6-18-2022

Designing for Conversational System Trustworthiness: The Impact of Model Transparency on Trust and Task Performance

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

Thiemo Wambsganss
University of St. Gallen, thiemo.wambsganss@epfl.ch

Andreas Janson
Institute of Information Management, andreas.janson@unisg.ch

DESIGNING FOR CONVERSATIONAL SYSTEM TRUSTWORTHINESS: THE IMPACT OF MODEL TRANSPARENCY ON TRUST AND TASK PERFORMANCE

Research Paper

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

Thiemo Wambsganss, École Polytechnique Fédérale de Lausanne, Lausanne, Switzerland, thiemo.wambsganss@epfl.ch

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

Abstract

Designing for system trustworthiness promises to address challenges of opaqueness and uncertainty introduced through Machine Learning (ML)-based systems by allowing users to understand and interpret systems' underlying working mechanisms. However, empirical exploration of trustworthiness measures and their effectiveness is scarce and inconclusive. We investigated how varying model confidence (70% versus 90%) and making confidence levels transparent to the user (explanatory statement versus no explanatory statement) may influence perceptions of trust and performance in an information retrieval task assisted by a conversational system. In a field experiment with 104 users, our findings indicate that neither model confidence nor transparency seem to impact trust in the conversational system. However, users' task performance is positively influenced by both transparency and trust in the system. While this study considers the complex interplay of system trustworthiness, trust, and subsequent behavioral outcomes, our results call into question the relation between system trustworthiness and user trust.

Keywords: Trust, Trustworthiness, Transparency, Machine Learning, Information Retrieval, Pedagogical Conversational Agents.

1 Introduction

Conversational systems have become a ubiquitous part of everyday life and are increasingly deployed for use in large-scale personal and social settings, e.g., to assist with daily tasks or to inform decision-making (Meshram et al., 2021; Stieglitz et al., 2021; Zhou et al., 2015). Despite their potential, challenges commonly arising in the context of such systems include user acceptance, as well as system mis- and disuse (Saddarizadeh et al., 2017; Zhou and Chen, 2018). An explanation for the amplification of such challenges can be traced back to the very nature of Machine Learning (ML)-based systems which are marked by opaqueness, complexity, and uncertainty introduced by, i.e., input data and the nature of statistical prediction models (Hamon et al., 2022; Miller, 2019; Zhou and Chen, 2018).

This is where the importance of trust comes into play since "[t]rust is an important mechanism for coping with the cognitive complexity that accompanies increasingly sophisticated technology" (Lee and See, 2004). Trust not only helps explain how humans deal with system-accompanying uncertainty, yet is fundamental to the design of such systems as it is tightly linked to the acceptance and enjoyment of interacting with the system (Gefen and Straub, 2004). In that sense, trust can directly affect people's willingness to interact, their involvement with the interaction, and their willingness to rely on information provided by a conversational system (Conrad et al., 2015; Pickard et al., 2016; Schuetzler

et al., 2018). For systems' most effective use, however, users must trust respective systems appropriately, often also referred to as trust calibration (Lee and See, 2004). For instance, students might blindly follow the recommendations of an erroneous learning system, ultimately impeding the learning process and outcomes. Next to ethical ramifications, trust misalignment is detrimental to the user and can be costly and harmful on organizational scale (Hidalgo et al., 2021). As the previous example illustrates, simply enhancing user trust is not desirable as users might be persuaded to over-trust the system (Zhang et al., 2020). More so, in educational and information retrieval settings, the user usually has a stake in receiving accurate, if not, any information provided by the system (Kratzwald and Feuerriegel, 2019).

Research in the field of information systems (IS) and Human Computer Interaction (HCI) has extensively explored trust in conversational systems in light of social cues, i.e., anthropomorphization of the appearance or conversational style, contributing to understand which noticeable cues of conversational systems ultimately influence users' perception and behavior (Araujo, 2018; Knote et al., 2021). However, conversational systems mimicking human features can cloud users' understanding of these systems being non-human and obscure privacy-related risk, thus potentially leading to overtrust (Aroyo et al., 2021; Puranam & Vanneste, 2021). In addition, taking into consideration the increasingly complex, opaque, and unpredictable nature of such systems, researchers have urged to explore trust more extensively as a function of system performance. In fact, "[t]rust […] has been shown to be a key mitigating factor in system use/disuse […] and importantly, […] subject to the user's perception of system performance." (Yu et al., 2016).

A promising way to address the beforementioned challenges and enhance system trustworthiness is to communicate model confidence or reliability rates (Bansal et al., 2019). By making the system's functioning and underlying model transparent to the user, uncertainty around the system design and users' sensemaking process could be facilitated. Researchers have started to investigate transparency as a means of trust calibration in crowdwork (Logg et al., 2019) or healthcare (Jussupow et al., 2021), yet oftentimes only regard perceptual or hypothetical behavioral outcomes. More so, mixed results have been found regarding the effectiveness of transparency statements in fostering user trust (Kästner et al., 2021). Little is known about how ML-based IS perform in learning and decision-making contexts. Current trust literature falls short on exploring how system trustworthiness not only affects user trust yet also behavioral interaction outcomes, i.e., task performance, and how to differentiate these two. The relation between the implementation of design features contributing to system trustworthiness and the desired implications of such features is not as straightforward (Jacovi et al., 2021). This study aims to provide empirical evidence on how transparency statements on a conversational system's confidence influence trust and subsequent task performance in the context of an information retrieval task. We seek to address the outlined objectives by answering the following research questions (RQs):

RQ1: *What is the effect of transparency statements on user trust and task performance?*

RQ2: *How do varied confidence rates of a trained conversational system alter the effects of such transparency statements?*

To answer our research questions, we conducted a 2x2 between-subject experiment to test whether transparency statements on system confidence (explanatory statement versus no explanatory statement) and system confidence (70% versus 90% intent modelling confidence) result in higher levels of trust and subsequent performance. We deployed the manipulations in four instantiations of a trained, pre-tested conversational system. In the context of a graduate university course, participants interacted with the system to retrieve course-relevant information and to answer course-related questions. We found that performance in the information retrieval task was positively influenced by participants being exposed to the transparency statements, as well as by higher levels of trust. However, we did not find a significant effect of transparency on trust nor a moderating effect of varying system confidence. Our results do not find support for the notion that measures of system trustworthiness, i.e., transparency or improving system confidence, necessarily lead to increased user trust. Nevertheless, both transparency and users' trust in the system seem to have an impact on

behavioral outcomes of the interaction, namely task performance. Our research contributes to the understanding of the relation between system trustworthiness and trust, as well as the relation between trust and subsequent behavioral outcomes.

2 Conceptual Background and Hypotheses Development

In the following, we review a task-specific class of conversational systems and lay out relevant work related to the importance of trust in conversational systems, the notion of system trustworthiness, as well as transparency statements and reliability rates as means of trustworthiness.

2.1 Conversational Systems for Information Retrieval Tasks

Conversational systems present a particular type of ML-based IS, distinguishing themselves through a dialogue-based interaction via text or speech (Pfeuffer et al., 2019; Rubin et al., 2010). Through Natural Language Processing (NLP), conversational systems can identify and respond to user intents (Shawar and Atwell, 2005). While commercially available conversational systems such as Amazon's Alexa assist with daily tasks and general information requests, such systems and their capabilities can also be implemented in domain-specific contexts (Knote et al., 2021). In fact, conversational systems are to be found in a variety of domains, finding application in frontline service applications such as customer retention management (Mozafari et al., 2021), healthcare (Wienrich et al., 2021), assistance in reading comprehension tasks (Schmitt et al., 2021) and problem-solving support in educational scenarios (Winkler et al., 2021).

