, M. T., Pierce, L. G., Beck, H. P., and Dawe, L. A. 2002. "The Perceived Utility of Human and Automated Aids in a Visual Detection Task," *Human Factors* (44:1), pp. 79–94. (https://doi.org/10.1518/0018720024494856).

Elkins, A. C., and Derrick, D. C. 2013. "The Sound of Trust: Voice as a Measurement of Trust During Interactions with Embodied Conversational Agents," *Group Decision and Negotiation* (22:5), pp. 897–913. (https://doi.org/10.1007/s10726-012-9339-x).

Endsley, M. R. 2017. "From Here to Autonomy: Lessons Learned from Human-Automation Research," *Human Factors* (59:1), pp. 5–27. (https://doi.org/10.1177/0018720816681350).

Faltings, B., Pu, P., Tran, B. D., and Jurca, R. 2014. "Incentives to Counter Bias in Human Computation," *Proceedings of the 2nd AAAI Conference on Human Computation and Crowdsourcing* (Hcomp), pp. 59–66. (https://doi.org/10.1145/2441776.2441828).

Feine, J., Gnewuch, U., Morana, S., and Maedche, A. 2019. "A Taxonomy of Social Cues for Conversational Agents," *International Journal of Human Computer Studies* (132:June), Elsevier Ltd, pp. 138–161. (https://doi.org/10.1016/j.ijhcs.2019.07.009).

Flender, J., Christmann, U., Groeben, N., and Mlynski, G. 1999. "Development and First Validation of a Scale of Passive Argumentational-Rhetorical Competence," *Zeitschrift Für Differentielle Und Diagnostische Psychologie* (20:4), pp. 309–325.

Fox, A. B., Rosen, J., and Crawford, M. 2009. "Distractions, Distractions: Does Instant Messaging Affect College Students' Performance on a Concurrent Reading Comprehension Task?," *Cyberpsychology and Behavior* (12:1), pp. 51–53. (https://doi.org/10.1089/cpb.2008.0107).

Gefen, D., and Straub, D. W. 2004. "Consumer Trust in B2C E-Commerce and the Importance of Social Presence: Experiments in e-Products and e-Services," *Omega* (32:6), pp. 407–424. (https://doi.org/10.1016/j.omega.2004.01.006).

Goddard, K., Roudsari, A., and Wyatt, J. C. 2012. "Automation Bias: A Systematic Review of Frequency, Effect Mediators, and Mitigators," *Journal of the American Medical Informatics Association* (19:1), pp. 121–127. (https://doi.org/10.1136/amiajnl-2011-000089).

Grove, W. M., and Meehl, P. E. 1996. "Comparative Efficiency of Informal (Subjective, Impressionistic) and

Formal (Mechanical, Algorithmic) Prediction Procedures: The Clinical-Statistical Controversy," *Psychology, Public Policy, and Law* (2:2), pp. 293–323. (https://doi.org/10.1037/1076-8971.2.2.293).

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

Haibe-Kains, B., Alexandru Adam, G., Hosny, A., Khodakarami, F., Society Board, M., Waldron, L., Wang, B., and Munk, P. (n.d.). "The Importance of Transparency and Reproducibility in Artificial Intelligence Research Affiliations," *John P.A. Ioannidis* (17), p. 20.

Highhouse, S. 2008. "Stubborn Reliance on Intuition and Subjectivity in Employee Selection," *Industrial and Organizational Psychology* (1:3), pp. 333–342. (https://doi.org/10.1111/j.1754-9434.2008.00058.x).

Hoff, K. A., and Bashir, M. 2015. "Trust in Automation: Integrating Empirical Evidence on Factors That Influence Trust," *Human Factors* (57:3), pp. 407–434. (https://doi.org/10.1177/0018720814547570).

Johnson, E. J., and Russo, J. E. 1984. "Product Familiarity and Learning New Information," *Journal of Consumer Research* (11:1), p. 542. (https://doi.org/10.1086/208990).

Jordan, M. I., and Mitchell, T. M. 2015. "Machine Learning: Trends, Perspectives, and Prospects," *Science* (349:6245), pp. 255–260. (https://doi.org/10.1126/science.aaa8415).

