Kares, Felix; König, Cornelius J.; Bergs, Richard; Protzel, Clea; Langer, Markus

Trust in hybrid human-automated decision-support

International Journal of Selection and Assessment

Provided in Cooperation with:
John Wiley & Sons

Suggested Citation: Kares, Felix; König, Cornelius J.; Bergs, Richard; Protzel, Clea; Langer, Markus (2023) : Trust in hybrid human-automated decision-support, International Journal of Selection and Assessment, ISSN 1468-2389, Wiley, Hoboken, NJ, Vol. 31, Iss. 3, pp. 388-402, https://doi.org/10.1111/ijsa.12423

Received: 30 April 2022 | Revised: 10 February 2023 | Accepted: 11 February 2023

DOI: 10.1111/ijsa.12423

RESEARCH ARTICLE

International Journal of Selection and Assessment | WILEY

Trust in hybrid human-automated decision-support

Felix Kares[1] | Cornelius J. König[1] | Richard Bergs[1] | Clea Protzel[1] | Markus Langer[2]

[1]Fachrichtung Psychologie, Universität des Saarlandes, Saarbrücken, Germany

[2]Fachbereich Psychologie, Philipps-Universität Marburg, Marburg, Germany

Correspondence
Felix Kares, Fachrichtung Psychologie, Universität des Saarlandes, Arbeits & Organisationspsychologie, Campus A1 3, 66123 Saarbrücken, Germany.
Email: felix.kares@uni-saarland.de

Funding information
Volkswagen Foundation, Grant/Award Number: AZ98513; DFG German Research Foundation, Grant/Award Number: DFGgrant389792660aspartofTRR248

Abstract

Research has examined trust in humans and trust in automated decision support. Although reflecting a likely realization of decision support in high-risk tasks such as personnel selection, trust in hybrid human-automation teams has thus far received limited attention. In two experiments ($N_1 = 170$, $N_2 = 154$) we compare trust, trustworthiness, and trusting behavior for different types of decision-support (automated, human, hybrid) across two assessment contexts (personnel selection, bonus payments). We additionally examined a possible trust violation by presenting one group of participants a preselection that included predominantly male candidates, thus reflecting possible unfair bias. Whereas fully-automated decisions were trusted less, results suggest that trust in hybrid decision support was similar to trust in human-only support. Trust violations were not perceived differently based on the type of support. We discuss theoretical (e.g., trust in hybrid support) and practical implications (e.g., keeping humans in the loop to prevent negative reactions).

KEYWORDS
artificial intelligence, decision-support, human-automation collaboration, personnel selection, trust

Practitioner points

(a) What is currently known about the topic of our study:
- Automated decision-support (DS) often fueled by artificial intelligence can be perceived more negatively in selection tasks than human DS
- The task context can modulate trust in automated DS
- Depending on the agent (human or system), trust violations can be perceived differently

(b) What our Paper adds to this:
- In both examined contexts (selection for bonus payments; personnel selection), system DS was trusted less compared to human DS
- Fairness issues in a decision negatively impacted trust but did not lead to different reactions based on the type of agent that produced them

- Hybrid DS is perceived on par with human DS and better than system DS when it comes to trust

(c) The implications of our study findings for practitioners:

- Full automation of HR related tasks can be perceived negatively, even if the task may seem suited for automation
- Avoid fairness issues in selection as they negatively impact trust regardless of the agent that produced them
- Adding a human to an automated DS (i.e., realizing hybrid DS) can alleviate negative effects on trust associated with full automation.

1 | INTRODUCTION

Trust is crucial in situations where decision-makers receive support, no matter whether this support comes from human colleagues or from algorithm-based, automated systems. Although there is extensive research on trust in, both, human and automated support (Baer & Colquitt, 2018; J. D. Lee & See, 2004), little is known about trust in an automated system and a human decision-maker working together. We propose that hybrid human-system decision support (DS) may be perceived as a trust agent combined of two single trust agents, and humans may need to trust the human, the system, as well as the combined work of these agents.

Initial evidence hints that a cooperation of this nature can be perceived more but also less favorably compared to single agents, depending on stakeholders (e.g., familiarity with technology; Gonzalez et al., 2022) or task characteristics (e.g., complexity; Nagtegaal, 2021). However, research on reactions to hybrid DS is only emerging and especially in the area of trust, there is to the best of our knowledge no research that explicitly examines reactions to hybrid DS. This is unfortunate because hybrid collaborations reflect a likely implementation of automated systems in future high-risk decision-making processes such as personnel selection, given that ethical guidelines (e.g., the European Commission's Ethics Guidelines for Trustworthy AI, 2019) and proposed legislation (European Commission, AI Act, 2021) call for keeping humans in the loop.

The goal of this paper is to shed light on trust in hybrid DS by examining different selection contexts, namely personnel selection (Study 1) and bonus payment (Study 2). Participants received a preselection of candidates that was produced either by a human, an automated, or a hybrid DS and the preselection either displayed an equal number of male and female candidates or predominantly male candidates, simulating a possible trust violation. We chose these contexts because we expected that people react differently in those contexts to a human versus a system providing decision-support (M. K. Lee, 2018).

2 | BACKGROUND AND HYPOTHESES DEVELOPMENT

2.1 | Interpersonal trust and trust in automation

Trust is essential for everyday work processes (Mayer et al., 1995). It increases work efficiency and enables predictability of situations where supervision or support by the trustor (the agent who trusts) are necessary (Lee & See, 2004). This applies to both trust in humans and trust in automated systems. Mayer et al. (1995, p. 712) define trust as "the willingness of a party to be vulnerable to the actions of another party based on the expectation that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party."

The trust process starts with an evaluation of the trustee's (the agent who is trusted) trustworthiness given prior information toward the trustee. According to Mayer et al. (1995), trustworthiness consists of the facets ability, integrity, and benevolence. Ability concerns the capacities of the trustee to fulfill a given task, high integrity means that a trustee adheres to a set of principles that the trustor finds acceptable, and benevolence is the extent to which a trustee is believed to want to do good to the trustor aside from an egocentric profit motive (Mayer et al., 1995). Trustworthiness of automation is also conceptualized with multiple facets (partly) similar to the facets for human trustees. For instance, conceptualizations of trustworthiness of automated systems always include a facet that concerns the perceived capabilities of an automated system, corresponding to the ability facet (Körber et al., 2018; J. D. Lee & See, 2004; Mcknight et al., 2011). Further proposed facets of trustworthiness for automated systems align with integrity as they highlight predictability, transparency, or alignment with moral standards and ethical values (e.g., lack of bias) of automated systems (Höddinghaus et al., 2021; Körber et al., 2018; J. D. Lee & See, 2004). Thus, we decided to include a measure of unbiasedness as well as transparency to reflect perceived integrity. Unbiasedness is especially important when it comes to the trustworthiness of automation contexts where systems decide about the fate of humans

(e.g., Langer, König, et al., 2022; Zerilli et al., 2022) and transparency has received attention as a facet of trustworthiness across contexts (e.g., Glikson & Woolley, 2020; Höddinghaus et al., 2021). Other facets align with benevolence arguing that trustors assess the purpose for which systems are developed as well as developers' intentions that manifest in system functioning (Höddinghaus et al., 2021; J. D. Lee & See, 2004).

It is assumed that perceived trustworthiness then determines the initial level of trust toward a trustee (Mayer et al., 1995). If trust is high enough, the trustor may (under consideration of further variables such as risks involved, trustors' propensity to trust) engage in trusting behavior and will afterwards evaluate the outcomes of the actions of the trustee. Based on this, trustworthiness will be reevaluated (Mayer et al., 1995). In the case of positive outcomes, trustworthiness might increase, in the case of trust violations it may decrease.

2.2 | Human and automated decision-support

There is increasing interest in the use of automated decision-support in personnel selection (Langer & Landers, 2021). Screening large quantities of information is a strength of automated DS systems that can make selection processes more time-efficient (e.g., Campion et al., 2016). Additionally, automated DS systems are realizations of mechanical combination of data and could thus increase the validity of job performance predictions compared to holistic data combination methods (Kuncel et al., 2013), although scientific evidence regarding validity is just emerging (Hickman et al., 2022). Yet, although there is potential for high perceived trustworthiness for automated decision-support, people might not necessarily perceive automated systems as capable of high-quality personnel selection (Langer et al., 2021).