When turning towards traditional question-answering systems deployed in educational and learning settings, the sophisticated interaction quality of conversational systems promises to provide more precise and personalized feedback to learners' requests (Wambsganss et al., 2021b). In fact, conversational systems allow users to retrieve relevant information in a simple, effective and adaptive manner (Kratzwald and Feuerriegel, 2019). While the impact of pedagogical conversational systems on certain perceptual and behavioral outcomes, including learning-related measures and performance, have been studied (Weber et al., 2021), trust measures have been only scarcely investigated in the educational domain and for information retrieval task despite learning scenarios presenting trust-relevant contexts where costs of system error are high, and decision outcomes are important to the individual user (Wollny et al., 2021).

2.2 Trust in Conversational Systems and Trust-Related Behavior

According to established trust theories and conceptualizations, we speak of a trust-relevant context in light of the possibility of a disadvantageous or undesirable event to the trustor (Gambetta, 1988; Rousseau et al., 1998). Trust itself is a cognitive reaction to reduce complexity, although undesirable outcomes are possible (Mayer et al., 1995). Per se, this definition implies that trust lies with the user and is a perceptual attitude the user holds towards a system (Kästner et al., 2021). More so, a context of trust requires consideration of alternatives and a choice by the trustor, which ultimately results in him or her choosing one action over the other (Luhmann, 2000). In the context of conversational systems, users presume a favorable behavior of the system despite the uncertainty of the system providing erroneous or unfavorable output (i.e., providing incorrect information, giving no answer at all).

Extant literature argues trust to be viewed as a second-order construct, and thus as an antecedent or even prerequisite for effective and sustainable system adoption and use (McKnight et al., 2011; Turel and Gefen, 2013). As a result, trust-related behavioral outcomes are important to understand how attitudes of trust translate into subsequent behavior. Söllner (2020), for instance, find that higher levels of trust in a decision support system lead to increased system usage and reliance on such systems. Ou et al. (2014) demonstrate that in the context of an online marketplace buyers' trust positively influences subsequent repurchases from sellers. In a similar vein, multiple studies have found a positive influence of trust on intentions to use and accept AI-based systems (Glikson and Woolley,

2020). For conversational systems and the educational domain specifically, however, little is known if and how trust in the system matters for learning-related outcomes.

Beyond the conceptual relation between trust and trust-related behavior, behavioral outcome variables offer the opportunity to contribute more profoundly to an empirical understanding and theory finding as such behavior can be objectively quantified and measured, i.e., through log data or measurable task outcomes (Hulland and Houston, 2021). Lee and See's (2002) conceptual framework for trust calibration provides a theoretical distinction between trust and subsequent, distinct behavioral outcomes. Most literature focuses on self-reported measures of trust (Kohn et al., 2021). In the best cases, these measures explore the underlying cognitive processes of user trust, in the worst cases, these measures do not provide a clear delineation of what specifically the term "trust" means in the context of their research. In both cases, however, there is a lacking differentiation between trust as a cognitive, self-reported construct and subsequent, potentially trust-related behavior. Studies that consider subsequent behavioral outcomes mostly focus on intentional variables such as intention to use or acceptance and reliance on such systems (Lane et al., 2016; Wang and Benbasat, 2005). So far, little research has investigated task outcome- and performance-related behavioral outcomes. In the context of our study context, performing well in the information retrieval task represents an important goal of the interaction. Based on findings from previous studies exploring the effect of trust on subsequent behavioral outcomes (Gefen and Straub, 2004; Pavlou and Gefen, 2004; Ou et al., 2014; Söllner, 2020), we pose the following hypothesis:

H1: *Greater levels of user trust in the conversational system have a positive effect on task performance.*

2.3 Transparency Statements and System Reliability as Means of System Trustworthiness

A crucial third dimension around trust next to user trust and trust-related behavior is system trustworthiness (Jacovi et al., 2021; Kästner et al., 2021; Lee and See, 2004). Trustworthiness encompasses external dimensions that lie with the trustee, in our case the conversational system. Causes or attributes of the trustee can influence previously mentioned cognitive processes of user trust. Measures for designing for system trustworthiness have been extensively studied in the HCI literature (Karsenty and Botherel, 2005; Yin et al., 2019; Zhang et al., 2020b). In light of ongoing advances and the unintended implications of ML-based systems, various institutions have proposed a set of guidelines on to increase system trustworthiness (Independent High-Level Expert Group on Artificial Intelligence, 2019). These include but are not limited to technical design mechanisms such as technical robustness and safety, as well as conversational design mechanisms such as transparency around system capabilities, limitations, and levels of accuracy.

2.3.1 Conversational Systems and Transparency

In the discussion around trust calibration, various researchers such as Hoffman et al. (2018) have pointed towards "[…] a need to explain […] so that users and decision makers can develop appropriate trust […]." In that sense, making elements such as overall system confidence or insights into underlying ML models transparent to the user can act as a means to help users distinguish cases they can trust from those they should not. Transparency measures, i.e., providing information about the accuracy and reliability rates of a system, are commonly brought up (Zhang et al., 2020). Different types of information can be provided to make transparent the inner workings of ML-based systems.

Various studies have investigated the effect of providing cues about the kind of information analyzed by the algorithm yet showing mixed results regarding their effectiveness (Langer et al., 2018; Langer and Landers, 2021; Newman et al., 2020). In an online learning context, a study by Kizilcec (2016) even shows that providing too much information on an algorithmic interface can eliminate trust. The effect of transparency statements on user perceptions is not as straightforward as some studies suggest a negative effect of making a system's performance transparent to the user (Castelo et al., 2019) while other studies suggest the opposite (Nagulendra and Vassileva, 2016; Yeomans et al., 2019). Early

research on recommender systems illustrates that transparency measures positively affect user trust, acceptance and satisfaction (Herlocker et al., 2000; Kulesza et al., 2013; Sinha and Swearingen, 2002). A more recent study illustrates that making transparent an algorithm's self-improving nature led users to rely more heavily on the algorithm's provided information as part of a judgment task (Berger et al., 2021). Despite a lack of established core findings on transparency, transparency measures directed at the user are claimed to "build a sense of trust in the technology" (Felzmann et al., 2021, p. 5). We hence expect transparency statements on our systems' confidence to reduce opaqueness around the functioning of our conversational system. We, therefore, hypothesize that:

H2: *Transparency statements on a conversational system's intent modelling confidence lead to enhanced user trust in the conversational system.*

2.3.2 Conversational Systems and Their Technical Reliability

Most current studies have explored systems' technical accuracy or reliability as a function of correct versus incorrect advice. In the realm of algorithmic aversion, erroneous advice has been named as a key factor for decreased trust in AI-based systems (Yin et al., 2019; Yu et al., 2016; Zhang et al., 2020). Based on the assumption that humans desire perfect predictions, coined as the perfection schema, errors are perceived as particularly negative (Dawes, 1979). Extant research has found that users overestimate perceived error rates of systems (Dzindolet et al., 2002; Hoff and Bashir, 2015) and small mistakes made by AI-based systems already lead to a significant decrease in trust (Dietvorst et al., 2015). Interestingly, individual studies demonstrate that erroneous recommendations can also positively affect trust-related behavioral outcomes (Liel and Zalmanson, 2020).

While research around erroneous algorithmic advice and algorithmic aversion provides insights into how suboptimal system performance affects user trust and subsequent behavior, performance of algorithms is usually compared to humans (Bigman and Gray, 2018). In the context of visual detection trials, Madhavan and Wiegmann (2004), for instance, deploy an algorithm of 70% actual reliability, yet explore the effect of framing this algorithm as a novice or as a human expert. They find that decision makers do not know about actual algorithm performance and thus assess the performance subjectively. Extant reviews find inconsistent effects of algorithm performance and call for further testing, for instance, by making users aware of actual reliability rates (Jussupow et al., 2020).

