How Much Reliability Is Enough? A Context-Specific View on Human Interaction With (Artificial) Agents From Different Perspectives

Ksenia Appelganc, Department of Business Psychology and Human Resource Management, University of Bremen, Bremen, Germany, **Tobias Rieger**, **Eileen Roesler** and **Dietrich Manzey**, Department of Psychology and Ergonomics, Technische Universität Berlin, Berlin, Germany

Tasks classically performed by human–human teams in today's workplaces are increasingly given to human–technology teams instead. The role of technology is not only played by classic decision support systems (DSSs) but more and more by artificial intelligence (AI). Reliability is a key factor influencing trust in technology. Therefore, we investigated the reliability participants require in order to perceive the support agents (human, AI, and DSS) as "highly reliable." We then examined how trust differed between these highly reliable agents. Whilst there is a range of research identifying trust as an important determinant in human–DSS interaction, the question is whether these findings are also applicable to the interaction between humans and AI. To study these issues, we conducted an experiment (N = 300) with two different tasks, usually performed by dyadic teams (loan assignment and x-ray screening), from two different perspectives (i.e., working together or being evaluated by the agent). In contrast to our hypotheses, the required reliability if working together was equal regardless of the agent. Nevertheless, participants trusted the human more than an AI or DSS. They also required that AI be more reliable than a human when used to evaluate themselves, thus illustrating the importance of changing perspective.

Introduction

In recent decades, artificial intelligence (AI) has not only entered our everyday lives but it has also become of critical importance to our working environments. Professionals in a wide range of fields like the finance sector (Bahrammirzaee, 2010; O'Neil, 2020) or healthcare (Bejnordi et al., 2017; McKinney et al., 2020) are now supported by AI: Medical practitioners are using AI to screen x-rays and help them detect cancer (McKinney et al., 2020), recruiters are being supported by AI finding appropriate candidates (Langer et al., 2019), and loan decisions (O'Neil, 2020) are made with the support of AI. As a result, the collaboration and decisions made by these teams of humans and AI have a major impact on many of our lives (for a review, see Langer & Landers, 2021).

The support of technical automated systems of humans at work is nothing new. Decision support systems (DSS) have been used for decades in a plethora of fields (Power, 2008) and are of particular importance within the field of human–automation interaction research (Endsley, 2017). According to Mosier and Manzey (2020), they are "designed as a support tool for humans …[and] provide at discrete points in time, either automatically or on demand, certain information about the state of the world that can improve informed decision-making" (p. 19). These classical DSSs usually have predefined algorithms and parameters, whereas the advancement of AI has led to a new kind of intelligent DSS that supports humans (Bini, 2018). To define AI in this paper we focus on the definition of a subset of AI, machine learning, that imitates human intelligence by using computational algorithms to recognize patterns in big data sets (Bini, 2018). Despite the omnipresence of AI, there is limited research on whether the presented technological differences

Address correspondence to Ksenia Appelganc, University of Bremen, Enrique-Schmidt-Straße 1, 28359 Bremen/Germany.
Email: appelganc@uni-bremen.de

Journal of Cognitive Engineering and Decision Making
Vol. 0, No. 0, ■■ ■, pp. 1-15
DOI:10.1177/15553434221104615
Article reuse guidelines: sagepub.com/journals-permissions

can also lead to differences in the perception and usage of AI by humans, compared to classical DSS. However, many earlier findings from human–automation interaction studies may well be an adequate starting point for a better understanding of human–AI interaction.

INTERSECTION of HUMANS WITH TECHNICAL SYSTEMS

An important factor of successful interaction in both human–human and human–DSS teams is trust (Hoff & Bashir, 2015). Trust has been widely researched and this research shows that it fundamentally influences the use of automated systems (Hoff & Bashir, 2015). Inappropriate levels of trust can lead to crucial behavioral consequences like misuse or disuse of DSSs which are typically investigated in human–automation research (Parasuraman & Riley, 1997). However, the question arises as to whether these findings can be applied to the research of interactions between humans and AI.

In order to approach this question, it is first necessary to delineate the differences between AI and classical DSS. One major difference that distinguishes AI from classical DSS is AI's ability to continuously learn and improve. This could plausibly lead to the perception of AI being a more competent and advanced expert system, thus being perceived as superior in comparison to classical DSS (e.g., Lerch et al., 1997). Furthermore, according to Legaspi et al. (2019), AI is usually considered to have some sort of agency. Originally, agency has been a unique human characteristic. Attributing this characteristic to a technical system could again contribute to a perception of increased superiority of AI over DSS. Whereas the comparison of DSS and humans has a long history in human–automation research (Dzindolet et al., 2002; Madhavan & Wiegmann, 2007), it is unclear what will change if technologies like AI combine the aforementioned features of both these agents, that is, DSS (still being a technical support system) and humans (ability to learn and agency).

According to Dzindolet et al. (2002), humans have different expectations of different interaction partners (i.e., human vs. automated aid). More precisely, humans already tend to initially expect automated aids to function flawlessly. This means they overestimate an automated aids' reliability, which is an effect referred to as *perfect automation schema*. In contrast, decision support by another human is supposed to be perceived as less reliable because people are aware of their own fallibility as humans (Dzindolet et al., 2002). Looking at the differences between AI and DSS, this perfect automation schema could be more pronounced for AI because AI could easily be seen as the greater expert system (Alfonseca et al., 2021; Heer, 2019).

A large body of research on trust in automation suggests that an aids' reliability is one of the main determinants influencing trust (e.g., Hancock et al., 2011; Hoff & Bashir, 2015; Madhavan & Wiegmann, 2007; Parasuraman & Riley, 1997). However, not only the actual (objective) reliability but also the subjectively perceived reliability of support systems needs to be considered. This is because the perceived reliability can differ from the actual reliability (Rieger et al., 2022) and might therefore lead to differences in trust. In terms of perceived and actual reliability, the perfect automation schema assumes that humans have different initial expectations of the actual reliability of different support systems (Dzindolet et al., 2002). Bearing in mind the awareness of how error-prone humans are, decision support provided by humans (e.g., advice by a colleague) is usually expected to be imperfect and, thus, considerably less than 100% reliable. In contrast, much more is expected of technical systems and specifically automated support systems which often give humans different ideas about a system's fallibility. Consequently, this might result in different expectations when it comes to the reliability of human aids compared to automated aids, regardless of the objective reliability.

Against this background of an assumed differing perception of reliability between humans and automated aids, the first goal of the current research was to examine how well a support agent (i.e., human, AI, and DSS) had to perform in order to be perceived as highly reliable. Whilst there is already some research comparing the interaction with human aids and classical

DSS (for a review, see Madhavan & Wiegmann, 2007), much less is known about how AI is perceived in this interaction. However, as argued above, one might expect the perfect automation schema to be even more pronounced in interaction with AI compared to DSS since AI-based systems have self-learning capabilities and ascribed agency (Bini, 2018; Legaspi et al., 2019). As a consequence, this might also be reflected in differing levels of trust, independent of whether the objective reliability might be the same. Moreover, other factors besides perceived reliability such as the reputation of the agent and attitudes towards it might also affect the extent to which humans will trust it (Hoff & Bashir, 2015). Therefore, our second research goal was to find out whether trust differs between agents even though they each fulfill the requirements of a *highly* reliable support system (i.e., human vs. DSS vs. AI).