Automated DS in personnel selection also involves possible adverse effects. Even though automation removes individual human decision-makers' biases from decision processes, the outputs of automated systems can also reflect unfair bias. In fact, it seems challenging to control for every possible way that bias and subgroup differences can be reflected in automated systems decision processes (Tay et al., 2022). Nevertheless, people might still expect systems to be less biased and more consistent than human decision-makers (Langer & Landers, 2021). Such expectations toward automated systems could affect the trustworthiness facet integrity. However, research has shown that people are inclined to rather trust human than automated decision-makers, especially in selection contexts (Langer, König, 2022; M. K. Lee, 2018). This might be because humans are perceived to be able to evaluate the characteristics of humans holistically, while systems are perceived as reductionistic (Newman et al., 2020).

Yet, such reactions may differ depending on the type of task that a system is designed for. Applicant selection may be perceived to be highly complex and to require more "human" skills (M. K. Lee, 2018; Newman et al., 2020). These task characteristics have been proposed to lead to less favorable attitudes toward systems for such tasks (M. K. Lee, 2018; Nagtegaal, 2021). However, there are also selection tasks that may be perceived as better suited to be conducted by automated systems; for example, the selection of employees for bonus payments. Such a selection could be conducted according to few clearly-defined rules because the characteristics of employees within the same company may be more easily comparable than those of applicants. For example, for the evaluation of applicants, even comparable grades could mean different things depending on the applicants' school or year of graduation. Furthermore, selecting employees for bonus payments may be seen as a task that requires mechanical skills. Performance criteria may be more quantitative in nature and people may expect that the selection for bonus payments should follow strict rules (Nagtegaal, 2021), whereas such strict rules may be perceived too reductionistic for the assessment of applicants (Newman et al., 2020. For more clearly-defined tasks that may also be perceived as more monotonous, humans could be perceived as more prone to errors than systems and the consistency that people ascribe to systems may be perceived as beneficial (Madhavan & Wiegmann, 2007).

To examine whether a different context indeed influences trust-related variables, we employ a personnel selection setting for Study 1 and change the context for Study 2 to a selection between employees for bonus payment.[1] Overall, we propose the following hypotheses that aim to replicate prior findings for personnel selection and go beyond them by examining differences based on the selection context:

Hypothesis 1.1. Trust-related variables (trustworthiness, trust, trusting behavior) are impacted positively in the personnel selection task if the recommendation comes from a human compared to an automated system.

Hypothesis 1.2. Trust-related variables are impacted negatively in the bonus payment task if the recommendation comes from a human compared to an automated system.

3 | HYBRID DECISION-SUPPORT

We propose that trust research can benefit from examining trust regarding hybrid DS. For instance, supervisors trust that production processes will be fulfilled efficiently by their employee-robot teams (Sheridan & Parasuraman, 2005) or trust that employees interact effectively with automated DS in managerial decision-making (Langer et al., 2021). Decisions of hybrid human-automation DS affect an increasing number of people making it necessary to investigate trust processes in relation to hybrid DS.

Cooperation between humans and automation promises benefits but not all of those may manifest in practice. Individual weaknesses could be mitigated and work outcomes can reflect the best of both

worlds (Jarrahi, 2018; Mosier & Manzey, 2019). For example, faulty (e.g., biased) system outputs could be detected by the human who can then adjust decisions. Similarly, systems could prevent the premature selection of candidates without considering all relevant information (Koivunen et al., 2019) and could improve decision quality (Kuncel et al., 2013). Nevertheless, hopes that are put in human-system collaborations may also be misguided. There is evidence that human-system collaboration can decrease decision quality in comparison to human-only decisions (Skitka et al., 1999) and in comparison to automation-only decisions (Schemmer et al., 2022). Reasons for this are that humans follow/reject system decisions in the wrong situations and in other situations system outputs may receive too much/too little weight (Green, 2022; Parasuraman & Riley, 1997).

On the one hand, even when trust in one agent is low, trust in the hybrid DS may remain high if perceptions toward the other agent buffers this lack of trust. This could indicate that trust in hybrid DS would at least be similar to trust in the single agent that trustors perceive to be more trustworthy. There is initial evidence that hybrid decision-making is perceived more favorably than fully automated and on-par with human decision-making when it comes to perceptions of the legitimacy of political decisions (Starke & Lünich, 2020) or procedural justice in personnel selection (Gonzalez et al., 2022). In contrast, hybrid DS could also lead people to perceive that either the human or the system in the hybrid team may deteriorate decision quality compared to single-agent decisions. In line with this, there is initial evidence showing that hybrid DS can also be perceived as less positive regarding procedural justice, a construct that is closely related to trust (Colquitt & Rodell, 2011), than single-agent decision-support (Nagtegaal, 2021). Given this discussion about possible benefits and disadvantages, we thus ask:

Research Question (RQ) 1: Is there a difference in trust-related variables in a hybrid human-automation DS compared to human or automated support?

We furthermore explore whether trustors perceive single agents that are part of the hybrid DS differently than single agents that act independently. Humans could be perceived as less qualified if they require automated support to perform their work (Arkes et al., 2007). Similarly, systems could be perceived as less useful if humans need to be involved in decision-making. In contrast, we could also imagine that people perceive single agents in a hybrid DS as more trustworthy because they are able to collaborate efficiently. Effective interaction with a system could be perceived as requiring specific skills and could increase perceptions of ability. As an exploratory part of this work we thus ask:

Exploratory Question (EQ)1: How do trustworthiness and trust perceptions toward the human component as a part of the hybrid DS differ from the human-only support?

EQ. 2: How do trustworthiness and trust perceptions toward the automated component as a part of the hybrid DS differ from the automated-only support?

3.1 | Trust violations

Whereas decision-support can benefit decision quality and efficiency, neither human nor automated decisions are perfect. Consequently, it is important to examine how trust violations affect trust for the different types of DS. Fairness issues can be observed for humans and for automated systems (Langer, König, et al., 2022) and can be perceived differently when coming from a human or an automated system (e.g., Bigman et al., 2022). Therefore, we decided to investigate possible fairness issues as a trust violation. Specifically, we employed an experimental manipulation where participants were either confronted with a predominantly male preselection (which may reflect biased decision-making) or an output with similar numbers of male and female applicants. For an unbalanced preselection, participants may suspect that the candidates' gender has been influential. Since disparate treatment as a consequence of characteristics that are irrelevant for selection may be perceived as a violation of social norms or even law (Cook, 2016), the unbalanced condition should negatively impact the trustworthiness facets of integrity and benevolence, resulting in more negative attitudes. We thus propose:

Hypothesis 2. The unbalanced preselection negatively affects trust-related variables compared to the unbalanced preselection.

Research has shown that trust violations may lead to different effects for humans versus for automated systems (de Visser et al., 2018) and this may depend on the task at hand (Langer, König, et al., 2022). For instance, people seem to expected consistency from automated systems (Dietvorst & Bharti, 2020; Parasuraman & Manzey, 2010) and for tasks that require consistency (e.g., monitoring tasks), trust violations seem to have a stronger impact on trust in systems than in humans (Dzindolet et al., 2003). In contrast, people are less likely to expect systems to apply social norms and ethical considerations (e.g., Bigman et al., 2022), so for tasks that may require a greater understanding of social norms and ethics (such as personnel selection; Langer, König, et al., 2022; Rieger et al., 2022) trust violations may have weaker effects. For example, if people believe that systems do not actively discriminate, they might react more strongly to trust violations that reflects unethical behavior if produced by a human trustee (Bigman et al., 2022). In the case of human support, unfair discrimination of applicants could be considered a breach of integrity (Villegas et al., 2019), benevolence, and possibly also competence, reducing trustworthiness perceptions. On the other hand, a trust violation of automated DS may not be detected or may elicit weaker reactions (Bigman et al., 2022; Langer et al., 2021). Therefore, we propose:

Hypothesis 3. An unbalanced recommendation provided by human support leads to more negative effects for trust-related variables than an unbalanced recommendation provided by an automated system.