More so, the nature of contemporary IS trained on predictive ML models require a more nuanced understanding of system performance. While accuracy provides a measurement describing the systematic error of a classification model over a certain distribution, the confidence level for a single case prediction might be interesting to consider when turning towards conversational modelling and its effect on user perception. Yu et al. (2019) considered different levels of system accuracy, exploring user interaction with an automation system at 10%, 20%, 30%,… up to 100% system accuracy. They found that the threshold for a certain level of user trust is at 70% system accuracy. Beyond their study, however, extant literature has rarely explored the effect of varying the threshold of confidence levels for single predictions of AI-based systems and their effects on user perceptions of trust and subsequent related behavior.

We view the variation of the intent modelling confidence as a technical design feature that allows us to influence the interaction of the user with the conversational system. More specifically, the system can be trained on various intent modelling confidences which represent a barrier to providing an answer to user intents. Lower confidence leads to more wrongly classified answers, however, there are fewer errors in the intent recognition overall since the system is providing answers for user requests where model confidence is low. On the other hand, higher confidences lead to more correctly classified answers, yet also a potentially more conservative provision of answers (Bird et al., 2009). According to Yu et al. (2019), 70% system accuracy of an AI-based system represents the threshold of user trust. Beyond their paper exploring system accuracy in the context of a decision-making task in a factory setting, little have researchers explored system performance as a function of accuracy or confidence levels. Jussupow et al. (2021) refer to 90% algorithmic accuracy as a usually accepted accuracy in the context of medical diagnosis decision making. In the context of conversational systems such as

chatbots, confidence levels depend on the size and quality of the training data for each intent. Gapanyuk et al.'s (2018) question-answering chatbot relies on a 85% confidence level threshold. In a comparable educational context where a chatbot is deployed to respond to college students' enquiries, Meshram et al. (2021) arrive at average confidence scores between 0.98 and 0.99. In the development process of a chatbot for helping users learn to code, Ilić et al. (2020) present three analyzed frameworks which exhibit confidence levels ranging from 81% to 100% on the training set. We thus assume that 90% is an acceptable and reliable confidence level in the context of an educational information retrieval task. While domain- and context-specific studies of acceptable accuracy levels give us an indication of reliable confidence levels, there is no straightforward understanding of how different intent modelling confidences will exactly perform and subsequently affect user perceptions and behavior in our study context. Yet we expect that system statements making transparent the overall intent modelling confidence of the agent will decrease users' trust in the agent if this system confidence is lower (70%) as opposed to higher (90%). Hence, we hypothesize the following:

H3: *Positive effects of transparency statements are decreased under conditions of lower overall system confidence.*

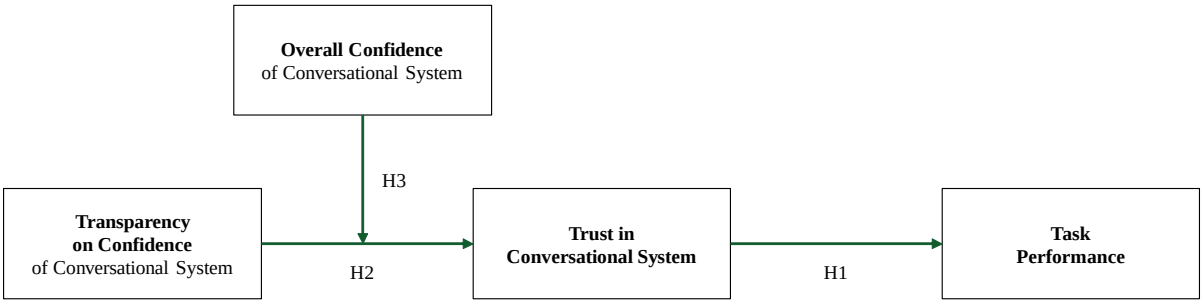

Figure 1. Conceptual research model.

3 Research Methodology

To explore the effects of 1) actual intent modelling system confidence and 2) transparency on a conversational system's actual confidence on user trust and task performance, we conducted a field experiment in the realm of an information retrieval task. A 2 (70% versus 90% system confidence) x 2 (explanatory statement versus no explanatory statement) between-subject design resulted in four treatments participants were randomly assigned to. For the conversational system, we developed, manipulated, and implemented a probabilistic, intent-based conversation structure.

3.1 Experimental Design and Procedure

We tested our hypotheses in the context of a university course. Before the start of the course and lectures, we provided students access to the conversational system to allow them to familiarize themselves with the course content and structure. The use of the conversational system was advertised as an alternative to traditional FAQ documents students usually receive to inform themselves about course-relevant information such as deadlines, deliverables, and used tools. Our experimental setup thus provided a field setting as part of which participation was voluntary. Setup and execution of the experiment were closely aligned and communicated with the lecturer of the course. In addition, we ensured that participation, as well as data shared and stored for the experiment were in line with the ethical standards and privacy guidelines of the university. As part of the interaction, students were provided six multiple choice (i.e., "What happens if I miss the first deadline for the feedback assignment?") and two open-ended questions (i.e., "What course deliverables are graded? Please provide an overview of all deliverables and what they contain.") on course-related matters. The experiment followed the sequence of 1) a pre-test phase, 2) an experiment phase, and 3) a post-test phase. As part of the overall study and experiment survey, we referred to the term chatbot instead of

conversational system. A chatbot is a specific, text-based instantiation of a conversational system and we believed students to be more familiar with this term (Shawar and Atwell, 2005). While the pre- and post-test phases remained stable across treatment groups, the experiment phase, including the interaction with the conversational system, varied across conditions according to our confidence and transparency manipulations.

Pre-Test: Participants were first informed about the aim of the overall study and the deployment of the system in the realm of the university course. They were incentivized to partake in the study and to perform well in the information retrieval task by raffling consumption vouchers to the university shop among the participants performing best in the information retrieval task. As part of 14 pre-survey questions, we collected control items and conducted an attention check.

Experiment: As part of the experiment phase, we presented students a number of simple questions on the course content and structure, as well as a link to the conversational system. We asked students to interact with the system in order to retrieve necessary information required to answer the presented course questions. The control group (CG) interacted with a conversational system of 90% system confidence, not disclosing any transparency on the confidence levels. Treatment group (TG) 1 used a system with 70% system confidence and no transparency statements, whereas TG2 and TG3 both exhibited transparency statements on the respective 70% or 90% overall system confidence.

Post-Test: The study concluded with a post-experiment questionnaire of 16 items as part of which we collected self-reported, perceptual outcome variables, open-ended questions on the interaction, additional control variables, and demographics.

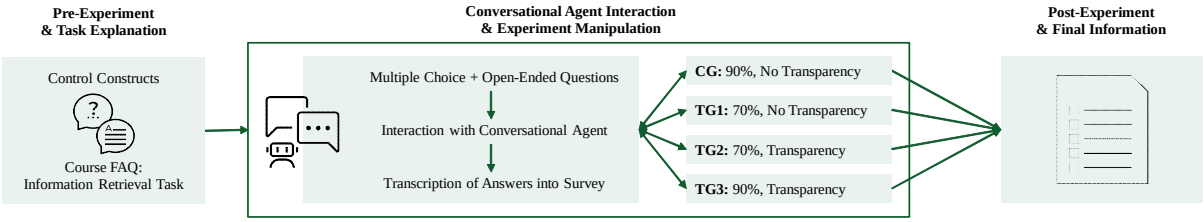

Figure 2. Experimental procedure.

3.2 Design and Manipulation of a Conversational System for Information Retrieval

As part of our experiment, participants interacted with a text-based system and were randomly assigned to one of four instantiations thereof. More specifically, the conversational system provided answers of either 70% or 90% overall confidence and statements making transparent its confidence levels. The four instantiations of the system were built based on the same backend and ML-model trained on over 70 intents on course content and structure. As a result, all four conversational systems were able to provide adaptive and personalized answers to the user.