COLLABORATION WITH or EVALUATION by (ARTIFICIAL) AGENTS

In dealing with issues of required reliability and trust of different decision support agents, one needs to take further into account that humans are not only working together with such agents but can also be evaluated by them (O'Neil, 2020). However, previous human–automation interaction research has typically only addressed issues of trust in automation from the perspective of humans making decisions in collaboration with some kind of automated support agent (Janssen et al., 2019). Langer and Landers (2021) describe this perspective as the "first party" perspective. In addition, they describe the "second party" perspective as an important stakeholder: For instance, second parties are job applicants who are evaluated by automated systems or patients who are diagnosed by automated systems. The effect of this change of perspective has not yet been the focus of human–automation research. Specifically, the question arises as to whether there are other reliability requirements when being evaluated by a human as opposed to a technical system. Whilst an assumed technological superiority is expected to shape our image of required reliability from an advice-taker's perspective, this may well be different from the perspective of the person who gets assessed by the system. For instance, Bonnefon and colleagues (2016) have found that even though people preferred that others use a certain kind of AI, they would not want to use it themselves. Moreover, a recent study comparing the preferences for different support agents (i.e., AI, DSS, and human) from the second party's perspective found that technological systems are favored over humans (Rieger et al., 2022). Surprisingly this stood in clear contrast to the first party's perspective where a greater trust was placed in the human aid (Rieger et al., 2022). Still, it remains unclear why the change of perspective led to opposing results. Perhaps, differences in the required reliability of different agents could be a reason for this. Whereas the assumed fallibility of a human (Dzindolet et al., 2002) can be compensated by the interaction partner in a collaborative setting, it has potentially drastic effects when being (exclusively) evaluated. Thus, another aim of the present research was to explore possible perspective-related differences on how automated decision-making (either performed by classical DSS or AI-based systems) is perceived compared to a human aid.

TASK FEATURES

Our final objective was to investigate the task features and the context in which the interaction took place. Thus far, the research on trust in automation (Dzindolet et al., 2002; Hoff & Bashir, 2015; Lyons et al., 2018) has mainly involved automated systems which support some sort of (visual) detection task (e.g., luggage screening or specific patterns in x-rays; e.g., Huegli et al., 2020, Rieger & Manzey, 2020; Rieger et al., 2021). These tasks usually have directly quantifiable results and may be perceived to suit the abilities of an automated system better than a human. This might be a possible explanation for the perfect automation schema. It is currently unclear if this will also hold true for tasks which are more subjective and are considered to be based more on human intuition (Castelo et al., 2019). According to Castelo et al. (2019), humans tend to trust algorithms more for objective than subjective

tasks, as subjective tasks leave more room for interpretation. However, as modern AI is making inroads into areas originally associated with human intuition (e.g., personnel selection, university applications) (O'Neil, 2016), we included not only an objective visual detection task (i.e., x-ray evaluation) but also a more subjective decision-making task (i.e., loan assignment). It is questionable whether required reliability will be comparable in this instance, even though the tasks differ in their objectivity.

Tasks like this vary not only in respect to their quantifiability but also in the situational risk associated with them (Stuck et al., 2021). Situational risk is domain-specific (Weber et al., 2002) and events involving health or public safety concerns are perceived as riskier than events associated with financial matters (Weber et al., 2002). Moreover, Jacovi et al. (2021) highlight the importance of risk in the trust assessment of human–AI interaction. That is because a perceived vulnerability and therefore differences in risk for a certain outcome can influence trust (Jacovi et al., 2021). Therefore, risk might also affect the required reliability as well as trust in support agents.

Furthermore, it is important to consider the perceived difficulty of a task as it also might play an important role. It has already been shown that humans tend to depend more on support if they perceive the task as more difficult (Maltz & Shinar, 2003). Additionally, the difficulty leads to a higher perception of the agent's reliability (Madhavan & Wiegmann, 2007). Since the tasks used in this experiment seem to clearly vary in regard to their objectivity and risk, we further wanted to examine the perceived task difficulty.

PRESENT RESEARCH

To address the issues mentioned above, we conducted an experiment to investigate required reliability and resulting trust in different (support) agents (i.e., human vs. DSS vs. AI). This was done within two task contexts and from the first and second party's perspective.

The first task involved a simulated x-ray screening in the context of radiology. This screening task can be considered to be an objective task as there are clearly correct and incorrect decisions based on an inspection of the x-rays. The second context involved the more subjective task, that is, a simulated decision about a private loan provided by a bank. This task is more subjective because decisions on providing a loan are regarded to require human empathy and are influenced by the relationship between the bank employee and the applicant. That is consistent with Castelo et al.'s (2019) description of the subjective task as needing intuition and being based more on personal opinion. The stimuli used for both tasks are illustrated in Figure 1.

We varied the perspectives by asking the participants to imagine working together with the support agent or being evaluated by the agent. More precisely, the participants should first imagine that they are the key decision-maker—with support from an agent—regarding an x-ray evaluation or a loan request (first party). Then, the participants were asked to change their role and to imagine having their x-ray or loan request evaluated by the respective agent (second party). For both perspectives, they were asked to indicate the required reliability.

Following the perfect automation schema (Dzindolet et al., 2002), we hypothesized that participants would require the lowest reliability, that is, fewest correctly evaluated cases, from the human as they are aware of their fallibility. Considering our presented AI (i.e., ability to learn) being more of an expert system than the classical DSS (i.e., based on predefined algorithms), we expected the highest required reliability for AI. We assumed this to be the case when working together as well as being evaluated by the agent. After the assessment of the required reliability, we asked participants how much they would trust the agent if the agent was known to be *highly* reliable, that is, it had yielded the exact rate of correct decisions as chosen previously. In contrast to the perfect automation schema (Dzindolet et al., 2002), we expected trust to be higher in humans compared to DSS and AI. This hypothesis was suggested by recent results of Rieger et al. (2022) who also conducted a study comparing trust in different support agents. They showed that participants trusted a human aid the most, although the human aid was exactly as reliable as the AI and