It remains unclear how these differences between automated and human DS regarding trust violations associated with fairness

issues affect trustworthiness evaluations of hybrid human-automation DS. Trustors may perceive that the collaboration should decrease the likelihood of errors which could lead to stronger reactions if an error does occur. If neither the human nor the system detected the error, the biased preselection by a hybrid DS may be perceived as the worst of both worlds, combining human inconsistency (Dietvorst & Bharti, 2020) with a low understanding of social norms resulting from delegating part of the task to a system (Bigman & Gray, 2018). This may result in more negative attitudes toward errors made by a hybrid system compared to those made by a single-agent DS. In contrast, if trustors believe that humans and systems buffer each other's weaknesses, trustors may be less likely to interpret trust violations based on possible unfair bias as actual violations. Participants may believe the DS to work without bias when two agents are involved, possibly attributing the gender imbalance to differences in the sample, rather than suspecting that gender affected the selection outcome. To explore the impact of trust violations for hybrid DS, we thus ask:

RQ2: Is there a difference in trust-related variables between the gender-balanced and the predominant male preselection if the preselection is provided by a hybrid human-automation system, a human or a system?

4 | STUDY 1—PERSONNEL SELECTION CONTEXT

4.1 | Method

4.1.1 | Sample

We conducted an a priori power analysis for a two-way analysis of variance with G*Power[2] (Faul et al., 2007) that indicated that a sample of 191 participants was needed to detect a small to medium-effect size of $\eta^2_p = 0.04$, with a power of $1-\beta = 0.80$. We recruited 283 participants from different social media platforms and forums that promote research studies. Psychology students received course credit for participating; other participants were informed that we will send them the study results if they are interested. Of the participants, 104 did not respond to all relevant items. Of the remaining 179, we excluded six participants whose response pattern indicated that they walked away from the study (i.e., responses took longer than an hour) or who responded in less than 3 min (indicating inattentiveness) as well as three participants who informed us that their data should not be used. The final sample consisted of 170 participants (68.2% female; 64.1% psychology students; $M_{age} = 25.81$, $SD_{age} = 11.36$; 48.8% employed). Roughly one-third (38.8%) of participants reported experience with personnel selection, 16.7% of those specified that it was in the role of the employer, 51.5% in the role of applicants, and 31.8% in both roles.

4.1.2 | Procedure

In a 3 (decision-support: automated vs. human vs. hybrid) × 2 (preselection: predominantly male vs. gender-balanced) between-subjects online experiment, participants were randomly assigned to one of the six groups. Participants were instructed to imagine that they are responsible for managing human resources (HR) in a fictional company. They then received a description of the job for which they should make hiring decisions. The job posting included a list of job requirements and descriptions regarding the job duties (Supporting Information: Material A). Participants then received the information that the most promising out of 103 applicants (52 male and 51 female) should be invited to the next selection stage. To reduce the burden for the task, we told participants that they would receive decision-support that preselects 12 applicants. Every preselected applicant was presented with a photo and an applicant profile. This profile contained two qualitative criteria (i.e., information about their strengths and weaknesses) and two numeric criteria (i.e., work reference performance and years of job experience). Specific values were assigned randomly in a way that there were no systematic differences between applicants—every applicant met the job requirements. The photos showed Caucasian individuals of similar age (to simplify study materials, we intentionally did not include older applicants or applicants with a different ethnical background; see Supporting Information: Material B for photos and applicant information). The preselection consisted of either six men and six women (balanced preselection) or of 10 men and 10 women (unbalanced preselection).

The DS was described as a "human resources employee," an "automated system," or a "cooperation between a human resources employee and an automated system." Participants first either accepted or rejected the preselection and were then asked to provide reasons for their decision. Afterwards, participants rated the fairness of the preselection, as well as the trustworthiness of and trust in the respective DS. In the hybrid DS condition, participants rated trustworthiness and trust for the hybrid DS, the human part of the hybrid DS, and the automated part of the hybrid DS.[3] We randomized the order of the trustworthiness and trust items in the hybrid DS condition. Participants then responded to items regarding their affinity for technology, and we gathered demographic information. Finally, participants were asked about experience in personnel selection and about whether they were in the role of the hiring manager or an applicant.

4.1.3 | Measures[4]

We adapted all items to relate to either a human HR employee, a system, or an HR employee cooperating with a system depending on the experimental condition (in the following sample items, we use "the trustee" as a placeholder). Unless stated otherwise, participants rated all items on a five-point scale ranging from 1 (strongly disagree) to 5 (strongly agree).

Trustworthiness

Trustworthiness was measured with four subscales capturing perceived ability, integrity, benevolence, and transparency. The scales for benevolence and integrity were originally used by Benbasat and Wang (2005) and were adapted and translated to German by Höddinghaus et al. (2021). They consisted of three items each, for example "I believe 'the trustee' would put my interest first" for benevolence and "I believe 'the trustee' provides an unbiased preselection" for integrity. The subscales for ability and transparency were developed by Höddinghaus et al. (2021). The subscale ability consists of six items. A sample item is "I believe 'the trustee' can process all the necessary data required for the preselection." The subscale transparency consists of three items. A sample item is "I think I could understand the decision-making processes of 'the trustee'."

Trust

To assess the trust of participants toward the DS, we used three items by Thielsch et al. (2018) that we adapted to capture trust toward different types of DS. A sample item is "I would completely trust 'the trustee'."

Trusting behavior

Trusting behavior was measured with the acceptance or rejection of the preselection. Responding with "Yes, I accept this preselection" indicates trusting behavior.

Additional variables

Participants completed the German version of the Affinity for Technology Interaction Scale with nine items (ATI; Franke et al., 2019) on a six-point scale from "not true at all" to "completely true." A sample item is, "I like to occupy myself in greater detail with technical systems."

4.1.4 | Manipulation check

To check whether participants perceived the preselection outputs (gender-balanced vs. predominantly male) differently, participants rated the perceived fairness of the preselection with the question "Do you think the preselection is fair?" on a five-point scale from "not fair at all" to "completely fair." This item was constructed in reference to Warszta (2012). We additionally asked, whether participants believed that there were more men, more women, or an equal number of men and women in the preselection.

4.2 | Results

Table 1 displays correlations, means and SDs of all variables in Study 1. Table 2 shows contingency tables for participants' rejection and acceptance of the preselection of applicants. We used linear regression and *t*-tests to analyze the data. For the regressions, the condition DS (automated vs. human vs. hybrid DS) was dummy coded so that the human DS was the reference group for the comparison with hybrid and automated DS.[5]

4.2.1 | Research questions and hypotheses

Manipulation check. The unbalanced preselection was perceived as less fair than the balanced one, $R^2 = 0.09$, $F(1, 168) = 16.34$, $p < .001$. This indicates that participants were aware of the difference in the gender distribution and perceived the condition with a disproportionate number of male applicants as unfair compared to the balanced preselection. Additionally, 91.44% of participants in the unbalanced groups correctly reported that there was an unequal number of male and female applicants.

Trust, trustworthiness and trusting behavior

H1.1 stated that trust-related variables will be positively impacted for the personnel selection task if the recommendation comes from a human compared to an automated system. Overall trustworthiness did not significantly differ between the conditions (see Table 3). Thus, we examined the trustworthiness facets (see Table 4) and found that participants perceived humans as more able than the system. Furthermore, the results indicated that participants reported more trust in the human than in the automated DS. Regarding trusting behavior, 35.6% of participants in the human and 38.7% in the automated group rejected the preselection (see Table 2), indicating no significant difference. Our results thus only partly support H1.1: although there were differences in a specific trustworthiness facet and for trust, overall trustworthiness and trusting behavior showed no significant differences between human and automated DS.