In order to implement our system confidence manipulation, the conversational system was trained on either 70% or 90% system confidence. As Bird et al. (2009) mention, 80% accuracy in the prediction of text-based labels is often a good rule-of-thumb for a threshold for an embedding in real-life scenarios. With our study we aimed to explore the impact of different confidence levels (relatively low versus relatively high) on user trust and behavior. Working with a trained ML-based system, we cannot predict actual performance in the field. However, by setting a confidence level difference of 20%, we are convinced to ensure a perceivable difference between a “low” and “high” performance of the conversational system.

Regarding the transparency manipulation, we integrated transparency statements into the answers of the respective system. We placed transparency statements 1) at the beginning of the interaction where the conversational system introduced itself, 2) at the end of every extensive answer on course-related questions, as well as 3) in recovery statements when the system did not know what to reply.

Transparency statements varied in length, and as the conversational system initiated the interaction, participants were exposed to at very least one transparency statement. Doing so, we ensured that a user would encounter transparency statements multiple times throughout the interaction.

Interaction flow, communication style, and appearance of the system were developed in a precedent step based on a literature review on human-computer interaction and educational technology, as well as thirteen user interviews. Based on our findings from literature and the user interviews, we designed a novel conversational system for educational settings (Wambsganss et al., 2021a). For the adaptive back-end functionality of our conversational system, we utilized a combination of different NLP- and ML-based techniques. In general, the system is built as a web app in HTML5 with CSS and JavaScript. The front end is connected to a python script that a) processes incoming user intents and b) provides predefined answers based on the incoming classifications. For the conversational interaction, we modelled 70 intents, including an introduction, frequently asked questions, and casual dialogue. The intents were then trained based on a "Naive Bayes classifier" in combination with semantic similarity matching. The conversational back end was implemented utilizing the frameworks chatterbot and spacy. The test-based conversational system Hermine was tested and evaluated in a pre-study with 45 students (Wambsganss et al., 2021a). By keeping appearance (i.e., agent avatar, corporate identity), navigation (i.e., help button), language (i.e., colloquial and personalized), and layout (i.e., device-agnostic) constant across conditions, we ensured that the effects could be explained due to conducted manipulations.

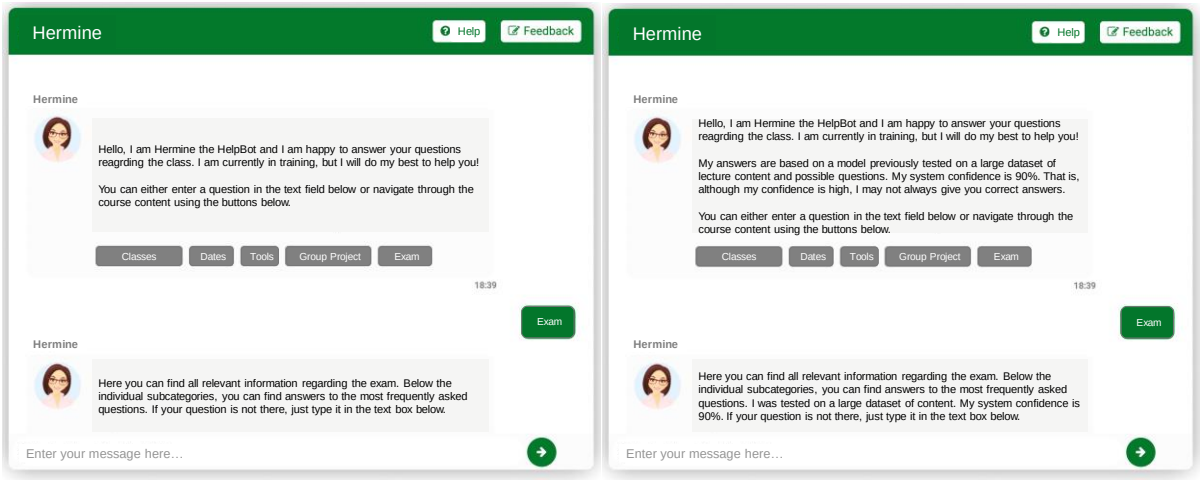

Figure 3. Conversational system without versus with transparency statements.

3.3 Measures

Our key measures include both perceptual and behavioral outcome variables on individual-user basis. Regarding self-reported variables. We measured trust in the system with three items (scale adapted from (McKnight et al., 2020), sample item: "For information on the course content and structure, I feel I can depend on the CA"; 7-point scale, from 1: "Strongly Disagree" to 7: "Strongly Agree", α_{Trust}= .87). We measured multiple control variables in both the pre- and post-experimental questionnaire, including participants' trusting disposition (Gefen and Straub, 2004), Big Five personality constructs (Rammstedt and John, 2007), and algorithmic familiarity (scale adapted from Johnson and Russo, 1984).

The measurement of our behavioral dependent variables was part of the information retrieval task in our experiment. The MC questions offered four to five pre-defined answer options, with one of them being correct. As a result, individual answer correctness for the six MC questions was measured as either correct (1) or incorrect (0). This grading was conducted for each MC question. In a subsequent

step, an overall performance score for the MC questions was calculated for each datapoint, summing the results for the six questions (range: 0-6, 6 = highest). Regarding the O-E course questions, we applied the following grading scheme in correspondence with the lecture and course representative to assess the performance of the students: Our grading scheme exhibited four gradation levels (1) Completely correct, 2) Correct, yet missing information, 3) Partially correct, partially incorrect, 4) Completely incorrect) according to which we assessed the accuracy and completeness of each answer provided to the O-E questions. Participants were not limited in the content and the amount of their answers. Necessary keywords as well as a first classification of exemplary answers were developed by a first annotator. In a second step, a second annotator classified the same selected answers and provided a revised selection of keywords and a grading scheme. Upon agreement, the second annotator proceeded to grade all open-ended questions independently. Answers only received a maximum of four points when all keywords and a sufficient explanation for each were given. One point was deducted when no full explanation was given, yet all keywords were mentioned. If keywords were missing or explanations were substantially lacking, the answer was graded with two points. Any answer less than that received one point. Similar to the procedure for the MC questions, we calculated an overall performance score for the two O-E questions for each datapoint (range: 1 – 8, 8 = highest). The overall task performance score represented the sum of the grade of both the MC and the O-E score.

3.4 Data Collection and Cleaning

We collected data as part of a pool of graduate students of a particular university course. As over 160 students were enrolled in the course, an important consideration before the data collection was whether a sufficient number of datapoints per cell to observe relatively stable effect sizes could be ensured. An a priori power analysis based on simulations in R for the 2x2 between-subject ANOVA design (thus, $u = 4$), given a large effect size ($f = 0.4$), common significance level (0.05), and power of test (0.80) suggests at least 18 (n = 18.043) datapoints in each treatment group.

We distributed our survey before the start of the course, advertising the conversational agent to be deployed as an informative tool and replacing a simple FAQ document. A total of 121 participants fully completed the study. To ensure attentive participation, students were incentivized by raffling gift vouchers to the university shop among the participants with the most correct answers to the questions of the information retrieval task. The voluntary participation represents a potential boundary of our study, as we expected only students who are generally motivated to better understand the course and who are curious about novel technology to partake in our study. We attempted to alleviate this limitation by making no other course information (i.e., FAQ document) available to the students at that point in time. Subjects who failed the attention check or who remarked having had technical difficulties with the conversational system were removed from our dataset. We further removed participants who exhibited abnormal completion time or completion patterns, leaving us with a final sample set of 104 subjects. A potential boundary represents the survey distribution within the course of students who most probably exhibit greater familiarity with ML-based systems as compared to the general public. Additional analyses on the control and demographic variables confirm participants' random assignment to the different experimental conditions. Specifically, there are no significant differences in trusting disposition, algorithmic familiarity, or personality traits among the four treatments (all $p > .1$). In addition, no differences were found regarding the demographic variables age and gender (all $p > .1$).