Jussupow, E., Spohrer, K., Heinzl, A., and Gawlitza, J. 2021. "Augmenting Medical Diagnosis Decisions? An Investigation into Physicians' Decision-Making Process with Artificial Intelligence," *Information Systems Research* (March). (https://doi.org/10.1287/isre.2020.0980).

Kehr, F., Kowatsch, T., Wentzel, D., and Fleisch, E. 2015. "Blissfully Ignorant: The Effects of General Privacy Concerns, General Institutional Trust, and Affect in the Privacy Calculus," *Information Systems Journal* (25:6), pp. 607–635. (https://doi.org/10.1111/isj.12062).

Knote, R., Janson, A., Söllner, M., and Leimeister, J. M. 2021. "Value Co-Creation in Smart Services: A Functional Affordances Perspective on Smart Personal Assistants," *Journal of the Association for Information Systems* (22:2), pp. 418–458. (https://doi.org/10.17705/1jais.00667).

Köchling, A., Riazy, S., Wehner, M. C., and Simbeck, K. 2020. "Highly Accurate, But Still Discriminatory: A Fairness Evaluation of Algorithmic Video Analysis in the Recruitment Context," *Business and Information Systems Engineering* (63:1), pp. 39–54. (https://doi.org/10.1007/s12599-020-00673-w).

Kuncel, N. R., Klieger, D. M., Connelly, B. S., and Ones, D. S. 2013. "Mechanical versus Clinical Data Combination in Selection and Admissions Decisions: A Meta-Analysis," *Journal of Applied Psychology* (98:6), pp. 1060–1072. (https://doi.org/10.1037/a0034156).

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

Lewicki, R. J., and Bunker, B. B. 2012. "Developing and Maintaining Trust in Work Relationships," *Trust in Organizations: Frontiers of Theory and Research* (October), pp. 114–139. (https://doi.org/10.4135/9781452243610.n7).

Li, H., Luo, X. (Robert), Zhang, J., and Xu, H. 2017. "Resolving the Privacy Paradox: Toward a Cognitive Appraisal and Emotion Approach to Online Privacy Behaviors," *Information and Management* (54:8), Elsevier B.V., pp. 1012–1022. (https://doi.org/10.1016/j.im.2017.02.005).

Liel, Y., and Zalmanson, L. 2020. *Association for Information Systems Association for Information Systems What If an AI Told You That 2 + 2 Is 5? Conformity to Algorithmic What If an AI Told You That 2 + 2 Is 5? Conformity to Algorithmic Recommendations Recommendations*, pp. 0–9. (https://aisel.aisnet.org/icis2020).

Logg, J. M., Minson, J. A., and Moore, D. A. 2019. "Algorithm Appreciation: People Prefer Algorithmic to Human Judgment," *Organizational Behavior and Human Decision Processes* (151:December 2018), Elsevier, pp. 90–103. (https://doi.org/10.1016/j.obhdp.2018.12.005).

Luhmann, N. 1979. *Trust and Power. Two Works by Niklas Luhmann*, Wiley, Chichester.

Mayer, R. C., Davis, J. H., and Schoorman, D. F. 1995. "An Integrative Model of Organizational Trust," *Academy of Management Review* (20:3), pp. 709–734.

McKnight, D. H., Liu, P., and Pentland, B. T. 2020. "Trust Change in Information Technology Products," *Journal of Management Information Systems* (37:4), Routledge, pp. 1015–1046. (https://doi.org/10.1080/07421222.2020.1831772).

Nass, C., and Lee, K. M. 2001. "Does Computer-Synthesized Speech Manifest Personality? Experimental

Tests of Recognition, Similarity-Attraction, and Consistency-Attraction," *Journal of Experimental Psychology: Applied* (7:3), pp. 171–181. (https://doi.org/10.1037/1076-898X.7.3.171).

Parasuraman, R., and Manzey, D. H. 2010. "Complacency and Bias in Human Use of Automation: An Attentional Integration," *Human Factors* (52:3), pp. 381–410. (https://doi.org/10.1177/0018720810376055).

Parasuraman, R., and Riley, V. 1997. "Humans and Automation: Use, Misuse, Disuse, Abuse," *Human Factors* (39:2), pp. 230–253.

Prahl, A., and Van Swol, L. 2017. "Understanding Algorithm Aversion: When Is Advice from Automation Discounted?," *Journal of Forecasting* (36:6), pp. 691–702. (https://doi.org/10.1002/for.2464).