RQ1 asked whether there is a difference in trust-related variables in a hybrid human-automation system compared to human or automated support. We found no significant differences and very small effect sizes between hybrid and human DS for trust, trustworthiness and all its facets (Tables 3 and 4). For the comparison of the automated and the hybrid DS group, we found significant differences indicating higher values for the hybrid group for both trust, $t(109) = 3.83$, $p < .001$, $d = 0.73$, and trustworthiness, $t(109) = 2.45$, $p = .02$, $d = 0.46$. With a rejection rate of 36.7%, the hybrid group showed no difference to the automated or human DS. Hybrid DS was thus perceived more positively than the automated DS and on-par with the human support.

EQ. 1 and EQ. 2 asked whether there are differences in trustworthiness and trust between the human or system part in the hybrid DS relative to the human- or system-only groups respectively. The human-only versus human as a part of hybrid DS comparison yielded no significant differences for trust, $t(106) = 0.65$, $p = .52$, $d = 0.13$, or trustworthiness, $t(106) = 0.93$, $p = .35$, $d = 0.18$. For the automated DS, there were also no significant differences for trust, $t(109) = 0.13$, $p = .89$, $d = 0.02$, and trustworthiness, $t(109) = 0.42$, $p = .67$, $d = 0.08$. In summary, for the comparison of human-only DS with the human part of hybrid support, trustworthiness and trust

TABLE 1 Means, standard deviations, and correlations of Study 1 and Study 2.

Variable	*M*	*SD*	1	2	3	4	5	6	7	8	9	10	11
1. System DS	0.36	0.48											
	[0.36]	[0.48]											
2. Hybrid DS	0.29	0.45	−.48**										
	[0.32]	[0.47]	[−.52**]										
3. Preselection balance	0.45	0.50	−.09	.05									
	[0.53]	[0.50]	[−.05]	[.02]									
4. Gender	0.68	0.47	.07	.04	−.10								
	[0.74]	[0.44]	[.01]	[−.06]	[−.12]								
5. Age	25.81	11.36	.05	−.10	.12	−.15							
	[24.94]	[7.49]	[.12]	[−.07]	[.02]	[−.08]							
6. Psychology student	0.64	0.48	−.02	.02	−.14	.25**	−.48**						
	[0.69]	[0.46]	[−.01]	[−.05]	[−.06]	[.03]	[−.25**]						
7. Employment	0.49	0.50	.02	−.05	−.05	−.17*	.11	−.30**					
	[0.56]	[0.50]	[.05]	[.02]	[.14]	[−.07]	[.29**]	[−.30**]					
8. Experience	2.73	1.26	−.08	.02	.16*	−.08	.25**	−.19*	−.06				
	[2.03]	[1.13]	[−.02]	[.11]	[.09]	[−.18**]	[.22**]	[−.26**]	[.18*]				
9. Trustworthiness	3.10	0.55	−.21**	.13	−.16*	−.03	.10	−.03	.17*	−.03			
	[3.29]	[0.68]	[−.19*]	[.17*]	[−.24**]	[−.03]	[.07]	[−.09]	[−.09]	[.17*]			
10. Trust	2.49	0.95	−.32**	.16*	−.17*	.05	.09	−.04	.06	.08	.71**		
	[2.95]	[0.97]	[−.27**]	[.21**]	[−.22**]	[−.11]	[.09]	[−.13]	[−.13]	[.12]	[.76**]		
11. Fairness	3.20	0.96	.08	−.19*	−.30**	−.00	.11	.09	.09	−.06	.48**	.42**	
	[3.32]	[1.05]	[.01]	[−.05]	[−.40**]	[.07]	[−.05]	[−.08]	[−.11]	[.07]	[.59**]	[.50**]	
12. ATI	3.24	1.08	−.06	−.00	.01	−.46**	.10	−.27**	.16*	.31**	−.06	−.08	−.09
	[3.22]	[1.00]	[−.01]	[.09]	[.20*]	[−.34**]	[.17*]	[−.29**]	[.05]	[.13]	[.10]	[.09]	[−.02]

Note: Study 2 values in brackets. ATI, affinity for technology; DS, decision support. DS type was dummy-coded: 0 = human DS, 1 = system or hybrid DS. Coding of preselection balance: 0 = gender-balanced, 1 = predominantly-male. Coding of Gender: 0 = male, 1 = female. Coding of psychology student: 0 = no, 1 = yes. Coding of whether participants were employed: 0 = no, 1 = yes. Experience codes experience with personnel selection (Study 1) or bonus payments (Study 2). N_{Study1} = 170, N_{Study2} = 154.

*p < .05; **p < .01.

TABLE 2 Percentage of participants who accepted and rejected the preselection in each condition and in Study 1 and Study 2.

	H			A			H-A		
	B	MD	T	B	MD	T	B	MD	T
Study 1									
Reject	15.2	20.3	35.5	22.6	16.1	38.7	12.2	24.5	36.7
Accept	37.3	27.2	64.5	38.7	22.6	61.3	38.8	24.5	63.3
Total	52.5	47.5	100.0	61.3	38.7	100.0	51.0	49.0	100.0
Study 2									
Reject	6.1	20.4	26.5	5.5	16.4	21.9	6.0	24.0	30.0
Accept	38.8	34.7	73.5	45.5	32.6	78.1	40.0	30.0	70.0
Total	44.9	55.1	100.0	51.0	49.0	100.0	46.0	54.0	100.0

Note: A, automated DS; B, gender-balanced preselection; H, human decision-support (DS); H-A, hybrid human-automated DS; MD, male dominant; N_{Study1} = 170; N_{Study2} = 154; T, total.

TABLE 3 Regression results for trust and overall trustworthiness in both Studies.

	Study 1						Study 2					
	Trust			Trustworthiness			Trust			Trustworthiness		
Predictors	Estimates	CI	*p*	Estimates	CI	*p*	Estimates	CI	*p*	Estimates	CI	*p*
(Intercept)	2.92	2.61, 3.24	<.001	3.25	3.06, 3.44	<.001	3.35	2.96, 3.73	<.001	3.56	3.29, 3.84	<.001
H versus A	−0.72	−1.15, −0.30	<.001	−0.20	−0.45, 0.06	.132	−0.55	−1.07, −0.04	.036	−0.28	−0.65, 0.09	.134
H versus H-A	0.01	−0.46, 0.48	.971	0.02	−0.27, 0.30	.901	0.10	−0.44, 0.64	.712	0.01	−0.37, 0.39	.961
Preselection	−0.46	−0.92, −0.00	.049	−0.18	−0.46, 0.10	.202	−0.55	−1.07, −0.03	.039	−0.46	−0.83, −0.09	.014
H versus A × preselection	0.20	−0.44, 0.85	.537	−0.10	−0.49, 0.29	.625	0.13	−0.58, 0.84	.717	0.13	−0.38, 0.63	.624
H versus H-A × preselection	0.05	−0.62, 0.73	.873	0.05	−0.36, 0.46	.807	0.16	−0.57, 0.89	.669	0.23	−0.28, 0.75	.374

Note: A, automated DS; H, human decision-support (DS); H-A, hybrid DS; DS type was dummy-coded: 0 = human DS, 1 = system or hybrid DS. Coding of preselection balance: 0 = gender-balanced, 1 = predominantly-male. $N_{Study1} = 170$, $N_{Study2} = 154$.

TABLE 4 Regression results for trustworthiness facets of Study 1.

	Ability			Benevolence			Unbiased			Transparency		
Predictors	Estimates	CI	*p*	Estimates	CI	*p*	Estimates	CI	*p*	Estimates	CI	*p*
(Intercept)	3.41	3.17, 3.65	<.001	3.15	2.86, 3.44	<.001	2.79	2.45, 3.13	<.001	3.35	3.13, 3.58	<.001
H versus A	−0.52	−0.85, −0.20	<.001	−0.20	−0.59, 0.18	.302	0.45	−0.01, 0.90	.054	0.03	−0.28, 0.34	.843
H versus H-A	−0.13	−0.49, 0.23	.483	−0.00	−0.43, 0.43	.986	0.21	−0.29, 0.71	.412	0.21	−0.14, 0.55	.241
Preselection	−0.31	−0.66, 0.04	.079	−0.35	−0.77, 0.06	.097	−0.00	−0.49, 0.48	.985	0.15	−0.19, 0.48	.392
H versus A × preselection	0.06	−0.44, 0.55	.813	0.02	−0.57, 0.61	.956	−0.36	−1.05, 0.33	.309	−0.35	−0.82, 0.12	.144
H versus H-A × preselection	0.31	−0.21, 0.83	.239	0.29	−0.33, 0.91	.357	−0.39	−1.12, 0.33	.288	−0.41	−0.91, 0.08	.101

Note: A, automated DS; H, human decision-support (DS); H-A, hybrid DS; DS type was dummy-coded: 0 = human DS, 1 = system or hybrid DS. Coding of preselection balance: 0 = gender-balanced, 1 = predominantly-male. $N = 170$.

were perceived similarly. The same is true for the analogous comparison for automated DS.