4 Results

To explore the effects of 1) system confidence and 2) transparency around a conversational system's confidence levels on user trust and task performance, we conducted a field experiment in the realm of an information retrieval task. We first conducted a manipulation check for the transparency manipulation, asking participants to what extent they agree with the following two statements: 1) "As part of the tasks, the chatbot revealed information about itself, namely a statement on the confidence

of the answers it provided and 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 validated the effectiveness of the transparency manipulation: An ANOVA on the first statement revealed a significant manipulation effect ($F = 50.36$, $p < .001$) with participants from the two treatment groups receiving transparency statements (70% system confidence, transparency; 90% system confidence, transparency) exhibiting a significantly higher confirmation of the first statement ($M_{Transparency} = 5.61$) than participants who were not exposed to the transparency-enhancing statements ($M_{NoTransparency} = 3.36$). Another ANOVA on the second statement strengthened this finding as no significant effect between the transparency present ($M_{Transparency} = 2.12$) versus transparency absent ($M_{NoTransparency} = 2.38$) treatments could be found.

Group	**N**	**Trust** (1 – 7, 7 = highest)	**Performance Overall** (1 – 14, 14 = highest)	**Performance (MC)** (0 – 6, 6 = highest)	**Performance (O-E)** (1 – 8, 8 = highest)
CG: 90%, No Transparency	26	3.97	10.1	5.23	4.85
TG1: 70%, No Transparency	27	4.31	10.7	5.44	5.26
TG2: 70%, Transparency	29	4.20	11.6	5.76	5.79
TG3: 90%, Transparency	22	4.36	11.4	5.95	5.45
Transparency Manipulation		ns $p > .1$	** $p < .01$	** $p < .01$	. $p < .1$
Interaction Effect (Transparency x Confidence)		ns $p > .1$	** $p < .01$	*** $p < .001$	. $p < .1$

Table 1. Means for perceptual and behavioral outcome variables across four groups.

Different from our initial conceptual model, Figure 4 also reports the direct relationship between transparency and task performance. Due to an unexpected finding beyond our key hypotheses, we include this relation in our updated conceptual model. We first turn towards perceptual outcomes, namely trust and our hypothesis 2. Participants who interacted with a conversational system transparently communicating its overall system confidence did not report significantly different levels of trust in the system as compared to participants who interacted with a conversational system where transparency statements were absent ($M_{NoTransparency} = 4.14$, $M_{Transparency} = 4.27$, $t(102) = -0.44$, $p = 0.66$). Following, we do not find support for H2. In general, no differences among the four treatment groups regarding trust in the conversational agent can be found ($F(3, 104) = 0.64$, $p = 0.42$).

Turning towards our behavioral outcome variable task performance, participants exposed to transparency statements performed significantly better in the information retrieval task than participants who were not exposed to such statements ($M_{NoTransparency} = 10.4$, $M_{Transparency} = 11.5$, $t(102) = -2.93$, $p < .01$). An ANOVA showed that there a significant differences across the four treatment groups ($F(1, 102) = 8.43$, $p < .01$). Namely, a pairwise comparison depicts a significant difference in task performance between the CG (90%, no transparency) and TG3 (70%, transparency) ($M_{90\%,NoTransparency} = 10.1$, $M_{70\%,Transparency} = 11.6$, $t(52.9) = 3.01$, $p < .1$).

Last, we do not find a significant interaction effect of our two manipulations on our self-reported outcome variable. Our moderation analysis regarding trust in the conversational system ($F(3,104) = 0.646$, $p > .1$) is insignificant.

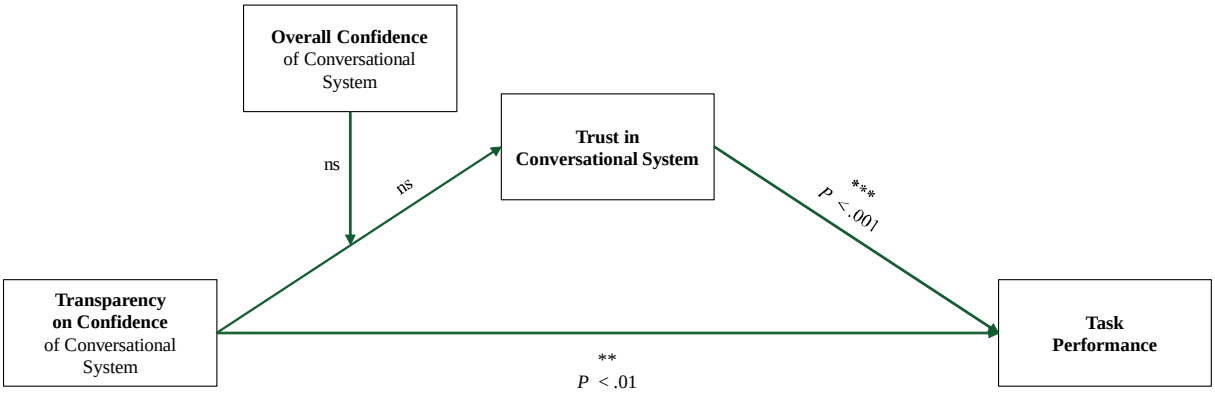

Figure 4. Conceptual model according to study results.

Further testing our theorizing, we estimated a moderated mediation (5,000 bootstrap samples) with the transparency manipulation as the independent variable, user trust (M = 4.21, SD = 1.43) as the mediator, the system confidence manipulation as the moderator, and users' task performance (M = 10.93, SD = 1.98) as the dependent variable.

The standardized regression coefficients between user trust and users' task performance were significant (β_{Trust} = 0.34, SE = 0.13, p < .01). In that sense, the path of the trust mediator on the dependent variable task performance is significant, supporting H1. In addition, we found a direct effect of transparency on system confidence on task performance ($\beta_{Transparency}$ = 1.05, SE = 0.36, t = 2.89, p < .01). Using bootstrapping procedures with 5000 estimations, the standardized indirect effect was insignificant since the bootstrap confidence interval includes zero ($\beta_{Indirect}$ = -0.17, 95%, CI = [-0.75; 0.19]). The standardized regression coefficients between transparency and user trust were insignificant ($\beta_{Transparency}$ = 0.89, SE = 0.92, t = 0.97, p = 0.34), as were the standardized regression coefficients between the interaction term (transparency x system confidence) and user trust ($\beta_{Interaction}$ = -0.50, SE = 0.57, t = -0.88, p = 0.38). We thereby find no support for H3.

5 Discussion of Results

As part of this study, we aimed to explore the interrelatedness of system trustworthiness, user trust, and subsequent behavioral outcomes. We suggest high system confidence and transparency statements on such confidence as two prominent attributes of system trustworthiness to influence user trust in a conversational question-answering system and subsequent user performance in an information retrieval task. In line with extant research on the positive effect of user trust on subsequent, trust-related behavior (Gefen and Straub, 2004), we find that users who trust the conversational system more also perform better in the information retrieval task. This finding strengthens the notion of trust being a crucial prerequisite of sustainable system use and, more importantly, driving desirable performance outcomes for the individual user.

Regarding our transparency manipulation, we do not find a significant effect of transparency statements on users' trust in the conversational system. In that sense, we cannot provide support for the idea of transparency as a means of system trustworthiness to foster trust. More interestingly, our insignificant results might point towards the notion that system trustworthiness per se does not lead to increased user trust. Kästner et al. (2021) provide three reasons for why there is not necessarily a relationship between trustworthiness and trust, including already maximum levels of trust in the system, explanations revealing a problem of the statement, and explanations being incomprehensible or even not useful to evaluate the system. Potentially, users being exposed to the confidence levels of the underlying system model may have been made aware of system issues they otherwise would have not considered. While the effectiveness of our transparency manipulation was successful, users might not have perceived the information communicated by the system as useful. Next to those three reasons, the study context might have been marked by a limited amount of potential risk or personal

damage as a fourth potential reason for the insignificant effect of transparency on user trust. Accordingly, students were not sufficiently personally involved to perceive the transparency statements as reassuring. However, the significant direct effect of transparency statements positively influencing task performance point towards the effectiveness of means of system trustworthiness in driving relevant interaction outcomes, in our case, users' performance in the information retrieval task.