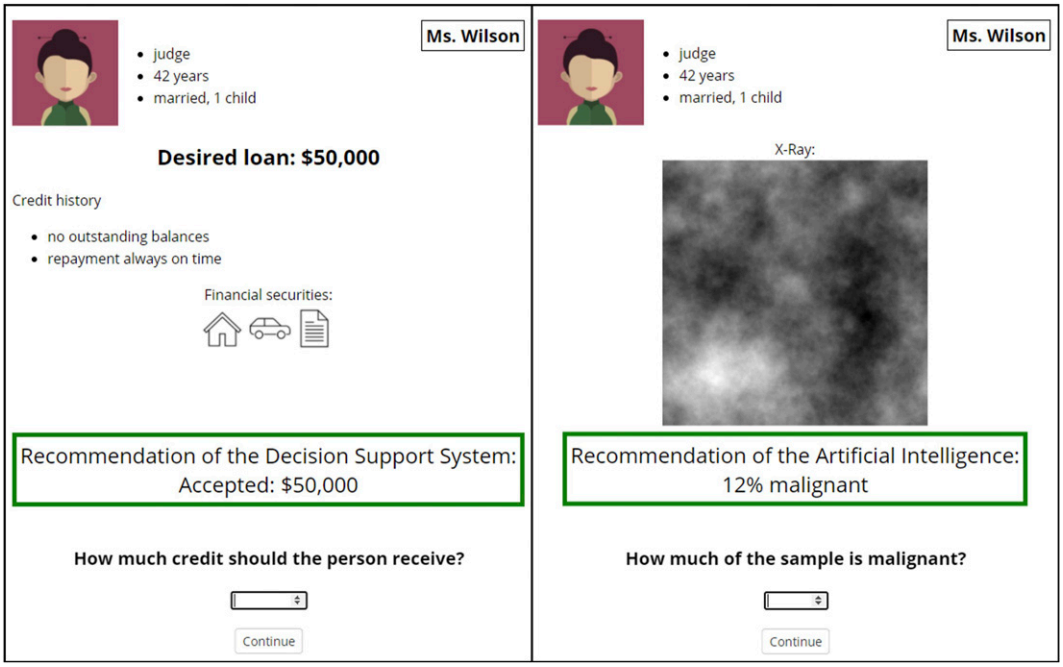

Figure 1. Depiction of the trials.
Note. Decision support system condition in the context of loan assignment (left) and the AI condition in the context of radiology (right). The human condition looked the same with "colleague" replacing the respective support agent.

DSS. The hypothesis is also supported by recent findings of Langer et al. (2021) who showed that participants trusted a human aid more than an automated aid in the context of personnel selection in an HR department.

Furthermore, we assumed that, independent of the specific perspective, the required reliability would be higher in the radiology task context because of the perceived higher risk associated with a wrong decision. In contrast, we expected trust to be lower in the radiology context, even if the aid worked highly reliably. This was assumed in line with earlier results (Rieger et al., 2022) because of the more fatal consequences in case of wrong decisions. Finally, we hypothesized a higher perceived difficulty for the x-ray task, as it is more objective and fits less to the abilities of the participant than the (automated) support.

METHOD

The experiment was pre-registered via the Open Science Framework (https://osf.io/h98cj/) and approved by the local ethics committee. The experiment was programmed using jspsych (de Leeuw, 2014) and was run on a JATOS (Lange et al., 2015) server so that participants could individually run the experiment in their browser.

Participants

We used the software program G*Power (Faul et al., 2007) to conduct an a priori power analysis. Our goal was to obtain close to .90 power to detect a small to medium effect size of .20 at the standard .05 alpha error probability. Results showed that a sample of 300 participants was required, given a between-subjects design.

341 participants took part in the online experiment. They were recruited via Prolific and received £0.80 for the approximately 10-minute experiment. Participants were randomly and equally assigned to one of six experimental conditions. Forty one participants had to be excluded from further analysis due to missed attention checks, resulting in a final sample of 300 participants (mean age = 33.55, SD = 5.36; 44%

female). There were no differences (p = .283) in the control variable disposition to trust technology between the participants in each condition.

Design

We used a 2 (context: loan assignment vs. radiology) × 3 (support agent: human vs. AI vs. DSS) between-subjects design. In the context of loan assignment, the task was to decide whether an applicant should get a desired loan, and if yes, what amount they should receive. The task in the radiology context was the estimation of the percentage of potentially malignant tissue in simulated x-rays of different individuals. The perception of the support agent was manipulated through framings including the designation of the agent, that is, "AI system," "DSS," or "colleague," as well as a brief description.

Procedure

First, participants were introduced to the aim of the study and started the experiment by giving their informed consent. Thereupon, they were given a short general introduction to their task. In the context of loan assignment, the description of the task included relevant information that should be taken into account when making a loan decision (income, job, family situation, debts, possessions, and place of residence). For the task in the radiology context, we used simulated $1/f^3$ noise as stimulus material as this resembles the power spectrum of real-world mammograms (Burgess et al., 2001). An example x-ray image and a grayscale continuum were presented with the critical cutoff that distinguishes between non-malignant and potentially malignant tissue. Additionally, participants were informed that the cutoff is very cautious and a percentage lower than 15% is normally not considered as concerning.

Subsequently, three examples were shown of the respective task including one correct and two incorrect decisions that were made. All examples were presented in the same format as the actual practice trials. The stimuli used for the two tasks were the same across all agent conditions and are illustrated in Figure 1. In the loan assignment context, the correct example was presented as an applicant (represented by a persona) who applied for a credit and correctly received the desired amount. In the two incorrect decision examples, a credit was either rejected although the creditworthiness of the applicant was high (first example), or approved although the application was inappropriate, that is, the applicant had a low creditworthiness (second example). In the context of radiology, the correct example showed a correct evaluation of 60% malignant tissue in a patient's x-ray that led to a correct treatment. The incorrect examples were presented as either an x-ray of a healthy patient that was evaluated with much more malignant tissue than there actually was (first example), or an unhealthy patient whose x-ray was evaluated with much less than the actual malignant tissue was (second example). Participants were informed that these examples are only presented for understanding the task and to get familiarized with the decision criteria. Each of the incorrect examples was accompanied by descriptions of possible consequences that could be detrimental to both, the bank or radiology practice and the applicant or patient, respectively. All examples were presented in the same format as the credit applications or x-rays that the participants received later when making their own decisions with the help of the support agent during the data collection.

The examples were then followed by specific framings of the support agent that would support the participant in making their decision (first party perspective). The framings were presented via written text. The AI was described as being based on deep neural networks that had the ability to continuously analyze previous decisions and outcomes. For the DSS, the description highlighted that the DSS was based on prior data and decided by using predefined parameters and algorithms. The human agent was described as a colleague who has been working in the field for many years and has considerable experience. Participants were informed that the agents were highly reliable in the practice trials, but that in an actual case the responsibility for making the final decision would be with the participant. Thus, they could disagree with the support agent. To test the framing, two attention checks followed, which asked the participants about the type of the support agent and the characterization that was given in the description before.

After the attention checks, participants got further information about how the task would look like and how exactly to perform it. In addition, they were told that the support agent's advice would be correct in all following cases and that this was just to show them how the interaction would work and to allow for a better understanding of the interaction with the support agent.

Participants then started with the first of a total of five trials which all required the participant to interact with the support agent from the first party perspective. Each trial was structured as follows: the loan applicant or patient was presented with the desired loan request or x-ray and personal information (which were kept the same for both contexts) for a minimum of 5 seconds. During these 5 seconds, participants were told that the support agent is also, at that time, evaluating (colleague) or processing (AI and DSS) the information. Subsequently, participants viewed the recommendation of the support agent and continued by pressing the spacebar. As shown in Figure 1, the applicant's or the patient's relevant information stayed visible during the whole trial in addition to the support agent's recommendation. After seeing the recommendation, participants were asked to type in their decision. Although participants were previously informed that the scenarios will all be done with the correct recommendation from their support agent, they still had to make the final decision. Specifically, they needed to type in the amount of the credit or the percentage of the malignant tissue. They continued with the next trials by pressing the spacebar again whenever they felt ready. The recommendation had a color-coded frame. In the loan assignment context, the colors indicated whether the loan was fully approved = green, partially approved = orange, or rejected = red. In the radiology context, they showed if the percentage of malignant tissue was under 15% = green, between 15% and 50% = orange, or above 50% = red. Note that these five trials were only included to give the participants a better understanding of the respective agents.