Trust violations

H2 proposed that the balanced preselection will lead to more positive effects for trust-related variables than the unbalanced preselection condition. Table 3 shows that there was no difference in overall trustworthiness perceptions but that participants trusted the DS less when the preselection was predominantly male. When it comes to trusting behavior, 51.5% of the predominantly male and 29.0% of the gender-balanced preselections were rejected. In a logistic regression, this difference failed to reach significance[6] (see Table 6). H2.1 was thus only partially supported.

H3 suggested that an unbalanced recommendation provided by human support will lead to more negative effects for trust-related variables than an unbalanced recommendation provided by an automated system. We found no significant interaction effects for any of the dependent variables (see Table 3 for trust and trustworthiness, Table 6 for trusting behavior). H3 was thus not supported.

RQ2 asked whether there are differences for trust-related variables between the gender-balanced and the predominant male preselection if the preselection is provided by a hybrid human-automation system, a human, or a system. The corresponding analyses showed no interaction effects (see Table 3 for trust and trustworthiness, Table 6 for trusting behavior).

4.3 | Discussion Study 1

Results showed that the human DS was trusted more than automated support (see also Langer, König, 2022). For trustworthiness, however, there was only a difference for a single facet: The human DS was perceived as more capable. Results further showed that hybrid DS was perceived on-par with human DS when it comes to trust-related variables and that both human and hybrid support were rated more favorably than automated DS. One interpretation of this finding could be that humans were perceived similarly regardless of being described as working alone or together with an automated system, because participants may have perceived the automated system as a tool rather than an integral part of a decision-making process. Finding no significant differences and only small effect sizes for the comparison of the human-only DS and the human part of the hybrid DS are in line with this.

The trust violation did not influence trustworthiness or trusting behavior. Whereas we could conclude that the unbalanced preselection fulfilled its role as a trust violation because it led to lower levels of trust, and descriptively leaned toward lower trusting behavior, we were surprised that trustworthiness or its facets were not affected by the unbalanced preselection. Also, trust violations were not perceived differently depending on the type of DS. In the overall discussion, we will further discuss this finding.

5 | STUDY 2—BONUS PAYMENT CONTEXT

5.1 | Method

5.1.1 | Sample

Again, we targeted a sample size of 191 participants. Recruitment and compensation of participants was parallel to Study 1. We recruited 227 participants and excluded 61 participants because they did not finish the questionnaire. Of the remaining 166, we excluded seven participants whose response pattern indicated that they walked away from the study (i.e., responses took longer than an hour) or who responded in less than three minutes (indicating inattentiveness) as well as five participants who informed us that their data should not be used. This resulted in a final sample of $N = 154$ participants (73.4% female; 69.5% psychology students; $M_{age} = 24.94$, $SD_{age} = 7.49$; 56.5% employed for $M = 19.27$ hrs, $SD = 13.06$). Roughly one fourth (24.0%) of participants reported experience with bonus payments, 5.4% of those specified that it was in the role of the employer, 91.9% in the role of applicants, and 2.7% in both roles.

5.1.2 | Procedure

The procedure was parallel to Study 1 except for the scenario presented.[7] Participants' task was to select employees for a fictional company's bonus program. They first received a description of a company that is specialized in finance and insurance and information on the bonus program with which employees could receive a bonus of up to 10% of their yearly salary (see Supporting Information: Material C). Additionally, criteria relevant to the selection for the bonus program were listed. Participants were then informed that they will be asked to accept or reject the preselection of employees. For each candidate, we randomly assigned a photo and, similar to Study 1, included two qualitative (strength and weakness) and two numeric criteria relevant to the bonus program (revenue and how many new customers were recruited). There were no salient differences between employees. Similar to Study 1, photos showed young adults of Caucasian ethnicity (see Supporting Information: Material D).

5.2 | Results

Table 1 displays correlations, means, and SDs for all variables in Study 2. Table 2 shows contingency tables for participants' rejection and acceptance of the preselection of applicants.[8]

TABLE 5 Regression results for trustworthiness facets of Study 2.

Predictors	Ability			Benevolence			Unbiased			Transparency		
	Estimates	CI	*p*	Estimates	CI	*p*	Estimates	CI	*p*	Estimates	CI	*p*
(Intercept)	3.63	3.30, 3.96	<.001	3.50	3.17, 3.83	<.001	3.30	2.88, 3.71	<.001	3.68	3.36, 4.00	<.001
H versus A	−0.40	−0.84, 0.05	.078	−0.35	−0.79, 0.10	.125	0.12	−0.44, 0.67	.681	−0.24	−0.67, 0.19	.266
H versus H-A	0.18	−0.28, 0.65	.436	−0.21	−0.67, 0.25	.371	0.07	−0.50, 0.65	.800	−0.16	−0.61, 0.29	.481
Preselection	−0.41	−0.86, 0.03	.069	−0.49	−0.93, −0.04	.032	−0.67	−1.22, −0.11	.019	−0.40	−0.83, 0.03	.070
H versus A × preselection	−0.05	−0.67, 0.56	.862	−0.11	−0.72, 0.50	.719	0.72	−0.05, 1.48	.065	0.33	−0.26, 0.92	.275
H versus H-A × Preselection	0.09	−0.54, 0.72	.780	0.15	−0.48, 0.77	.640	0.56	−0.23, 1.34	.163	0.39	−0.21, 1.00	.200

Note: A, automated DS; H, human decision-support (DS); H-A, hybrid DS; DS type was dummy-coded: 0 = human DS, 1 = system or hybrid DS. Coding of preselection balance: 0 = gender-balanced, 1 = predominantly-male. $N = 154$.

5.2.1 | Research questions and hypotheses

Manipulation check

The unbalanced preselection was perceived as less fair than the balanced one, $R^2 = 0.16$, $F(1, 152) = 28.32$, $p < .001$. This indicates that participants were aware of the difference in the gender distribution and perceived the condition with a disproportionate number of male applicants as unfair compared to the balanced preselection.

Trust, trustworthiness and trusting behavior

H1.2 proposed that trust-related variables will be negatively impacted for the bonus payment task if the recommendation comes from a human compared to an automated system. Table 3 shows the results of the corresponding regression analysis. Overall trustworthiness as well as its subsequently examined facets (see Table 5) did not significantly differ between the groups, but we found that the human DS was trusted more than the automated DS. Regarding trusting behavior, 26.5% of participants in the human and 21.9% in the automated condition rejected the preselection indicating no significant difference. H1.2 was thus not supported.

In RQ1 we asked whether differences in trust-related variables can be observed between the hybrid and the human or automated DS. We found no significant differences and only minor effect sizes between hybrid and human DS for all variables. For the comparison of hybrid with automated DS, hybrid DS was perceived more favorably with respect to trust, $t(103) = 3.62$, $p < .001$, $d = 0.71$, and trustworthiness, $t(103) = 2.54$, $p = .013$, $d = 0.49$. The hybrid group showed no difference in rejection rates (30.0%) compared to the automated or human DS. In response to RQ1, the hybrid DS was thus perceived more favorably than automated and on-par with human support.