Potential explanations raised for the lacking link between system trustworthiness and trust might also explain the insignificant effect of intent modelling confidence. In fact, students might associate less risk with receiving incorrect or no answer ("I can find this information at some later point in time the course or ask the professor") than initially assumed. With trust being claimed to be highly situation- and context-specific (Holthausen et al., 2020), the question arises whether modelling confidence levels deemed as acceptable in certain contexts and for certain tasks also hold for our information retrieval task in an educational setting. In addition, as our conversational system was based on a predictive model, mentioned challenges around ML-based systems such as uncertainty in outcomes also apply to our deployed conversational system. We could not control for the number and quality of answers the system provided. As a result, there might not have been a noticeable difference between 70% and 90% intent modelling confidence in the eyes of the students. We therefore cannot demonstrate convergence with extant literature having explored accuracy and reliability rates as a driving factor of user trust and subsequent behavior. More so, we cannot confirm previous suggestions of reliability thresholds for trust, as proposed by Yu et al.(2019), for instance.

Hypothesis	Key Findings	Results
H1: *Greater levels of user trust in the system have a positive effect on task performance.*	Users trusting the conversational system more, performed significantly better in the information retrieval task overall, in the multiple choice, and the open-ended questions.	*** $p <$ .001
H2: *Transparency statements on a conversational system's intent modelling confidence lead to enhanced user trust in the conversational system.*	No significant differences regarding user trust in the conversational system could be found between users who were exposed to the transparency statements as compared to those who were not. However, we find a significant direct effect of our mediation analysis of transparency on task performance.	ns ** $p < .01$
H3: *Positive effects of transparency statements are decreased under conditions of lower overall system confidence.*	Our moderated mediation does not find a significant moderating effect of intent modelling confidence on the relationship between transparency statements and user trust in the conversational system.	ns

Table 2. Review of key hypotheses.

6 Implications, Limitations, and Future Research

The results of this study contribute to current discussions on theoretical notions of trust, the implications of specific system trustworthiness measures, as well as the practical implementation of model-based conversational systems.

Our theoretical contribution to the rich literature body on trust within IS and HCI research is twofold. First, our results strengthen extant findings for trust in technology and trust in contemporary, ML-based conversational systems in particular. Our findings emphasize that attitudinal user trust significantly drives behavioral user outcomes beyond intentions to use and reliance (Logg, 2017; McKnight et al., 2011), thereby highlighting the importance of considering user trust as a crucial aspect in system design, as well as distinguishing it both on a conceptional and on an operational level from subsequent behavioral outcomes (Lee and See, 2004).

Second, our research follows a second theoretical differentiation between user trust and specific qualities of system trustworthiness. Our empirical analyses hint towards the claim that measures for trustworthiness might not be suitable for promoting user trust and that there is no automatic cause-and-effect relationship between the two (Jacovi et al., 2021; Lee and See, 2002). Ultimately, we follow recent calls "[inviting] the research community to explore AI explainability specifically for trust calibration [and a] different set of goals in addition to metrics suggested in the current literature such as faithfulness, improved human understanding or acceptance." (Zhang et al., 2020) by looking at two manipulable qualities of system trustworthiness, namely transparency and reliability, as well as the ultimate interaction goal of task performance. Users having to decide whether to trust information provided by ML-based prediction models is a typical interaction context with conversational systems. Thus, we expect our findings to be transferrable to a number of comparable interaction contexts and information retrieval tasks.

From a practitioner's perspective, the results of this study yield insightful implications for the development of and interaction with conversational systems. Understanding how users react to conversational question-answering systems is relevant for commercial developers of such systems, as well as other organizations deploying question-answering systems for educational purposes. While our study does not find prove for transparency statements increasing user trust, we show that both transparency statements and user trust have a positive direct effect on users' task performance. With the ultimate interaction goal in mind, practitioners and developers should leverage transparency measures and ensure that their users trust the respective system at hand to provide effective and successful interaction experiences. While integrating the proposed transparency statements is demonstrated to positively affect users' performance, simply deploying higher levels of intent modelling confidences in underlying interaction models of the conversational system is not as straightforward. Without being able to provide context- or system-agnostic recommendations, training probabilistic models on certain confidences and reliability rates requires extensive testing in the field for a particular use case (D'Amour et al., 2020).

Despite suggested implications and contributions of this study, the presented research and related findings should be interpreted in consideration of several limitations. First, we assumed our experimental context to be a context of trust. Students might associate too little risk with receiving incorrect or no answer from the conversational system. Future research should control for participants' assessment of the situation, i.e., whether they fear unfavorable outcomes or a risky personal involvement. In addition, future research settings could be designed for potential damage or unfavorable outcomes, i.e., the possibility of receiving a reduction in grade. Other potential confounding factors such as existing trust in the university might have inferred the results. Second, our chosen manipulations present two out of many attributes of the system. Despite holding affiliations and university branding constant across conditions, future research could control for students' institutional trust and trust in the university, program, or faculty. The effectiveness of our reliability manipulation could be strengthened by providing participants with an initial baseline confidence level which is stated to be reliable for a specific context and task. Future research could also look at alternative means of driving system trustworthiness, as well as alternative operationalizations of transparency (i.e. illustration of input data used) and confidence (i.e., greater variance of confidence levels). Third, while both transparency statements and user trust positively affect task performance, none of our manipulations were found to influence user trust. Future research could consider additional attributes which might help understand what drives user trust. In a similar vein, underlying cognitive mechanisms which help explain the effect of the transparency statements on task performance should be explored by adding mediating variables such as time spent on task and reported cognitive dissonance.

Acknowledgements

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

References

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*, Elsevier Ltd, 85, 183–189.

Aroyo, A., de Bruyne, J., Dheu, O., Fosch-Villaronga, E., Gudkov, A., Hoch, H., Jones, S., Lutz, C., Sætra, H., Solberg, M. and Tamò-Larrieux, A. (2021). "Overtrusting robots: Setting a research agenda to mitigate overtrust in automation." Paladyn, *Journal of Behavioral Robotics*, 12 (1), 423-436.

Bansal, G., Nushi, B., Kamar, E., Weld, D. S., Lasecki, W.S. and Horvitz, E. (2019). "Updates in Human-AI Teams: Understanding and Addressing the Performance/Compatibility Tradeoff." *Proceedings of the AAAI Conference on Artificial Intelligence*, *33*(01), 2429-2437. https://doi.org/10.1609/aaai.v33i01.33012429.

Bigman, Y.E. and Gray, K. (2018). "People are averse to machines making moral decisions.", *Cognition*, Elsevier B.V., 181, 21–34.

Bird, S., Klein, E. & Loper, E. (2009). *Natural Language Processing with Python*, 1st. Edition. O'Reilly Media, Inc.

Castelo, N., Bos, M. W. and Lehmann, D.R. (2019). "Task-Dependent Algorithm Aversion", *Journal of Marketing Research*, SAGE Publications Ltd, 56 (5), 809–825.

Conrad, F. G., Schober, M.F., Jans, M., Orlowski, R. A., Nielsen, D. and Levenstein, R. (2015). "Comprehension and engagement in survey interviews with virtual agents,", *Frontiers in Psychology*, 6. URL: 10.3389/fpsyg.2015.01578.

D'Amour, A., Heller, K., Moldovan, D., Adlam, B., Alipanahi, B., Beutel, A., Chen, C., *et al.* (2020). "Underspecification Presents Challenges for Credibility in Modern Machine Learning", available at: http://arxiv.org/abs/2011.03395.

