Langer, Markus; König, Cornelius J.; Back, Caroline; Hemsing, Victoria

Article — Published Version

Trust in Artificial Intelligence: Comparing Trust Processes Between Human and Automated Trustees in Light of Unfair Bias

Journal of Business and Psychology

Suggested Citation: Langer, Markus; König, Cornelius J.; Back, Caroline; Hemsing, Victoria (2022) : Trust in Artificial Intelligence: Comparing Trust Processes Between Human and Automated Trustees in Light of Unfair Bias, Journal of Business and Psychology, ISSN 1573-353X, Springer US, Vol. 38, Iss. 3, pp. 493-508, https://doi.org/10.1007/s10869-022-09829-9

This Version is available at:
https://hdl.handle.net/10419/313191

Journal of Business and Psychology (2023) 38:493–508
https://doi.org/10.1007/s10869-022-09829-9

ORIGINAL PAPER

Trust in Artificial Intelligence: Comparing Trust Processes Between Human and Automated Trustees in Light of Unfair Bias

Markus Langer[1] · **Cornelius J. König**[1] · **Caroline Back**[1] · **Victoria Hemsing**[1]

Accepted: 13 June 2022 / Published online: 28 June 2022

Abstract

Automated systems based on artificial intelligence (AI) increasingly support decisions with ethical implications where decision makers need to trust these systems. However, insights regarding trust in automated systems predominantly stem from contexts where the main driver of trust is that systems produce accurate outputs (e.g., alarm systems for monitoring tasks). It remains unclear whether what we know about trust in automated systems translates to application contexts where ethical considerations (e.g., fairness) are crucial in trust development. In personnel selection, as a sample context where ethical considerations are important, we investigate trust processes in light of a trust violation relating to unfair bias and a trust repair intervention. Specifically, participants evaluated preselection outcomes (i.e., sets of preselected applicants) by either a human or an automated system across twelve selection tasks. We additionally varied information regarding imperfection of the human and automated system. In task rounds five through eight, the preselected applicants were predominantly male, thus constituting a trust violation due to potential unfair bias. Before task round nine, participants received an excuse for the biased preselection (i.e., a trust repair intervention). The results of the online study showed that participants have initially less trust in automated systems. Furthermore, the trust violation and the trust repair intervention had weaker effects for the automated system. Those effects were partly stronger when highlighting system imperfection. We conclude that insights from classical areas of automation only partially translate to the many emerging application contexts of such systems where ethical considerations are central to trust processes.

Keywords Artificial intelligence · Trust · Personnel Selection · AI ethics · Errors

Introduction

Recent years have seen an upsurge in the use of automated systems based on artificial intelligence (AI) to support or even automate decision-making. Whereas classical application areas of automation were production or monitoring (Endsley, 2017), AI-based automated systems are now employed in tasks that affect the fate of individuals such as in medicine, jurisdiction, or management (Grgić-Hlača et al., 2019; Longoni et al., 2019; Raisch & Krakoswki, 2021). In management alone, such systems are on the verge of changing decision processes in personnel selection, performance management, or promotion (Cheng & Hackett, 2021; Tambe et al., 2019). Consequently, managers can increasingly choose to assign tasks to either human trustees (the party that is trusted) or automated systems as trustees. This warrants the need for managers as trustors (the party that trusts) to assess the trustworthiness of humans and automated systems to decide whether to rely on a respective trustee to perform a certain task.

Although research has shown similarities in trustworthiness assessments and trust processes for humans and automated systems as trustees (de Visser et al., 2018; Glikson & Woolley, 2020), this research predominantly stems from classical application contexts of automation (Rieger et al., 2022). There, trustworthiness assessments focus on classical performance measures associated with effectiveness and efficiency (e.g., prediction accuracy). In comparison, there is scarce research on trust processes in contexts where automated systems support decisions that affect individuals' fates (e.g., personnel selection; Langer et al., 2021). In such contexts, practitioners, researchers, and policy-makers are

✉ Markus Langer
Markus.Langer@uni-saarland.de

[1] Fachrichtung Psychologie, Universität des Saarlandes, Arbeits- & Organisationspsychologie, Campus A1 3, 66123 Saarbrücken, Germany

commonly concerned about ethical issues when using automated decision support (Jobin et al., 2019; Martin, 2019). In particular, fairness is important in such contexts (Raghavan et al., 2020) and determines trustworthiness assessments beyond classical performance measures. In fact, fairness issues (e.g., automated systems producing biased outputs) have presumably led to companies like Amazon losing trust and abandoning automated systems for managerial decisions (Dastin, 2018).