EQ. 1 and EQ. 2 asked whether there are differences in trust-related variables between the human or system part in the hybrid DS group relative to the human- or system-only condition respectively. For the human DS, perceptions did not differ between the hybrid and the human-only DS condition for trust, $t(97) = 0.44$, $p = .66$, $d = 0.09$, and trustworthiness, $t(97) = 0.28$, $p = .78$, $d = 0.06$. For the automated-only versus automated as a part of hybrid DS comparison, there were no difference for trust, $t(103) = 1.09$, $p = .28$, $d = 0.21$, or trustworthiness, $t(103) = 0.18$, $p = .86$, $d = 0.03$.

Trust violations

H2 stated that a DS that provides a predominantly male preselection would be perceived more negatively for trust-related variables. Participants' overall trustworthiness as well as trust was rated lower when the preselection was predominantly male as opposed to when it was gender-balanced. Regarding trusting behavior, participants rejected 20.3% of the unbalanced and 5.9% of the balanced preselection. This difference failed to reach significance in a logistic regression (see Table 6). H2 was thus again only partially supported.

With H3, we proposed that an unbalanced recommendation provided by human support will lead to a more negative impact on trust-related variables than an unbalanced recommendation provided

TABLE 6 Results for the logistic regression on trusting behavior.

Predictors	Study 1			Study 2		
	Odds ratios	CI	*p*	Odds ratios	CI	*p*
(Intercept)	0.41	0.18, 0.86	.024	0.16	0.04, 0.46	.003
H versus A	1.43	0.52, 4.05	.494	0.76	0.13, 4.51	.753
H versus H-A	0.77	0.22, 2.54	.673	0.95	0.16, 5.69	.953
Preselection	1.83	0.63, 5.52	.270	3.73	0.96, 18.73	.075
H versus A × preselection	0.67	0.15, 2.99	.598	1.12	0.14, 9.01	.914
H versus H-A × preselection	1.73	0.34, 9.01	.510	1.43	0.18, 11.49	.730

Note: A, automated DS; H, human decision-support (DS); H-A, hybrid DS; DS type was dummy-coded: 0 = human DS, 1 = system or hybrid DS. Coding of preselection balance: 0 = gender-balanced, 1 = predominantly-male. $N_{Study1} = 170$, $N_{Study2} = 154$.

by an automated system. We found no interaction effects (see Tables 3 and 6), disconfirming H3.

RQ2 asked whether there are differences between trust-related variables when the preselection was balanced or unbalanced between the human and the hybrid DS. The corresponding analyses showed no interaction effects (see Table 3 for trust and trustworthiness, Table 6 for trusting behavior).

5.3 | Discussion Study 2

The results of Study 2 were almost identical to Study 1, except for the trust violation now influencing overall trustworthiness and the absence of differences for trustworthiness facets in the comparison of human and automated DS. However, for the facet ability the findings descriptively point in the same direction as in Study 1, indicating that there was a tendency to perceive humans more able to conduct the bonus payment preselection compared to automated systems. The largely similar results indicate that hybrid DS was again perceived on-par with human DS and that contrary to our expectations, a different task context did not lead to different perceptions of human relative to automated support. This finding could be fuel for the search for reasons for task-related effects on reactions to automated decisions (Langer & Landers, 2021). Possibly, because we kept the selection criteria in the two tasks comparable, each consisting of two qualitative and two numerical criteria, the contexts may have been perceived similar in terms of complexity and (un-)suitability for automated systems. An alternative explanation is that due to our Vignette study and our participants having little experience in either personnel selection or bonus payment contexts, participants were not aware of differences between the tasks that could have led to different reactions regarding automated decisions.

6 | OVERALL DISCUSSION

The goal of this paper was to shed light on trust in hybrid DS in two application contexts. The main findings of our studies are that (a) hybrid DS led to higher trust and trustworthiness perceptions than automated support and was on-par with human DS, (b) the impact of trust violations did not differ based on DS type, and (c) the context alone did not elicit different reactions to DS types.

The finding that hybrid DS was on-par with human support when it comes to trust and trustworthiness is consistent with research in domains other than trust showing similar results for healthcare service usage likelihood or procedural justice in selection (Gonzalez et al., 2022; Longoni et al., 2019). Whereas Gonzalez et al. (2022) examined effects on justice from the perspective of applicants, our results showed that similar effects can be found for trust from the perspective of decision-makers. Consistent with the fact that hybrid DS was on-par with human DS, it was perceived more favorably with respect to trustworthiness and trust than automated DS. This is also a notable finding because participants received no additional information regarding the system or regarding the interaction of humans and systems in the hybrid DS condition. This could also imply that any kind of human contribution (e.g., from simple monitoring to close collaboration in decision-making processes with automated systems) to decision-processes involving any kind of automated systems is sufficient to make trust in hybrid DS similar to trust in humans alone. In this regard, it could be interesting to see whether including further information about the automated system (e.g., describing the technology in greater detail; see also Langer, Hunsicker, et al., 2022) or the kind of interaction between human and system (e.g., varying the involvement of the human decision-maker) would change the attitudes of participants toward hybrid DS.

The consequences of hybrid DS being perceived as similar to human DS are manifold. It can mean that people react more positively to automated systems in high-risk tasks as soon as there is a human involved, an idea that is reflected in many ethical (Jobin et al., 2019) and legal documents (e.g., the EU's proposal for an AI Act) arguing for human oversight or humans in the loop in high-risk application contexts where automated systems may be used. Also, if people find it more acceptable to use automated systems when humans interact with these systems, this can be beneficial for the successful implementation of automated systems which has then the potential to increase decision quality and efficiency. However, having a human interacting with an automated system does not necessarily mean that decisions of the hybrid DS will actually become more

trustworthy – there are cases where humans may trust too much or not enough in automated systems for their decisions (Green, 2022; Parasuraman & Manzey, 2010).

Contrary to previous evidence (Arkes et al., 2007), we did not find that collaborations between humans and systems affect perceptions of the single parts of the hybrid DS. For instance, in contrast to Arkes et al. (2007), the human as part of the hybrid DS was not rated less favorably than human-only support. There are many possible reasons for this: differences in the perspective of participants, application context, perceived risk, or competence expectancy. The study by Arkes et al. (2007) was conducted in a medical context, where doctors diagnosed their patients with or without consulting an automated DS, and their participants were in the role of patients. In our studies, participants were in the role of decision-makers receiving hybrid DS. Furthermore, health decisions could be perceived as a more high-risk context than HR selection tasks, calling for more scrutiny in the assessment of decision-makers. Finally, attitudes toward system usage could also have changed in the time between the work by Arkes at al. (2007) and our experiment.

We also did not find the expected greater difference in trust and trustworthiness of the balanced compared to unbalanced preselection for human relative to automated DS (in contrast to emerging evidence by e.g., Bigman et al., 2022; Bonezzi & Ostinelli, 2021; Jago & Laurin, 2022; Langer, König, et al., 2022). One explanation for this could be that the effect of an unbalanced preselection on trust and trustworthiness in only a single decision situation is small, which made it less likely to find significant interaction effects. Potentially, such effects can only be observed if there are repeated unbalanced outputs that make the possible existence of unfair bias more salient (see Langer, König, 2022). Another explanation could be that only Caucasian photos were used in both studies. This lack of diversity could have influenced the perceptions of the balanced condition also reflecting biased decision-making because no ethnical minority applicants were represented. However, since the balanced preselection was overall perceived as significantly fairer, as fairer than the middle category of the scale ("neither fair nor unfair"), and since the unbalanced preselection was perceived as less fair than the middle category, participants at least perceived the balanced condition not to be especially problematic.

6.1 | Main practical implication

If proposals for regulation such as the European AI Act are implemented, hybrid DS will be a likely implementation of automated systems in high-risk situations such as HR management. In line with recent research highlighting the importance of examining perceptions of hybrid DS (Gonzalez et al., 2022; Langer & Landers, 2021; Nagtegaal, 2021), our study shows that humans trust hybrid DS similar to humans deciding independently. This means that even though there seems to be skepticism toward full automation for certain areas in HR (Langer & Landers, 2021), automation collaborating with a human may be more positively perceived in HR. HR managers could thus benefit from automated tools in selection without having to fear negative perceptions by other important stakeholders (e.g., applicants, as shown by Gonzalez et al., 2022; and as shown in our studies). As a downside, finding more trust in hybrid DS can mean that the implementation of automated systems may be less likely to be subject to scrutiny when there remains human oversight (Green, 2022). This can be problematic given that the collaboration of human and system does not necessarily results in better decisions – contrarily, adding human decision-makers can reduce decision quality in contrast to fully automated decisions (Schemmer et al., 2022).