Dawes, R. M. (1979). "The Robust Beauty of Improper Linear Models in Decision Making,", *American Psychologist*, 34 (7), 571-582.

Dietvorst, B. J., Simmons, J. P. and Massey, C. (2015). "Algorithm aversion: People erroneously avoid algorithms after seeing them err", *Journal of Experimental Psychology: General*, American Psychological Association Inc., 144 (1), 114–126.

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*, Human Factors and Ergonomics Society, 44 (1), 79–94.

Gambetta, D. G. (1988). "Can We Trust Trust?," *Trust*, Basil Backwell, New York, 213 – 237.

Gapanyuk, Y., Chernobrovkin, S., Leontiev, A., Latkin, I., Belyanova, M., & Morozenkov, O. (2018). "A Hybrid Chatbot System Combining Question Answering and Knowledge-Base Approaches," *AIST*.

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*, Vol. 32 No. 6, pp. 407–424.

Glikson, E. and Woolley, A. W. (2020). "Human trust in artificial intelligence: Review of empirical research," *Academy of Management Annals*, Academy of Management, 14 (2), 627–660.

Hidalgo, C. A., Orghian, D., Albo-Canalds, J., de Almeide, F. and Martin, N. (2021). *How Humans Judge Machines*. MIT Press.

Ilić, A. Ličina, A. and Savić, D. (2020). "Chatbot development using Java tools and libraries," 2020 24th International Conference on Information Technology (IT), 1-4. DOI: 10.1109/IT48810.2020.9070294.

Independent High-Level Expert Group on Artificial Intelligence. (2019). Ethics guidelines for trustworthy AI. Brussels: European Commission. URL: https://ec.europa.eu/newsroom/dae/document.cfm?doc_id=60419

Hamon, R., Junklewitz, H., Sanchez, I., Malgieri, G. and P. De Hert, P. (2002). "Bridging the Gap Between AI and Explainability in the GDPR: Towards Trustworthiness-by-Design in Automated Decision-Making," in *IEEE Computational Intelligence Magazine*, 17 (1), 72-85. DOI: 10.1109/MCI.2021.3129960.

Hoff, K. A. and Bashir, M. (2015). "Trust in automation: Integrating empirical evidence on factors that influence trust", *Human Factors*, SAGE Publications Inc., 57 (3), 407–434.

Hoffman, R. R., Klein, G. and Mueller, S. T. (2018). "Explaining explanation for "explainable AI", *Proceedings of the Human Factors and Ergonomics Society*, 1, 197–201.

Holthausen, B. E., Wintersberger, P., Walker, B. N. and Riener, A. (2020). "Situational Trust Scale for Automated Driving (STS-AD): Development and Initial Validation," in: *12th International Conference on Automotive User Interfaces and Interactive Vehicular Applications (AutomotiveUI '20)*. Association for Computing Machinery, New York, NY, USA, 40–47. DOI: https://doi.org/10.1145/3409120.3410637

Hulland, J. and Houston, M. (2021). "The importance of behavioral outcomes", *Journal of the Academy of Marketing Science*, Springer, 1 May.

Jacovi, A., Marasović, A., Miller, T. and Goldberg, Y. (2021). "Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI," in: *FAccT 2021 - Proceedings of the 2021 ACM Conference on Fairness, Accountability, and Transparency*, Association for Computing Machinery, Inc, 624–635.

Johnson, E.J. and Russo, J.E. (1984). "Product Familiarity and Learning New Information", *Journal of Consumer Research*, 11 (1), 542.

Jussupow, E., Izak, B., and Heinzl, A. (2020). "Why Are We Averse towards Algorithms? A Comprehensive Literature Review on Algorithm Aversion," in: *28th European Conference on Information Systems - Liberty, Equality, and Fraternity in a Digitizing World, ECIS 2020, Marrakech, Morocco, June 15-17, 2020: Proceedings*, edited by Frantz Rowe, RP 168. Atlanta, GA: AISeL. https://aisel.aisnet.org/ecis2020_rp/168.

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*, INFORMS Inst. for Operations Res. and the Management Sciences, 32 (3), 713–735.

Karsenty, L. and Botherel, V. (2005). "Transparency strategies to help users handle system errors", *Speech Communication*, Elsevier, 45 (3), 305–324.

Kästner, L., Langer, M., Lazar, V., Schomäcker, A., Speith, T. and Sterz, S. (2021). "On the Relation of Trust and Explainability: Why to Engineer for Trustworthiness". URL: https://doi.org/10.1109/REW53955.2021.00031.

Kizilcec, R. F. (2016). "How much information? Effects of transparency on trust in an algorithmic interface," in: *Conference on Human Factors in Computing Systems - Proceedings*, Association for Computing Machinery, 2390–2395.

Knote, R., Janson, A., Söllner, M., & 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), 418–458.

Kohn, S. C., de Visser, E. J., Wiese, E., Lee, Y.-C. and Shaw, T. H. (2021). "Measurement of Trust in Automation: A Narrative Review and Reference Guide," in: *Frontiers in Psychology*, Frontiers Media SA, 12. URL: https://doi.org/10.3389/fpsyg.2021.604977.

Kratzwald, B. and Feuerriegel, S. (2019). "Putting Question-Answering Systems into Practice," *ACM Transactions on Management Information Systems*, 9 (4), 1–20. URL: https://doi.org/10.1145/3309706.

Lane, D., Venkatesh, V., Thong, J. Y. L. and Xu, X. (2016). "Unified Theory of Acceptance and Use of Technology: A Synthesis and the Road Ahead," *Journal of the Association for Information Systems*, 17.

Langer, M., König, C. J. and Fitili, A. (2018). "Information as a double-edged sword: The role of computer experience and information on applicant reactions towards novel technologies for personnel selection," *Computers in Human Behavior*, Elsevier Ltd, 81, 19–30.

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

Lee, J. D. and See, K. A. (2002). *Trust in Computer Technology and the Implications for Design and Evaluation*. AAAI Technical Report FS-02-02, available at: www.aaai.org.

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

Liel, Y. and Zalmanson, L. (2020). "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," *International Conference on Information Systems, ICIS 2020 - Making Digital Inclusive: Blending the Local and the Global.*

Logg, J. M. (2017), *Theory of Machine: When Do People Rely on Algorithms?* Harvard Business School Working Paper, No. 17-086, March 2017.

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, 90–103. URL: https://doi.org/10.1016/j.obhdp.2018.12.005.

Luhmann, N. (2000). "Familiarity, Confidence, Trust: Problems and Alternatives," in Gambetta, Diego (ed.) *Trust:* Making and Breaking Cooperative Relations, electronic edition, Department of Sociology, University of Oxford, 6, 94-107.

Madhavan, P. and Wiegmann, D. A. (2004). "A New Look at the Dynamics of Human-Automation Trust: Is Trust in Humans Comparable to Trust in Machines?," in *Proceedings of the Human Factors and Ergonomics Society Annual Meeting*, SAGE Publications, 48 (3), 581–585.

Mayer, R. C., Davis, J. H. and Schoorman, D. F. (1995). "An Integrative Model of Organizational Trust," *The Academy of Management Review*, 20.

McKnight, D. H., Carter, M., Thatcher, J. B. and Clay, P. F. (2011). "Trust in a specific technology: An investigation of its components and measures," *ACM Transactions on Management Information Systems*, 2 (2). URL: https://doi.org/10.1145/1985347.1985353.

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

Meshram, S., Naik, N., Megha, V. R., More, T., & Kharche, S. (2021). "Conversational AI: Chatbots," in : *2021 International Conference on Intelligent Technologies (CONIT)*, 1-6.

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

Mozafari, N., Weiger, W.H. and Hammerschmidt, M. (2021). "Trust me, I'm a bot – repercussions of chatbot disclosure in different service frontline settings," *Journal of Service Management*, Emerald Group Holdings Ltd. URL: https://doi.org/10.1108/JOSM-10-2020-0380.