After completing all five trials, the dependent variables were collected by means of questionnaires which were all presented one by one. This first comprised the first party's perspective with the required reliability and trust. Subsequently, participants were asked to take the second party's perspective. Here again, required reliability was measured. Additionally, participants needed to decide if they would prefer to be evaluated by a purely human or human–automation team. Then, participants filled out the disposition to trust technology questionnaire (Lankton, et al., 2015) that we used as a control variable. At last, they answered demographic questions (age and gender).

Dependent Variables

First party perspective. In order to assess the required reliability, participants were told to imagine working together with their support agent on the respective task (i.e., loan assignment and radiology). They should then decide how many cases out of 100 cases needed to be evaluated correctly by the support agent in order to perceive the support agent as highly reliable (0–100 cases). In a second step, participants were asked how much they would trust the support agent (0 = not at all to 100 = completely) if the support agent correctly evaluated the number of cases they indicated beforehand as highly reliable (i.e., referring to their individual assessment).

Second party perspective. The required reliability was measured by instructing participants to imagine that a decision would be made concerning their own loan application or their own x-ray exclusively by the respective agent. Here, we again asked participants how many cases needed to be evaluated correctly to accept their application or x-ray to be evaluated (0–100 cases) by the agent. They could also choose the additional option that they would never accept being exclusively evaluated by this agent ("never"—option). Furthermore, participants should decide if they would rather have their application assessed or x-ray evaluated by a team of human–human or human–automation (seven-point semantic differential scale). This was done to measure the preference for a purely human or a mixed human–automation team. Participants who experienced working with a colleague were presented with the option

human–human or human–automation while participants who worked with AI or DSS could choose between human–human and the support agent they have been working with (i.e., human–AI or human–DSS).

Task difficulty. Additionally, the difficulty of the task was considered to investigate the influence of this specific task feature. The difficulty was assessed by asking participants how difficult they perceived the task without the support of their agent (seven-point Likert scale from "Not difficult at all" to "Extremely difficult").

RESULTS

All 300 participants who were included in the analysis had passed the attention check. We analyzed all dependent variables except for trust by performing a 2 (context) × 3 (support agent) between-subjects ANOVA. Additional post-hoc tests using Bonferroni corrected *p*-values for multiple comparisons were used for further analysis if needed.

First Party Perspective: Collaboration with a Support Agent

The analysis of required reliability when working with the support agent is visualized in Figure 2. It showed a significant main effect of context ($F(1,294) = 7.79$, $p = .006$, $\eta_G^2 = .026$) with a higher required reliability for the x-ray assessment task ($M = 85.8$, $SE = 1.04$) compared to the loan decision task ($M = 80.7$, $SE = 1.50$). In contrast to our assumptions, no significant differences in required reliability emerged between the support agents, $F(2,294) = 1.55$, $p = .215$, $\eta_G^2 = .010$. To further support this null-effect, we conducted a parallel Bayesian ANOVA using the BayesFactor package. This analysis revealed that given the present data, a null-effect is about 7 times more likely than an alternative hypothesis ($BF_{01} = 7.06$), further strengthening the claim that the required reliability was equivalent between the support agents.

To analyze subjective trust, we conducted a 3 (support agent) × 2 (context) ANCOVA with required reliability as the covariate. The estimated marginal means of the corresponding trust measures controlled for required reliability are presented in Figure 3. The analysis revealed the opposite finding compared to the analysis of the required reliability. While context showed no significant effect on trust ($F(1, 293) = 0.69$, $p = .406$, $\eta_G^2 = .002$), there was a significant effect of support agent on trust while controlling for the effect of differences in individually required reliability, $F(2,293) = 15.75$, $p < .001$, $\eta_G^2 = .097$. Post-hoc contrasts revealed that even with control of the individual differences in required reliability trust differed between the support agents. That is, mean trust in the support by another human ($M = 89.1$, $SE = 1.08$) was significantly higher ($p = .010$) than trust in AI ($M = 84.6$, $SE = 1.08$), and also significantly higher ($p < .001$) compared to trust in DSS ($M = 80.6$, $SE = 1.08$). Furthermore, mean trust in DSS was significantly lower ($p = .030$) than mean trust in AI. Finally, as expected, there was also a significant main effect of context on task difficulty, $F(1,294) = 164.18$, $p < .001$, $\eta_G^2 = .358$. Participants perceived the x-ray assessment task ($M = 4.5$, $SE = 0.12$) as more difficult than the loan decision task ($M = 2.0$, $SE = 0.14$).

Second Party Perspective: Evaluation by an (Artificial) Agent

Figure 4 shows the analysis of the reliability that participants required to accept an exclusive evaluation by the respective support agent. As hypothesized, there was a main effect of context, $F(1,204) = 6.27$; $p = .013$, $\eta_G^2 = .030$. Similar to what we found for the first party's perspective, participants, on average, require a higher reliability when having their x-rays ($M = 90.8$, $SE = 1.12$) exclusively evaluated by the support agent compared to the evaluation of loan applications ($M = 85.2$, $SE = 1.78$). In addition, the main effect of support agent was significant, $F(2,204) = 4.11$; $p = .018$, $\eta_G^2 = .039$. In line with our hypothesis, there was a significantly higher required reliability ($p = .014$) for the exclusive evaluation by the AI ($M = 91.5$, $SE = 1.53$) compared to the human ($M = 83.3$, $SE = 1.95$). However, no significant differences in the required reliability were found between the AI and the DSS ($p = .569$), as well as between the DSS and the human ($p = .373$). Chi-square tests of

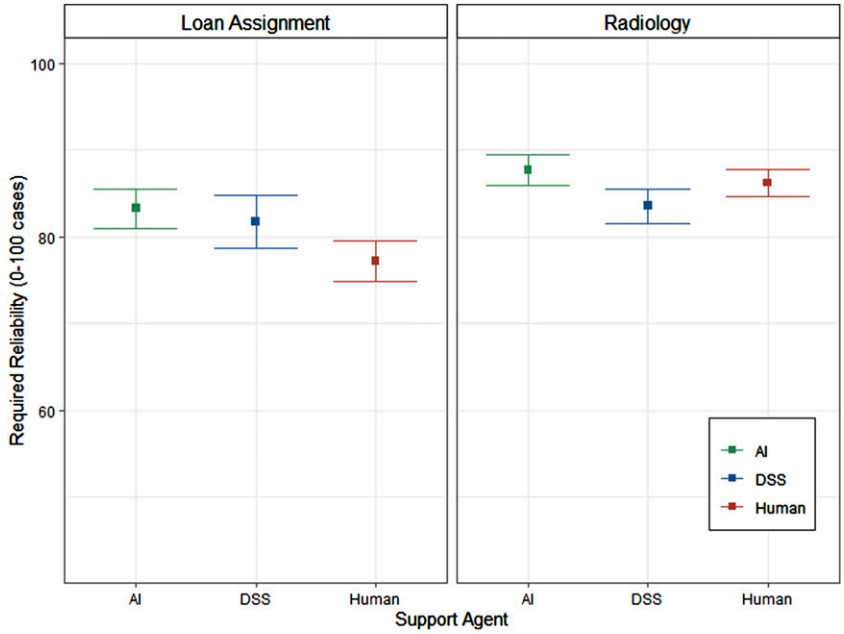

Figure 2. Analysis of required reliability.
Note. Means and standard errors for required reliability.