Rahwan, I., Cebrian, M., Obradovich, N., Bongard, J., Bonnefon, J. F., Breazeal, C., Crandall, J. W., Christakis, N. A., Couzin, I. D., Jackson, M. O., Jennings, N. R., Kamar, E., Kloumann, I. M., Larochelle, H., Lazer, D., McElreath, R., Mislove, A., Parkes, D. C., Pentland, A. 'Sandy,' Roberts, M. E., Shariff, A., Tenenbaum, J. B., and Wellman, M. 2019. "Machine Behaviour," *Nature* (568:7753), Springer US, pp. 477–486. (https://doi.org/10.1038/s41586-019-1138-y).

Rammstedt, B., and John, O. P. 2007. "Measuring Personality in One Minute or Less: A 10-Item Short Version of the Big Five Inventory in English and German," *Journal of Research in Personality* (41:1), pp. 203–212. (https://doi.org/10.1016/j.jrp.2006.02.001).

Robinette, P., Howard, A. M., and Wagner, A. R. 2017. "Effect of Robot Performance on Human-Robot Trust in Time-Critical Situations," *IEEE Transactions on Human-Machine Systems* (47:4), IEEE, pp. 425–436. (https://doi.org/10.1109/THMS.2017.2648849).

Seabrook, J. 2019. "The Next World: Where Will Predictive Text Take Us?," *The New Yorker*.

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:April 2020), Elsevier Ltd, p. 102551. (https://doi.org/10.1016/j.ijhcs.2020.102551).

Söllner, M., Hoffmann, A., Hoffmann, H., and Wacker, A. 2012. "Understanding the Formation of Trust in IT Artifacts," in *Proceedings of the International Conference on Information Systems (ICIS)*, Orlando, Florida, USA.

Söllner, M., Hoffmann, A., and Leimeister, J. M. 2016. "Why Different Trust Relationships Matter for Information Systems Users," *European Journal of Information Systems* (25:3), pp. 274–287. (https://doi.org/10.1057/ejis.2015.17).

Waardenburg, L., Sergeeva, A., and Huysman, M. 2018. "Predictive Policing: How Algorithms Inscribe the Understanding of Crime in Police Work," *Academy of Management Proceedings* (April 2018), p. 132.

Wambsganss, T., Kung, T., Sollner, M., and Leimeister, J. M. 2021. "Arguetutor: An Adaptive Dialog-Based Learning System for Argumentation Skills," *Conference on Human Factors in Computing Systems - Proceedings*. (https://doi.org/10.1145/3411764.3445781).

Wambsganss, T., Niklaus, C., Cetto, M., Söllner, M., Leimeister, J. M., and Handschuh, S. 2020. "AL : An Adaptive Learning Support System for Argumentation Skills," in *ACM CHI Conference on Human Factors in Computing Systems*, pp. 1–14.

Wambsganss, T., Niklaus, C., Söllner, M., Handschuh, S., and Leimeister, J. M. 2021. *Supporting Cognitive and Emotional Empathic Writing of Students*, pp. 4063–4077. (https://doi.org/10.18653/v1/2021.acl-long.314).

Whitecotton, S. M. 1996. "The Effects of Experience and a Decision Aid on the Slope, Scatter, and Bias of Earnings Forecasts," *Organizational Behavior and Human Decision Processes* (66:1), pp. 111–121. (https://doi.org/10.1006/obhd.1996.0042).

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

Yin, M., Vaughan, J. W., and Wallach, H. 2019. "Understanding the Effect of Accuracy on Trust in Machine Learning Models," *Conference on Human Factors in Computing Systems - Proceedings*, pp. 1–12. (https://doi.org/10.1145/3290605.3300509).

Zhang, Y., Vera Liao, Q., and Bellamy, R. K. E. 2020. "Efect of Confidence and Explanation on Accuracy and Trust Calibration in AI-Assisted Decision Making," *FAT* 2020 - Proceedings of the 2020 Conference on Fairness, Accountability, and Transparency*, pp. 295–305. (https://doi.org/10.1145/3351095.3372852).

Zierau, N., Wambsganss, T., Janson, A., Schöbel, S., and Leimeister, J. M. 2020. "The Anatomy of User Experience with Conversational Agents: A Taxonomy and Propositions of Service Clues," in *International Conference on Information Systems (ICIS)*.