6.2 | Limitations

First, our experiments were conducted with predominantly non-experts in selection and our experiments only involved a simulation of selection tasks—the selection situation may thus not involve particular risk for our participants. Higher risks could affect how real HR managers perceive different types of DS, for example when it comes to the perceived ability of support or importance of the task. Whereas possible fairness issues in our study did not have consequences, in practice, fairness issues—even in a single round of selecting applicants or employees for bonus payments—can have serious consequences (e.g., lawsuits) and significant trust violations. Moreover, task-dependent effects could also be more pronounced in real scenarios.

Second, we only examine initial trust in DS, thus the dynamic nature of trust is not reflected in our studies. Attitudes could change after repeated interactions and in cases where decision-makers have to live with the decision consequences. Nevertheless, because initial trust assessments may anchor attitudes (Hoff & Bashir, 2015) and make them difficult to change, examining initial assessment also remains relevant.

7 | CONCLUSION

Our studies demonstrate that, whereas fully-automated decisions involve possible negative reactions, the level of trust in hybrid DS may be no different than the level of trust in human-only support. To implement DS that benefits from the possible advantages of automated systems without having to fear negative reactions by important stakeholders, organizations could therefore consider keeping a human in the loop for high-risk decisions instead of implementing full automation. However, it is advisable to weigh the benefits and risks, as involving a human does not necessarily lead to better decisions. We are looking forward to research that assesses whether our findings replicate in different contexts and for real-life high-stakes decisions, and research examining why people sometimes prefer humans to remain an integral part of decision-making instead of fully automating decisions.

ACKNOWLEDGMENTS

Work on this paper was funded by the Volkswagen Foundation grant number AZ98513 and by the DFG grant 389792660 as part of TRR248. Both studies were preregistered (Study 1: https://aspredicted.org/blind.php?x=bm7vy8; Study 2: https://aspredicted.org/blind.php?x=zh56zp). Open Access funding enabled and organized by Projekt DEAL.

DATA AVAILABILITY STATEMENT

The data that support the findings of this study are available from the corresponding author upon reasonable request.

ORCID

Cornelius J. König http://orcid.org/0000-0003-0477-8293
Richard Bergs http://orcid.org/0000-0001-9697-2311
Markus Langer https://orcid.org/0000-0002-8165-1803

ENDNOTES

[1] Both studies were preregistered (Study 1: aspredicted.org/blind.php?x=bm7vy8; Study 2: aspredicted.org/blind.php?x=zh56zp). We do not report any of the exploratory hypotheses listed in the preregistration. Note that none of them were supported, results can be made available upon request.

[2] In the preregistration, we report power analyses for an ANOVA. However, we realized that using regression allowed us to display the results more concisely. Regarding the findings, this should not make a difference as ANOVA and regression are analogous.

[3] Participants in the hybrid DS condition responded to the item "I believe 'the trustee' can process all the necessary data required for the preselection" three times—the term "trustee" was replaced with either (a) "the cooperation between a human resources employee and an automated system," (b) "the human resources employee," or (c) "the automated system."

[4] For exploratory purposes, we asked participants whether they perceived the human or system to be male or female. Additionally, participants responded to four questions regarding their experience with algorithms.

[5] All analyses were conducted with and without the inclusion of ATI as a control variable. Because the result pattern did not differ, we only report the analyses that do not include ATI.

[6] For exploratory purposes, we included the factor participant gender in a logistic regression with the same specifications and found significant effects of gender for trusting behavior but not for the other trust-related variables. We found a significant main effect for participant gender, OR = 7.90, 95% CI [2.11, 51.71]), $p = .008$, and there was a significant interaction between participants' gender and the preselection balance, OR = 0.14, 95% CI [0.02, 0.74]), $p = .033$. Women were more likely to reject the preselection, this was especially the case for the balanced preselection (which was rejected by 7.7% of men and 39.7% of women). For the unbalanced preselection, the rejection rates were similar (42.9% for men and 45.8% for women). Because we only examined this exploratorily, and because this finding was not replicated in Study 2, we decided not to further discuss it.

[7] For exploratory purposes we measured whether participants prefer analytical or intuitive decision making. This was measured with two of the five subscales from the General Decision-Making Style questionnaire (Scott & Bruce, 1995). Results can be made available upon request.

[8] We only report the results without the ATI scale. The inclusion of this control variable only changed the result for the trustworthiness factor benevolence, where the preselection balance was no longer significant in the corresponding regression.

REFERENCES

Arkes, H. R., Shaffer, V. A., & Medow, M. A. (2007). Patients derogate physicians who use a computer-assisted diagnostic aid. *Medical Decision Making*, *27*(2), 189–202. https://doi.org/10.1177/0272989X06297391

Baer, M., & Colquitt, J. A. (2018). Moving toward a more comprehensive consideration of the antecedents of trust. In R. H. Searle, A.-M. I. Nienaber, & S. B. Sitkin (Eds.), *Routledge companion to trust* (pp. 163–182). Routledge.

Benbasat, I., & Wang, W. (2005). Trust in and adoption of online recommendation agents. *Journal of the Association for Information Systems*, *6*(3), 72–101. https://doi.org/10.17705/1jais.00065

Bigman, Y. E., & Gray, K. (2018). People are averse to machines making moral decisions. *Cognition*, *181*, 21–34. https://doi.org/10.1016/j.cognition.2018.08.003

Bigman, Y. E., Wilson, D., Arnestad, M. N., Waytz, A., & Gray, K. (2023). Algorithmic discrimination causes less moral outrage than human discrimination. *Journal of Experimental Psychology: General*, *152*(1), 4–27. https://doi.org/10.1037/xge0001250

Bonezzi, A., & Ostinelli, M. (2021). Can algorithms legitimize discrimination? *Journal of Experimental Psychology: Applied*, *27*(2), 447–459. https://doi.org/10.1037/xap0000294

Campion, M. C., Campion, M. A., Campion, E. D., & Reider, M. H. (2016). Initial investigation into computer scoring of candidate essays for personnel selection. *Journal of Applied Psychology*, *101*(7), 958–975. https://doi.org/10.1037/apl0000108

Colquitt, J. A., & Rodell, J. B. (2011). Justice, trust, and trustworthiness: A longitudinal analysis integrating three theoretical perspectives. *Academy of Management Journal*, *54*(6), 1183–1206. https://doi.org/10.5465/amj.2007.0572

Cook, M. (2016). *Personnel selection: Adding value through people—a changing picture*. John Wiley & Sons.

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., Peterson, S. A., Pomranky, R. A., Pierce, L. G., & Beck, H. P. (2003). The role of trust in automation reliance. *International Journal of Human-Computer Studies*, *58*(6), 697–718. https://doi.org/10.1016/S1071-5819(03)00038-7

European Commission. (2021). *Proposal for a Regulation Laying down Harmonised Rules on Artificial Intelligence (Artificial Intelligence Act)*. European Commission.