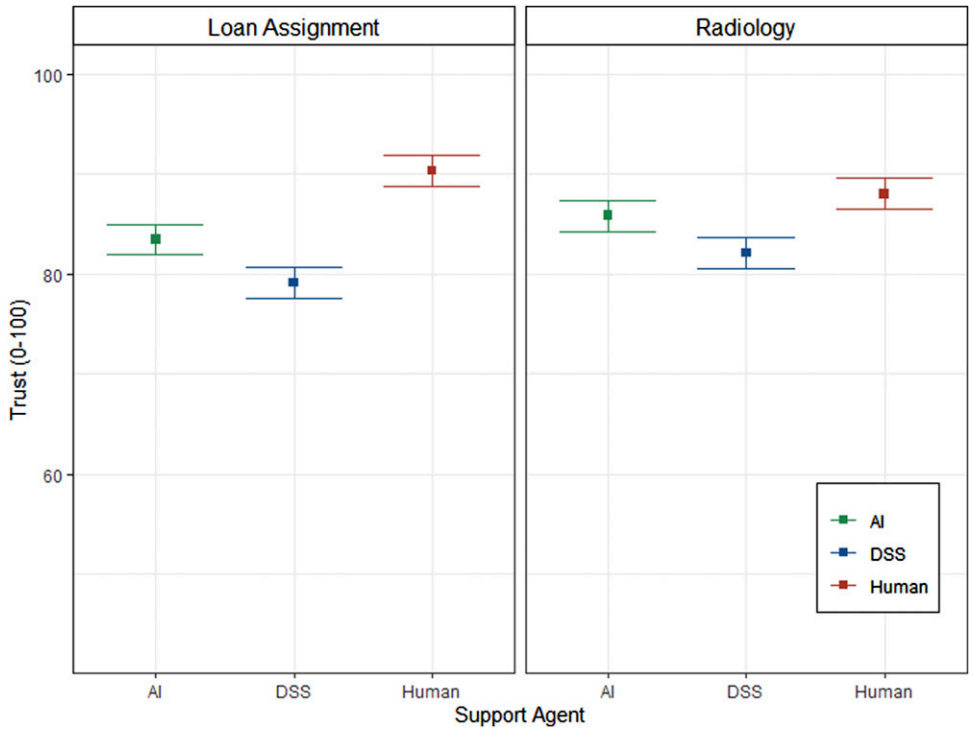

Figure 3. Analysis of Trust with Required Reliability as Covariate.
Note. Estimated marginal means and standard errors for trust.

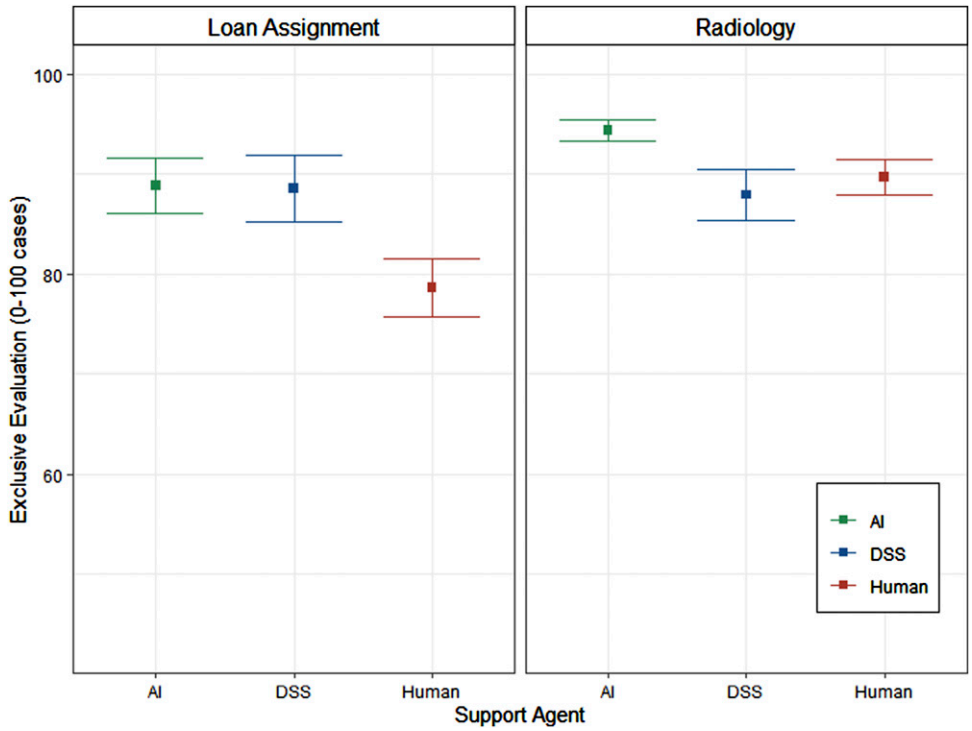

Figure 4. Analysis of required reliability for exclusive evaluation.
Note. Means and standard errors for required reliability in exclusive evaluation.

independence with Yates's continuity correction revealed no significant association between support agent and "never"—option (i.e., decision to never exclusively being evaluated by the support agent) (p = .368) as well as no significant association between context and "never"—option (p = .166).

The final analysis addressed the preference for being evaluated by a human–human or a human–automation team. The results are depicted in Figure 5. There was a main effect of context (F (1,294) = 21.61, p < .001, η_G^2 = .068) indicating a stronger preference for a team of human and automation in the context of radiology (M = 4.0, SE = 0.16) compared to loan assignment (M = 2.9, SE = 0.18). Moreover, there was a main effect of support agent, F (2,294) = 17.06, p < .001, η_G^2 = .104. Participants who experienced working with the DSS preferred to be evaluated by a team of human and automation (M = 3.7, SE = 0.20) significantly more (p < .001), compared to participants who worked with another human (M = 2.5, SE = 0.21). This was also the case (p < .001) for participants who worked with an AI (M = 4.1, SE = 0.20), compared to participants who worked with a human. There was no significant difference (p = .578) in preference between participants who worked with a DSS or an AI.

DISCUSSION

The present experiment addressed issues of subjectively required reliability and trust in different (support) agents, dependent on the task context and from two perspectives. On the one hand, we investigated from the first party's perspective if different support agents need different levels of reliability to be considered highly reliable and if trust differs if they have the self-defined high reliability. On the other hand, we included the perspective of the second party (i.e., being evaluated) and were interested in potential effects of different agents on the willingness to be evaluated by them.