Thus, there is a lack of insight regarding trust processes in tasks where there is potential for violations of ethical standards such as fairness. For instance, although it is a matter of life and death to adequately trust air-traffic-control systems, such systems usually violate operators' trust through misses or false alarms (Parasuraman & Riley, 1997). In such contexts, illustrative for most research on trust in automation, fairness issues do not play an obvious role. However, such issues are salient throughout many novel application areas of automated systems (Raghavan et al., 2020). Therefore, it remains unclear whether effects found in classical trust in automation research translate to novel application contexts where accuracy is only one of many factors determining trust in those systems.

This study compares trust processes between human trustees and automated system as trustees in a novel application context for automated systems and in light of a potential violation of ethical standards. Specifically, participants took part in twelve rounds of a personnel selection task, where they received decision support (i.e., a preselection of applicants) from either a human colleague or an automated system. After the fourth task, participants repeatedly received potentially biased preselection outputs (i.e., trustees predominantly preselected male applicants) constituting trust violations due to violations of ethical standards. Before round nine, participants received an excuse for the biased preselection as a trust repair intervention. We investigate initial perceptions of trustworthiness, trust, and trust behavior, as well effects of the trust violation and the trust repair intervention. Furthermore, we examine expectations of perfection associated with automated systems (high performance, high consistency in performance) as a driver of possible differences in trust processes between human and automated trustees (Madhavan & Wiegmann, 2007).

Background and Hypothesis Development

Interpersonal Trust and Trust in Automation

Trust processes are central when trustors consider delegating tasks to trustees and when they receive decision-support (Bonaccio & Dalal, 2006). In those cases, the task fulfillment or the work output is important to trustors (Mayer et al., 1995). Consequently, relying on trustees' work outputs involves risks because trustees might not fulfill the trustors' expectations. This is true for interpersonal trust processes and in the case of trust in automation (J. D. Lee & See, 2004).

As a context where trust in decision-support is important, we chose to examine personnel selection as a context demanding ethically sensitive decisions where AI[1]-based automated systems are already a viable option for decision-support (Hickman et al., 2021). Although it is challenging to get reliable data on the actual use of automated systems in personnel selection practice, the number of vendors offering automated solutions for selection (e.g., as listed by Raghavan et al., 2020), growing interest in the validity of automated solutions (Hickman et al., 2021), and research with HR managers who have already used automated solutions (Li et al., 2021) indicate that AI-based selection is increasingly common. In a recent (likely non-representative) poll[2] among around 200 German HR managers, 29% indicated that they already use (11%), are in the pilot phase (6%), or are planning to use (11%) automated systems in the analysis of CVs and 20% indicated that they already use (2%), are in the pilot phase (3%), or are planning to use (15%) automated systems in the ranking of candidates. These numbers, however, have to be interpreted cautiously since automated selection may only be valuable for companies who need automated support to screen large numbers of applicants (Li et al., 2021).

Theoretical models on interpersonal trust (Mayer et al., 1995) and trust in automation (J. D. Lee & See, 2004) show that, independently of whether support is provided by a human or an automated system, main concepts in trust processes are trustworthiness, trust, and trust behavior. Trustworthiness reflects perceptions of the characteristics of a trustee. Trustor's trustworthiness perceptions arise from known or perceived characteristics of the trustee as well as trustees' performance in a task (J. D. Lee & See, 2004; Mayer et al., 1995). In a personnel selection task, this means that trustors will assess trustworthiness in relation to their goals for the task.

Trustworthiness of humans and automated systems is usually conceptualized with several facets (J. D. Lee & See, 2004; Mayer et al., 1995). In this study, we examine three facets of trustworthiness that trustors may consider for human and automated trustees: ability, integrity, and benevolence (Höddinghaus et al., 2020; Wang & Benbasat,

[1] When referring to AI, we refer to a broad range of methods that may be used to automate decision-making such as rule-based methods and to methods from the area of machine learning and deep learning (Cheng & Hackett, 2019).