European Commission, Directorate-General for Communications Networks, Content and Technology. (2019). *Ethics guidelines for trustworthy AI*. Publications Office of the European Union. https://data.europa.eu/doi/10.2759/346720

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

Franke, T., Attig, C., & Wessel, D. (2019). A personal resource for technology interaction: Development and validation of the affinity for technology interaction (ATI) scale. *International Journal of Human–Computer Interaction*, *35*(6), 456–467. https://doi.org/10.1080/10447318.2018.1456150

Glikson, E., & Woolley, A. W. (2020). Human trust in artificial intelligence: Review of empirical research. *Academy of Management Annals*, *14*(2), 627–660. https://doi.org/10.5465/annals.2018.0057

Gonzalez, M. F., Liu, W., Shirase, L., Tomczak, D. L., Lobbe, C. E., Justenhoven, R., & Martin, N. R. (2022). Allying with AI? Reactions toward human-based, AI/ML-based, and augmented hiring processes. *Computers in Human Behavior*, *130*, 107179. https://doi.org/10.1016/j.chb.2022.107179

Green, B. (2022). The flaws of policies requiring human oversight of government algorithms. *Computer Law & Security Review*, *45*, 105681. https://doi.org/10.1016/j.clsr.2022.105681

Hickman, L., Bosch, N., Ng, V., Saef, R., Tay, L., & Woo, S. E. (2022). Automated video interview personality assessments: Reliability, validity, and generalizability investigations. *Journal of Applied Psychology*, *107*(8), 1323–1351. https://doi.org/10.1037/apl0000695

Höddinghaus, M., Sondern, D., & Hertel, G. (2021). The automation of leadership functions: Would people trust decision algorithms. *Computers in Human Behavior*, *116*, 106635. https://doi.org/10.1016/j.chb.2020.106635

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

Jago, A. S., & Laurin, K. (2022). Assumptions about algorithms' capacity for discrimination. *Personality and Social Psychology Bulletin*, *48*(4), 014616722110161. https://doi.org/10.1177/01461672211016187

Jarrahi, M. H. (2018). Artificial intelligence and the future of work: Human-AI symbiosis in organizational decision making. *Business Horizons*, *61*(4), 577–586. https://doi.org/10.1016/j.bushor.2018.03.007

Jobin, A., Ienca, M., & Vayena, E. (2019). The global landscape of AI ethics guidelines. *Nature Machine Intelligence*, *1*(9), 389–399. https://doi.org/10.1038/s42256-019-0088-2

Koivunen, S., Olsson, T., Olshannikova, E., & Lindberg, A. (2019). Understanding decision-making in recruitment: Opportunities and challenges for information technology. *Proceedings of the ACM on Human-Computer Interaction*, *3*(GROUP), 242. https://doi.org/10.1145/3361123

Körber, M., Baseler, E., & Bengler, K. (2018). Introduction matters: Manipulating trust in automation and reliance in automated driving. *Applied Ergonomics*, *66*, 18–31. https://doi.org/10.1016/j.apergo.2017.07.006

Kuncel, N. R., Klieger, D. M., Connelly, B. S., & Ones, D. S. (2013). Mechanical versus clinical data combination in selection and admissions decisions: A meta-analysis. *Journal of Applied Psychology*, *98*(6), 1060–1072. https://doi.org/10.1037/a0034156

Langer, M., Hunsicker, T., Feldkamp, T., König, C. J., & Grgić-Hlača, N. (2022). *"Look! It's a computer program! It's an algorithm! It's AI!": Does terminology affect human perceptions and evaluations of algorithmic decision-making systems?* CHI Conference on Human Factors in Computing Systems, p. 581. https://doi.org/10.1145/3491102.3517527

Langer, M., König, C. J., Back, C., & Hemsing, V. (2022). Trust in artificial intelligence: Comparing trust processes between human and automated trustees in light of unfair bias. *Journal of Business and Psychology*, 1–16. Advance online publication. https://doi.org/10.1007/s10869-022-09829-9

Langer, M., König, C. J., & Busch, V. (2021). Changing the means of managerial work: Effects of automated decision support systems on personnel selection tasks. *Journal of Business and Psychology*, *36*(5), 751–769. https://doi.org/10.1007/s10869-020-09711-6

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*, *123*:106878. https://doi.org/10.1016/j.chb.2021.106878

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50_30392

Lee, M. K. (2018). Understanding perception of algorithmic decisions: Fairness, trust, and emotion in response to algorithmic management. *Big Data & Society*, *5*(1), 205395171875668. https://doi.org/10.1177/2053951718756684

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

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

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review*, *20*(3), 709–734. https://doi.org/10.2307/258792

Mcknight, D. H., Carter, M., Thatcher, J. B., & Clay, P. F. (2011). Trust in a specific technology: an investigation of its components and measures. *ACM Transactions on Management Information Systems*, *2*(2), 1–25. https://doi.org/10.1145/1985347.1985353

Mosier, K. L., & Manzey, D. (2019). Humans and automated decision aids: A match made in heaven? M. Mouloua, P. A. Hancock & J. Ferraro, (Eds.), *Human performance in automated and autonomous systems* (pp. 19–42). CRC Press. https://doi.org/10.1201/9780429458330-2

Nagtegaal, R. (2021). The impact of using algorithms for managerial decisions on public employees' procedural justice. *Government Information Quarterly*, *38*(1), 101536. https://doi.org/10.1016/j.giq.2020.101536

Newman, D. T., Fast, N. J., & 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*, *160*, 149–167. https://doi.org/10.1016/j.obhdp.2020.03.008

Parasuraman, R., & Manzey, D. H. (2010). Complacency and bias in human use of automation: An attentional integration. *Human Factors: The Journal of the Human Factors and Ergonomics Society*, *52*(3), 381–410. https://doi.org/10.1177/0018720810376055

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

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

Schemmer, M., Hemmer, P., Nitsche, M., Kühl, N., & Vössing, M. (2022). A meta-analysis of the utility of explainable artificial intelligence in human-ai decision-making. *Proceedings of the 2022 AAAI/ACM Conference on AI, Ethics, and Society*, 617–626. https://doi.org/10.1145/3514094.3534128

Scott, S. G., & Bruce, R. A. (1995). Decision-making style: The development and assessment of a new measure. *Educational and Psychological Measurement*, *55*(5), 818–831. https://doi.org/10.1177/0013164495055005017

Sheridan, T. B., & Parasuraman, R. (2005). Human automation interaction. *Reviews of Human Factors and Ergonomics*, *1*(1), 89–129. https://doi.org/10.1518/155723405783703082

Skitka, L. J., Mosier, K. L., & Burdick, M. (1999). Does automation bias decision-making. *International Journal of Human-Computer Studies*, *51*(5), 991–1006. https://doi.org/10.1006/ijhc.1999.0252

Starke, C., & Lünich, M. (2020). Artificial intelligence for political decision-making in the European Union: Effects on citizens' perceptions of input, throughput, and output legitimacy. *Data & Policy*, *2*, e16. https://doi.org/10.1017/dap.2020.19

Tay, L., Woo, S. E., Hickman, L., Booth, B. M., & D'Mello, S. (2022). A conceptual framework for investigating and mitigating machine-learning measurement bias (MLMB) in psychological assessment. *Advances in Methods and Practices in Psychological Science*, *5*(1):251524592110613. https://doi.org/10.1177/25152459211061337

Thielsch, M. T., Meeßen, S. M., & Hertel, G. (2018). Trust and distrust in information systems at the workplace. *PeerJ*, *6*:e5483. https://doi.org/10.7717/peerj.5483

Villegas, S., Lloyd, R. A., Tritt, A., & Vengrouskie, E. F. (2019). Human resources as ethical gatekeepers: Hiring ethics and employee selection. *Journal of Leadership, Accountability and Ethics*, *16*(2), 80–88. https://doi.org/10.33423/jlae.v16i2.2024

de Visser, E. J., Pak, R., & Shaw, T. H. (2018). From 'automation' to 'autonomy': The importance oftrust repair in human–machine interaction.*Ergonomics*, *61*(10), 1409–1427. https://doi.org/10.1080/00140139.2018.1457725

Warszta, T. (2012). *Application of Gilliland's model of applicants' reactions to the field of web-based selection* [PhD Thesis, Christian-Albrechts-Universität zu Kiel, Germany].

Zerilli, J., Bhatt, U., & Weller, A. (2022). How transparency modulates trust in artificial intelligence. *Patterns*, *3*(4), 100455. https://doi.org/10.1016/j.patter.2022.100455

SUPPORTING INFORMATION

Additional supporting information can be found online in the Supporting Information section at the end of this article.

How to cite this article: Kares, F., König, C. J., Bergs, R., Protzel, C., & Langer, M. (2023). Trust in hybrid human-automated decision-support. *International Journal of Selection and Assessment*, *31*, 388–402. https://doi.org/10.1111/ijsa.12423