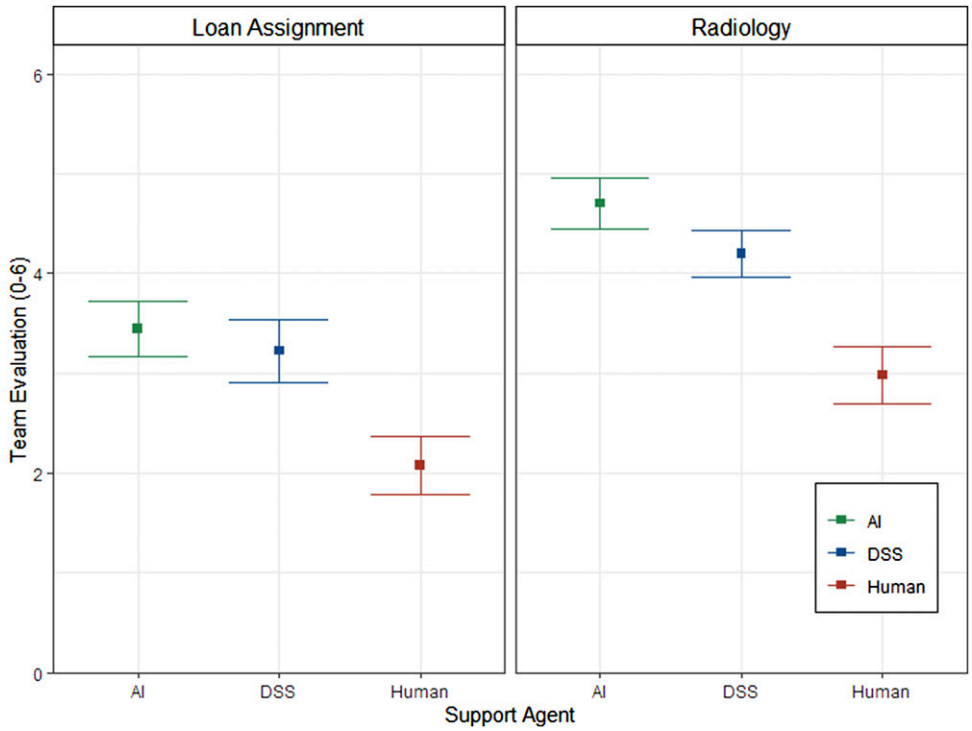

Figure 5. Analysis of team preference.
Note. Means and standard errors for team evaluation (human–human team = 0; human–automation team = 6).

First Party Perspective: Being Supported by an (Artificial) Agent

Surprisingly, we did not find any differences in the required reliability between the three agents. For all agents the rate of correct decisions required to perceive the agent as highly reliable was around 81% for the context of loan assignment and 86% for the context of radiology. This finding stands in contrast to the perfect automation schema (Dzindolet et al., 2002) which would have suggested that demands on reliability would be higher for the two automated systems compared to a human support. Perhaps, there is more of a general concept of high reliability that leads to certain performance expectations regardless of the type of agent. Moreover, technical systems (AI and DSS) were also not required to have near perfect reliability, as would be predicted by earlier research (Dzindolet et al., 2002; Madhavan & Wiegmann, 2007).

As expected, our findings also do not support the perfect automation schema in terms of trust as trust was highest for the human when controlling for required reliability. This holds true even for AI which was expected to induce a higher perfect automation schema than DSS. Interestingly, trust towards AI was higher than towards DSS when controlling for the required reliability. One reason might be the fact that AI shares some commonalities of cognitive abilities previously uniquely associated with humans, namely, learning capability and agency (Heer, 2019; Legaspi et al., 2019). Perhaps, the discrepancy of our results and the previous results of Dzindolet et al. (2002) could be because we described the human agent explicitly as an experienced colleague. This might have led to an attribution of an expertise to the human and therefore a perception of the human being superior compared to the automated aids. In any case, our research suggests that the perceived expertise of the human and automated aids should be considered in future research.

Notably, the findings mentioned above apply to both contexts (i.e., loan assignment and radiology), as the factor context never interacted with the factor agent. Further, in terms of context, we hypothesized that the objective task with more risk (i.e., the radiology task) would result in a higher required reliability and lower trust (Weber et al., 2002). The required reliability was influenced by the task context as expected. In line with our hypotheses, participants wanted the agents to have a higher actual reliability in the radiology context than in the context of loan assignment for the agents to be perceived as highly reliable advice. This can be explained by the tremendous consequences of a mistake that could, in the worst case, lead to death in the context of radiology. In contrast to the required reliability and our hypothesis, we could not find any differences in trust for the two contexts. When the support agent had the self-defined high reliability, there were no differences in trust between the two contexts. However, this might be due to the differences in self-chosen reliability in the first step. Regardless, the results might look different if the reliability is not self-chosen but defined by the system characteristics or technical limitations (Castelo et al., 2019; Dietvorst & Bharti, 2020). This is especially relevant as the reliability of a system is given and not self-chosen in real working environments.

Our assumptions that the more objective classical recognition task would be perceived as more difficult were confirmed. According to findings of Maltz and Shinar (2003), participants depend more on automation when the task is more difficult. They emphasize the importance of assessing the difficulty of the task in order to prevent possible performance degradation due to complacency (Parasuraman, 2000) in easy tasks. Our results directly contradict this assumption as trust was highest in the human agent even for the more difficult task. This suggests that the task difficulty per se might not be an important determinant of how much humans trust the advice of a support agent.

Second Party Perspective: Evaluation by an (Artificial) Agent

The change of perspective from first to second party had a considerable impact on the perception of the different agents. When asked for the required reliability to accept being exclusively evaluated by the agent, participants expressed higher demands for the two automated agents compared to the human. This effect was particularly pronounced for the AI. The higher required reliability of the automated aids could explain the results of Rieger et al. (2022) that showed a preference for an evaluation by the automated agents. This preference could be due to the fact that a higher reliability is expected from the automated agents.

Having in mind that new technological advancements should bring advancements in terms of performance and safety, it makes sense that a novel technology such as AI should be superior to human performance if oneself is evaluated (McKinney et al., 2020). In line with the findings of Bonnefon et al. (2016) and Rieger et al. (2022), our results illustrate that changing the perspective does make a difference and further research needs to consider the second party. The second party's view is especially important as this is a group that is targeted often without giving a consent to the interaction with the AI and can usually not escape this interaction other than quit a job or not apply for certain positions (Langer & Landers, 2021). Moreover, similar to the first party, participants required higher reliability in the context of radiology. This was assumed and can again be explained by the potentially fatal consequences of an incorrectly evaluated x-ray.

Even though these findings on exclusive evaluation are interesting, in the real world, such decisions are often jointly made by two agents (Cymek, 2018; Grießhaber & Mörike, 2021; Mosier & Manzey, 2020). Again, the question arises how people want to be evaluated in this situation. Unsurprisingly, our results suggest that context also matters when deciding between being evaluated by a mixed human–automation team or by a purely human team. Specifically, our data point to a preference for a human–automation team in the context of radiology, consistent with the expectation that this task would fit the strengths of automation (Castelo et al., 2019). In contrast, there was no preference for either a purely human team or human–automation team in the context of loan assignment. Moreover, the support agent which the participants got to know directly from

interaction in their own condition also impacted which kind of team was preferred. That is, participants who interacted with a human had a higher preference for a human–human team than participants in the other two conditions. In contrast, both groups who interacted with technical systems preferred a mixed team over a human–human team. This shows that experiencing and interacting with novel technological advancements, and potentially realizing their benefits, can foster better human–technology interaction and facilitate acceptance and use of new technologies.

Limitations and Future Research

Of course, the present study does not come without limitations. First, participants of the study were not actual professionals working as radiologists or in the finance sector. This is especially important for the first party perspective as earlier research (e.g., Chavaillaz et al., 2019; Navaro et al., 2021) suggests at least some differences between experts and novices when interacting with technologies. However, there is also evidence suggesting that some trust-related phenomena are rather similar between experts and novices (e.g., Mosier et al., 2020). Regardless, the present findings need further investigation and corroboration with experts to further strengthen the practical implications. However, the missing expertise is only a limitation for the first party's perspective. The expertise is not relevant when someone is evaluated by an artificial agent or indirectly influenced in some other way. In contrast, the second party is often not necessarily experienced or even aware of the influence of the technology (Langer & Landers, 2021).