[2] The poll is only available in German at https://www.ethikbeirat-hrtech.de/wp-content/uploads/2021/11/Umfrage_zur_Automatisierung_in_der_Personalarbeit_Update_Nov21.pdf

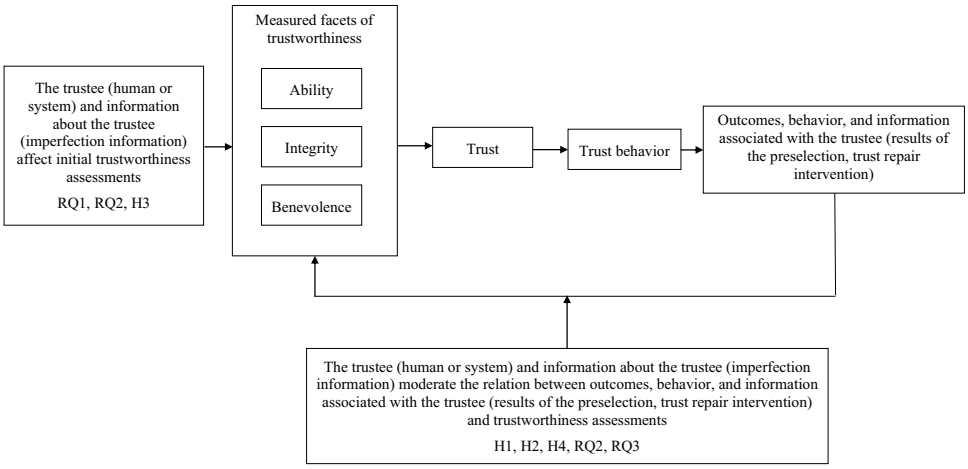

Fig. 1 The trust process investigated in the current study (figure based on Mayer et al., 1995). H = hypothesis, RQ = research question

2005), because these facets are in line with the main facets of trustworthiness proposed in Mayer et al.'s (1995) theoretical model of trust. In the case of personnel selection, ascribing high *ability* means that trustors believe that trustees can successfully select suitable applicants (Langer et al., 2021). *Integrity* captures whether trustors believe that trustees provide unbiased recommendations, and trustors may evaluate whether trustees follow ethical standards that they value (Den Gilliland, 1993; Hartog & De Hoogh, 2009). Finally, *benevolence* is the perception that trustees consider trustors' interests, goals, and values (Höddinghaus et al., 2020).

The second main concept within trust processes is trust. Mayer et al., (1995 p. 712) define trust as "a willingness to be vulnerable towards the actions of trustees without explicitly controlling or supervising trustees." In interpersonal trust and trust in automation, the relation between trustworthiness and trust is mostly straightforward: All else being equal, if trustors estimate trustworthiness of trustees to be comparably high, their trust in this trustee will be comparably strong.

Trust behavior is the behavioral outcome of trust (J. D. Lee & See, 2004). For example, trust behavior manifests when trustors actually delegate tasks to trustees or when they follow advice provided by trustees. This implies that trustors accept the risk that they may need to respond to a failure in a task or may receive bad advice (Mayer et al., 1995). Depending on the respective outcome (e.g., the work output or the quality of advice), trustors will re-evaluate trustees' trustworthiness, starting another cycle in the trust process (J. D. Lee & See, 2004; Mayer et al., 1995).

Thus, trust processes are dynamic (Glikson & Woolley, 2020). Specifically, there is an initial level of trustworthiness, trust, and trust behavior that can increase over time if trustors perceive trustees to perform successfully (J. D. Lee & See, 2004; Mayer et al., 1995). However, trusting someone or something always runs the risk of unfulfilled expectations, and trust violations (e.g., trustees' producing low quality work outcomes) can reduce trustworthiness, trust, and trust behavior (Kim et al., 2006). It is, however, possible to repair trust (e.g., through excuses) which can then rebuild, and positively affect perceptions of trustworthiness, trust, and trust behavior (de Visser et al., 2018; Tomlinson & Mayer, 2009).