Nagulendra, S. and Vassileva, J. (2016). "Providing awareness, explanation and control of personalized filtering in a social networking site," *Information Systems Frontiers*, 18 (1), 145-158.

Newman, D. T., Fast, N. J. and Harmon, D. J. (2020). "When eliminating bias isn't fair: Algorithmic reductionism and procedural justice in human resource decisions," *Organizational Behavior and Human Decision Processes*, Academic Press Inc., 160, 149–167.

Ou, C. X., et al. (2014). "Swift Guanxi in Online Marketplaces: The Role of Computer-Mediated Communication Technologies," *MIS Quarterly*, 38 (1), 209–30. URL: https://www.jstor.org/stable/26554875.

Pavlou, P. A., & Gefen, D. (2004). "Building Effective Online Marketplaces with Institution-Based Trust," *Information Systems Research*, *15*(1), 37–59. URL: https://doi.org/10.1287/isre.1040.0015.

Pfeuffer, N., Benlian, A., Gimpel, H. and Hinz, O. (2019). "Anthropomorphic Information Systems," *Business and Information Systems Engineering*, Gabler Verlag, 61 (4), 523–533.

Pickard, M. D., Roster, C. A. and Chen, Y. (2016). "Revealing sensitive information in personal interviews: Is self-disclosure easier with humans or avatars and under what conditions?," *Computers in Human Behavior*, 65, 23-30. URL: https://doi.org/10.1016/j.chb.2016.08.004.

Puranam, P., & Vanneste, B. (2021). "Artificial Intelligence, Trust, and Perceptions of Agency," 2021/42/STR, Available at SSRN: https://ssrn.com/abstract=3897704 or http://dx.doi.org/10.2139/ssrn.3897704.

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), 203–212.

Rousseau, D. M., Sitkin, S. B., Burt, R. S., and Camerer, C. (1998). "Not So Different After All: A Cross-Discipline View of Trust," *Academy of Management Review*, 23(3), 393 - 404.

Rubin, V. L., Chen, Y. and Thorimbert, L. M. (2010). "Artificially intelligent conversational agents in libraries," *Library Hi Tech*, 28 (4), 496–522.

Saffarizadeh, K., Boodraj, M., & Alashoor, T. M. (2017). "Conversational Assistants: Investigating Privacy Concerns, Trust, and Self-Disclosure," *Thirty Eighth International Conference on Information Systems.*

Schmitt, A., Wambsganss, T., Soellner, 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,* Austin, United States of America.

Schuetzler, R. M, Giboney, J. S., Grimes, G. M. and Nunamaker, J. F. (2018). "The influence of conversational agent embodiment and conversational relevance on socially desirable responding," *Decision Support Systems*, 114, 94-102. URL: https://doi.org/10.1016/j.dss.2018.08.011.

Shawar, B. A. and Atwell, E.S. (2005). "Using corpora in machine-learning chatbot systems," *International Journal of Corpus Linguistics*, John Benjamins Publishing Company, 10 (4), 489–516.

Söllner, M. (2020). "Same same but different? A Two-Foci perspective on trust in information systems," *Proceedings of the Annual Hawaii International Conference on System Sciences*, IEEE Computer Society, 5129–5138.

Stieglitz, S., Mirbabaie, M., Möllmann, N. & Rzyski, J. (2021). "Collaborating with Virtual Assistants in Organizations: Analyzing Social Loafing Tendencies and Responsibility Attribution," *Information Systems Frontiers*, 1-26. DOI: 10.1007/s10796-021-10201-0.

Sumikawa, Y., Fujiyoshi, M., Hatakeyama, H. and Nagai, M. (2019). "Supporting creation of FAQ dataset for e-learning Chatbot," *Smart Innovation, Systems and Technologies*, 142, Springer Science and Business Media Deutschland GmbH, 3–13.

Turel, O. and Gefen, D. (2013). "The Dual Role Of Trust In System Use," *Journal of Computer Information Systems*, 54 (1), 2-10. DOI: 10.1080/08874417.2013.11645666.

Wambsganss, T., Haas, L. and Söllner, M. (2021a). "Towards the Design of a Student-Centered Question-Answering System in Educational Settings," in *Twenty-Ninth European Conference on Information Systems (ECIS 2021).*

Wambsganss, T., Kung, T., Sollner, M. and Leimeister, J. M. (2021b). "Arguetutor: An adaptive dialog-based learning system for argumentation skills," *Conference on Human Factors in Computing Systems - Proceedings*, Association for Computing Machinery. URL: https://doi.org/10.1145/3411764.3445781.

Wang, W. and Benbasat, I. (2005). "Trust in and Adoption of Online Recommendation Agents," *Journal of the Association for Information Systems*, 6 (3). DOI: 10.17705/1jais.00065.

Weber, F., Wambsganss, T., Rüttimann, D. and Söllner, M. (2021). "Pedagogical Agents for Interactive Learning: A Taxonomy of Conversational Agents in Education," *International Conference on Information Systems,* Austin, United States of America.

Wienrich, C., Reitelbach, C. and Carolus, A. (2021). "The Trustworthiness of Voice Assistants in the Context of Healthcare Investigating the Effect of Perceived Expertise on the Trustworthiness of Voice Assistants, Providers, Data Receivers, and Automatic Speech Recognition," *Frontiers in Computer Science*, Frontiers Media SA, 3. URL: https://doi.org/10.3389/fcomp.2021.685250.

Winkler, R., Söllner, M. and Leimeister, J. M. (2021). "Enhancing problem-solving skills with smart personal assistant technology," *Computers and Education*, Elsevier 165. URL: https://doi.org/10.1016/j.compedu.2021.104148.

Wollny, S., Schneider, J., Di Mitri, D., Weidlich J., Rittberger, M. and Drachsler, H. (2021). "Are We There Yet? - A Systematic Literature Review on Chatbots in Education," *Front. Artif. Intell*, 4. DOI: 10.3389/frai.2021.654924.

Xu, W. (2019). "Toward Human-Centered AI: A Perspective from Human-Computer Interaction," *Interactions*, 26 (4), 42-46. DOI:10.1145/3328485.

Yeomans, M., Shah, A., Mullainathan, S. and Kleinberg, J. (2019). "Making sense of recommendations," *Journal of Behavioral Decision Making*, John Wiley and Sons Ltd, 32 (4), 403–414.

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*, Association for Computing Machinery. URL: https://doi.org/10.1145/3290605.3300509.

Yu, K., Berkovsky, S., Taib, R., Zhou, J. and Chen, F. (2019). "Do I trust my machine teammate? An investigation from perception to decision," *International Conference on Intelligent User Interfaces, Proceedings IUI*, Vol. Part F147615, Association for Computing Machinery, 460–468.

Yu, K., Taib, R., Berkovsky, S., Zhou, J., Conway, D. and Chen, F. (2016). "Trust and Reliance based on system accuracy," *UMAP 2016 - Proceedings of the 2016 Conference on User Modeling Adaptation and Personalization*, Association for Computing Machinery, Inc, 223–227.

Zhang, Y., Vera Liao, Q. and Bellamy, R. K. E. (2020). "Effect 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*, Association for Computing Machinery, Inc, 295–305.

Zhou, J. and Chen, F. (2018). "Towards Trustworthy Human-AI Teaming under Uncertainty," International Joint Conference on Artificial Intelligence. URL: http://hdl.handle.net/10453/136189.

Zhou, J., Sun, J., Chen, F., Wang, Y., Taib, R., Khawaji, A. and Li, Z. (2015). "Measurable decision making with GSR and pupillary analysis for intelligent user interface," *ACM Transactions on Computer-Human Interaction*, Association for Computing Machinery, 21 (6). URL: https://doi.org/10.1145/2687924.