Second, although we recorded actual behavioral dependence on the support agent, we did not analyze it. This approach was chosen, as the presented trials served more or less as an explanation of the task. Assessing behavior was beyond the scope of the current experiment, which aimed to investigate initial differences in required reliability and trust.

Last, all support agents were 100% reliable in the practice trials as we did not want to influence participants through a failure experience. In addition, when asked for trust, participants were presented with support agents that were all *highly* reliable. However, in real life applications, the implementation of automation and AI is justified by their higher reliability compared to humans. Especially, by now AI surpasses even experienced humans, like radiologists, in their performance (Hosny et al., 2018). Therefore, investigating if participants also trust the human more even if they experience the human as less reliable might prove important. In addition, further research should investigate if these trust differences could also result in actual behavioral differences.

CONCLUSION

The most important finding of the present research is that participants required an equally high level of reliability for each of the different agents (i.e., AI, DSS, and human) to consider them as *highly* reliable. Despite this fact, trust towards these agents differed when controlling for the chosen reliability, with higher trust towards human support agents than towards both technical support agents. This decoupling of the direct relationship between reliability and trust offers food for thought, as the kind of support agent seems to play an important role here. Moreover, when changing the perspective, the risk associated with a task context is important, and future research should systematically investigate the impact of risk. Finally, in contrast to human–human teams, human–technology teams offer the opportunity for symbiotic balancing of individual weaknesses—but to unlock this opportunity, it seems necessary to have prior positive experience with any given technology.

ORCID iD

Ksenia Appelganc https://orcid.org/0000-0001-7577-3583
Tobias Rieger https://orcid.org/0000-0002-5097-6613
Eileen Roesler https://orcid.org/0000-0002-7243-4882
Dietrich Manzey https://orcid.org/0000-0002-6977-7024

References

Alfonseca, M., Cebrian, M., Anta, A. F., Coviello, L., Abeliuk, A., & Rahwan, I. (2021). Superintelligence cannot be contained:

Lessons from computability theory. *Journal of Artificial Intelligence Research*, *70*, 65–76. https://doi.org/10.1613/jair.1.12202, https://jair.org/index.php/jair/article/view/12202

Bahrammirzaee, A. (2010). A comparative survey of artificial intelligence applications in finance: artificial neural networks, expert system and hybrid intelligent systems. *Neural Computing and Applications*, *19*(8), 1165–1195. https://doi.org/10.1007/s00521-010-0362-z

Bejnordi, B. E., Veta, M., van Diest, P. J., van Ginneken, B., Karssemeijer, N., Litjens, G., van der Laak, J. A. W. M., Hermsen, M., Manson, Q. F., Balkenhol, M., Geessink, O., Stathonikos, N., van Dijk, M. C., Bult, P., Beca, F., Beck, A. H., Wang, D., Khosla, A., Gargeya, R., & Venâncio, R. (2017). Diagnostic assessment of deep learning algorithms for detection of lymph node metastases in women with breast cancer. *JAMA*, *318*(22), 2199. https://doi.org/10.1001/jama.2017.14585

Bini, S. A. (2018). Artificial intelligence, machine learning, deep learning, and cognitive computing: what do these terms mean and how will they impact health care? *The Journal of Arthroplasty*, *33*(8), 2358–2361. https://doi.org/10.1016/j.arth.2018.02.067

Bonnefon, J. F., Shariff, A., & Rahwan, I. (2016). The social dilemma of autonomous vehicles. *Science*, *352*(6293), 1573–1576. https://doi.org/10.1126/science.aaf2654

Burgess, A. E., Jacobson, F. L., & Judy, P. F. (2001). Human observer detection experiments with mammograms and power-law noise. *Medical Physics*, *28*(4), 419–437. https://doi.org/10.1118/1.1355308

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

Chavaillaz, A., Schwaninger, A., Michel, S., & Sauer, J. (2019). Expertise, automation and trust in X-ray screening of cabin baggage. *Frontiers in Psychology*, *10*, 256. https://doi.org/10.3389/fpsyg.2019.00256, https://www.frontiersin.org/articles/10.3389/fpsyg.2019.00256/full

Cymek, D. H. (2018). Redundant automation monitoring: Four eyes don't see more than two, if everyone turns a blind eye. *Human Factors*, *60*(7), 902–921. https://doi.org/10.1177/0018720818781192

de Leeuw, J. R. (2014). jsPsych: A javaScript library for creating behavioral experiments in a Web browser. *Behavior Research Methods*, *47*(1), 1–12. doi: 10.3758/s13428-014-0458-y

Dietvorst, B. J., & Bharti, S. (2020). People reject algorithms in uncertain decision domains because they have diminishing sensitivity to forecasting error. *Psychological Science*, *31*(10), 1302–1314. https://doi.org/10.1177/0956797620948841

Dzindolet, M. T., Pierce, L. G., Beck, H. P., & Dawe, L. A. (2002). The perceived utility of human and automated aids in a visual detection task. *Human Factors*, *44*(1), 79–94. https://doi.org/10.1518/0018720024494856

Endsley, M. R. (2017). From here to autonomy: Lessons learned from human–automation research. *Human Factors*, *59*(1), 5–27. https://doi.org/10.1177/0018720816681350

Faul, F., Erdfelder, E., Lang, AG, & Buchner, A. (2007). G*Power 3: A flexible statistical power analysis program for the social, behavioral, and biomedical sciences. *Behavior Research Methods*, *39*(2), 175–191. https://doi.org/10.3758/BF03193146

Grießhaber, N. M., & Mörike, F. (2021, September). *Von bedrohung bis kollaborative partnerschaft: Mensch und computer 2021*. https://doi.org/10.1145/3473856.3474585

Hancock, P. A., Billings, D. R., Schaefer, K. E., Chen, J. Y., De Visser, E. J., & Parasuraman, R. (2011). A meta-analysis of factors affecting trust in human-robot interaction. *Human Factors*, *53*(5), 517–527. https://doi.org/10.1177/0018720811417254

Heer, J. (2019). Agency plus automation: Designing artificial intelligence into interactive systems. *Proceedings of the National Academy of Sciences*, *116*(6), 1844–1850. https://doi.org/10.1073/pnas.1807184115

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

Hosny, A., Parmar, C., Quackenbush, J., Schwartz, L. H., & Aerts, H. J. W. L. (2018). Artificial intelligence in radiology. *Nature Reviews Cancer*, *18*(8), 500–510. https://doi.org/10.1038/s41568-018-0016-5

Huegli, D., Merks, S., & Schwaninger, A. (2020). Automation reliability, human–machine system performance, and operator compliance: A study with airport security screeners supported by automated explosives detection systems for cabin baggage screening. *Applied Ergonomics*, *86*, 103094. https://doi.org/10.1016/j.apergo.2020.103094, https://www.sciencedirect.com/science/article/pii/S0003687020300570?via%3Dihub