Differences in Trust Processes Over Time

Whereas the basic concepts as well as dynamics (e.g., effects of trust violations or trust repair interventions) exist for interpersonal trust and for trust in automation, research has indicated differences for humans versus systems as trustees (de Visser et al., 2018; Madhavan & Wiegmann, 2007). Based on classical trust in automation research, Madhavan and Wiegmann (2007) propose a framework where they contrast trust in human versus automated systems. Their basic assumption is that humans are perceived as adaptable to situations, whereas automated systems are perceived as invariant. Furthermore, they propose that trustors expect perfection from automated systems because they are developed for a purpose and should work near-to-perfection for this purpose, whereas humans are expected to be imperfect. According to Madhavan and Wiegmann, those expectations lead to trustors being more observant and less forgiving of errors by automated compared to human trustees. Whereas Madhavan and Wiegmann (2007) mostly refer to the effects of trust violations, other research emphasizes that expectations of consistency may also affect trust repair processes (de Visser et al., 2016, 2018). For instance, trust rebuilding may be more difficult for automated trustees.

To summarize, differences in expectations of consistency and perfection likely affect initial trustworthiness assessments. Additionally, there might be different reactions to trust violations and trust repair interventions depending on the nature of the trustee (de Visser et al., 2016). Figure 1 summarizes the main concepts and their relation in trust

processes and highlights our hypotheses and research questions in the trust process as proposed by Mayer et al. (1995). The following sections provide further rationale for our proposed hypotheses and research questions.

Initial Trustworthiness Assessments

There is research proposing that initial trustworthiness assessments of automated systems might be comparably higher than for human trustees. Specifically, Madhavan and Wiegmann (2007) model proposes that humans believe that systems should work as intended, leading to high levels of initial trustworthiness (see also Dzindolet et al., 2003; Parasuraman & Manzey, 2010). Consequently, trustors should start with lower levels of trustworthiness for human trustees because of more uncertainty about their abilities, values, and intentions (Madhavan & Wiegmann, 2007).

In contrast to the propositions by Madhavan and Wiegmann (2007), there is research indicating that trustworthiness assessments could be lower for automated systems compared to human trustees (Rieger et al., 2022), which may especially be the case in tasks that affect the fate of individuals such as personnel selection (Langer & Landers, 2021; M. K. Lee, 2018). This might be partly because people believe automated systems are less capable of fulfilling such tasks (M. K. Lee, 2018). Specifically, people may believe that automated evaluation of individuals' characteristics does not adequately capture task complexity, is dehumanizing, and does not consider ethical peculiarities in such contexts (Grove & Meehl, 1996; Newman et al., 2020).

Overall, research stemming from classical areas of automation would suggest high levels of initial trustworthiness for automated systems (Madhavan & Wiegmann, 2007). In novel application contexts, there are arguments indicating that initial assessments of system trustworthiness could be higher or lower compared to human trustees (Langer & Landers, 2021; M. K. Lee, 2018). Considering that trust and trust behavior should result from initial trustworthiness assessments, we ask.

> Research Question (RQ) 1: Is there an initial difference for trustworthiness assessments, trust, and trust behavior between the automated system and the human trustee?[3]

Implications of Trust Violations

Different expectations of humans and automated systems as trustees might moderate the re-evaluation of trustworthiness after trust violations. In fact, trustworthiness assessments seem to suffer more for automated systems than for human trustees when trustors experience the first error (Bahner et al., 2008; Dzindolet et al., 2003). It is commonly assumed that this first-error-effect results from people expecting that systems should show consistently high performance (Dietvorst & Bharti, 2020; Parasuraman & Manzey, 2010). When recognizing first errors, trustors realize that systems are not perfect and that system performance varies, thereby strongly negatively affecting trustworthiness assessments (Dzindolet et al., 2003). In contrast, trustors might not expect human trustees to constantly work near-to-perfection; thus, errors are expected, leading to comparably weaker trust violation effects (Madhavan & Wiegmann, 2007).