Jacovi, A., Marasović, A., Miller, T., & Goldberg, Y. (2021, March). Formalizing trust in artificial intelligence: Prerequisites, causes and goals of human trust in AI. In Proceedings of the 2021 ACM conference on fairness, accountability, and transparency, Virtual Event Canada, 3–10 March, 2021, 624–635. https://doi.org/10.1145/3442188.3445923

Janssen, C. P., Donker, S. F., Brumby, D. P., & Kun, A. L. (2019). History and future of human-automation interaction. *International Journal of Human-Computer Studies*, *131*, 99–107. https://doi.org/10.1016/j.ijhcs.2019.05.006, https://www.sciencedirect.com/science/article/pii/S1071581919300552

Lange, K., Kühn, S., & Filevich, E. (2015). Just another tool for online studies" (JATOS): An easy solution for setup and management of web servers supporting online studies. *PLOS ONE*, *10*(6), Article e0130834. https://doi.org/10.1371/journal.pone.0130834

Langer, M., König, C. J., Back, C., & Hemsing, V. (2021). *Trust in Artificial Intelligence: Comparing trust processes between human and automated trustees in light of unfair bias*. https://doi.org/10.31234/osf.io/r9y3t

Langer, M., König, C. J., & Papathanasiou, M. (2019). Highly automated job interviews: Acceptance under the influence of stakes. *International Journal of Selection and Assessment*, *27*(3), 217–234. https://doi.org/10.1111/ijsa.12246

Langer, M., & 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*. https://doi.org/10.1016/j.chb.2021.106878

Lankton, N. K., McKnight, D. H., & Tripp, J. (2015). Technology, humanness, and trust: Rethinking trust in technology. *Journal of the Association for Information Systems*, *16*(10), 1. https://doi.org/10.1038/s41467-019-12170-0

Legaspi, R., He, Z., & Toyoizumi, T. (2019). Synthetic agency: Sense of agency in artificial intelligence. *Current Opinion in Behavioral Sciences*, *29*, 84–90. https://doi.org/10.1016/j.cobeha.2019.04.004, https://www.sciencedirect.com/science/article/pii/S2352154618301700

Lerch, F. J., Prietula, M. J., & Kulik, C. T. (1997). The Turing effect: The nature of trust in expert systems advice. In *Expertise in context: Human and machine* (pp. 417–448). MIT Press.

Lyons, J., Ho, N., Friedman, J., Alarcon, G., & Guznov, S. (2018). Trust of learning systems: Considerations for code, algorithms, and affordances for learning. In *Human and machine learning* (pp. 265–278). Springer. https://doi.org/10.1007/978-3-319-90403-0_13

Madhavan, P., & Wiegmann, D. A. (2007). Similarities and differences between human–human and human–automation trust: An integrative review. *Theoretical Issues in Ergonomics Science*, *8*(4), 277–301. https://doi.org/10.1080/14639220500337708

Maltz, M., & Shinar, D. (2003). New alternative methods of analyzing human behavior in cued target acquisition. *Human Factors*, *45*(2), 281–295. https://doi.org/10.1518/hfes.45.2.281.27239

McKinney, S. M., Sieniek, M., Godbole, V., Godwin, J., Antropova, N., Ashrafian, H., Back, T., Chesus, M., Corrado, G. S., Darzi, A., Etemadi, M., Garcia-Vicente, F., Gilbert, F. J., Halling-Brown, M., Hassabis, D., Jansen, S., Karthikesalingam, A., Kelly, C. J., King, D., & Shetty, S. (2020). International evaluation of an AI system for breast cancer screening. *Nature*, *577*(7788), 89–94. https://doi.org/10.1038/s41586-019-1799-6

Mosier, K., & Manzey, D. (2020). Humans and automated decision aids: A match made in heaven? In M. Mouloua & P. A. Hancock (Eds.), *Human performance in automated and autonomous systems: Current theory and methods* (1st Ed., pp. 19–42). CRC Press.

Navarro, J., Allali, S., Cabrignac, N., & Cegarra, J. (2021). Impact of pilot's expertise on selection, use, trust, and acceptance of automation. *IEEE Transactions on Human-Machine Systems*, *51*(5), 432–441. https://doi.org/10.1109/THMS.2021.3090199

O'Neil, C. (2016). *Weapons of math destruction: How big data increases inequality and threatens democracy*. Crown.

O'Neil, C. (2020). AI, ethics, and the law. In *Work in the future* (pp. 145–153). Springer International Publishing. https://doi.org/10.1007/978-3-030-21134-9_15

Parasuraman, R. (2000). Designing automation for human use: Empirical studies and quantitative models. *Ergonomics*, *43*(7), 931–951. https://doi.org/10.1080/001401300409125

Parasuraman, R., & Riley, V. (1997). Humans and automation: Use, misuse, disuse, abuse. *Human Factors*, *39*(2), 230–253. https://doi.org/10.1518/001872097778543886

Power, D. (2008). Decision support systems: A historical overview. In *Handbook on decision support systems 1. International handbooks information system*. Springer. https://doi.org/10.1007/978-3-540-48713-5_7

Rieger, T., Heilmann, L., & Manzey, D. (2021). Visual search behavior and performance in luggage screening: effects of time pressure, automation aid, and target expectancy. *Cognitive Research: Principles and Implications*, *6*(12), 1–12. https://doi.org/10.1186/s41235-021-00280-7

Rieger, T., & Manzey, D. (2020). Human performance consequences of automated decision aids: The impact of time pressure. *Human factors*, 1–18. https://doi.org/10.1177/0018720820965019

Rieger, T., Roesler, E., & Manzey, D. (2022). Challenging presumed technological superiority when working with (artificial) colleagues. *Scientific Reports*, *12*(1), 3768. https://doi.org/10.1038/s41598-022-07808-x

Stuck, R. E., Tomlinson, B. J., & Walker, B. N. (2021). The importance of incorporating risk into human-automation trust. *Theoretical Issues in Ergonomics Science*, 1–17. https://doi.org/10.1080/1463922X.2021.1975170

Weber, E. U., Blais, A.-R., & Betz, N. E. (2002). A domain-specific risk-attitude scale: measuring risk perceptions and risk behaviors. *Journal of Behavioral Decision Making*, *15*(4), 263–290. https://doi.org/10.1002/bdm.414

Ksenia Appelganc is a researcher and lecturer at the Department of Business Psychology and Human Resource Management, University of Bremen. She earned a master's in human factors at the Technische Universität Berlin and is currently working on a PhD addressing the issues of team performance in human agent teams in a Mars habitat.

Tobias Rieger is a researcher and lecturer at the Department of Psychology and Ergonomics, Technische Universität Berlin, Germany. He earned a master's in psychology at the University of Freiburg in 2018 and is currently working on a PhD addressing issues of human performance consequences of automation.

Eileen Roesler is a researcher and lecturer at the Department of Psychology and Ergonomics, Technische Universität Berlin, Germany. She earned a master's in psychology focusing on human performance in sociotechnical systems at the Technische Universität Dresden in 2018 and is currently working on her PhD addressing the effects of anthropomorphism in human–robot interaction.

Dietrich Manzey is a university professor of work, engineering and organizational psychology in the Department of Psychology and Ergonomics, Technische Universität Berlin, Germany. He earned his PhD in experimental psychology at the University of Kiel, Germany, in 1988 and his habilitation in psychology at the University of Marburg, Germany, in 1999.