However, research suggesting this moderating effect also predominantly stems from classical application areas of automation, where trust violations are mainly associated with trustee characteristics that relate to the trustworthiness facet *ability*. For instance, in such studies, trustees would miss alarms, produce false alarms, or provide less than perfect predictions (Hoff & Bashir, 2015). In personnel selection, trust violations can also be ability-associated, meaning that a trustee recommends unsuitable applicants. Additionally, trustees may produce ethically questionable outcomes (i.e., discriminating against minority applicants) that can affect trustworthiness assessments (Bonezzi & Ostinelli, 2021). In cases where trust violations are based on violations of ethical considerations, people might have stronger negative reactions for human trustees as they may believe that automated systems do not actively discriminate against specific groups of people (Bigman et al., 2022; Jago & Laurin, 2022). However, previous research on trust in automation would suggest that errors by automated systems result in strong negative effects on trustworthiness which is why we propose:

> Hypothesis 1[4]: After a trust violation, trustworthiness, trust, and trust behavior in the automated system condition will decrease more compared to the human trustee condition.

[3] Research questions and hypotheses were preregistered under https://aspredicted.org/sj9ud.pdf.

[4] In the preregistration, we mentioned moral reasoning as a dependent variable. Since we only measured moral reasoning for exploratory purposes, we did not include results for this variable. Additionally, we measured perceived transparency and flexibility as additional facets of trustworthiness but do not include them in the paper since the theoretical model by Mayer et al. (1995) states ability, integrity, and benevolence are the main facets of trustworthiness. Additional results are available under https://osf.io/j5wc9/. In the preregistration, we also mentioned the following research question: "Will participants realize trust violations by the automated agent later?" but omitted those analyses as we realized they do not provide insights beyond the other hypotheses.

Implications of Trust Repair Interventions

Similar moderating effects associated with the nature of the trustee might occur for trust repair interventions. Specifically, systems are deployed with a specific set of functions and level of quality. Any system improvements might require system updates or a new system (Höddinghaus et al., 2020). In contrast, human trustees are more adaptable. If there are trust repair interventions for human trustees (e.g., excuses), trustors might assume that the trustee will do their best to not let this error happen again (de Visser et al., 2018; Tomlinson & Mayer, 2009). If there are trust repair interventions associated with automated trustees, trustors might still believe that a system will produce similar errors in future as they may perceive systems as invariant (Dietvorst et al., 2015). Thus, we conclude,

> Hypothesis 2: After the trust repair intervention, trustworthiness, trust, and trust behavior in the automated system condition will increase less compared to the human trustee condition.

Differences in Trustworthiness Facets

The previously mentioned differences between human and automated trustees may also influence single facets of trustworthiness. Regarding ability, trustors might start off with higher levels of ability assessments for automated systems, but those could suffer more strongly from trust violations and might be less affected by trust repair interventions (Madhavan & Wiegmann, 2007). Regarding integrity, there is evidence that automated systems might be assessed as more consistent and less biased compared to human trustees (Bonezzi & Ostinelli, 2021; Jago & Laurin, 2022; Langer & Landers, 2021; M. K. Lee, 2018). In addition, for human trustees, trust violations associated with ethical considerations (e.g., a biased preselection) might have stronger effects as people might be more outraged by such trust violations by human trustees (Bigman et al., 2022). For trust repair effects, it is possible to assume stronger effects for human trustees as people believe that humans can learn from their mistakes (Tomlinson & Mayer, 2009). However, it is also possible that humans may not believe that biased human trustees can change (i.e., assuming that biases could reflect stable attitudes). For benevolence, humans might be perceived as more benevolent than automated systems because they are more likely to be able to consider trustors' interests (Höddinghaus et al., 2020). However, we are not aware of research that investigates the benevolence of human and automated systems with respect to trust violations and trust repair interventions. This list of tentative assumptions regarding the facets of trustworthiness subsume under.

> RQ2: Is there an initial difference and different effects for trust violations and trust repair interventions for the facets of trustworthiness for human and automated systems as trustees?

Expectations of Perfection as a Driver of Differences Between Human and Automated Trustees?

Research argues that differences between human trustees and automated systems as trustees are driven by expectations of perfection (Hoff & Bashir, 2015; Madhavan & Wiegmann, 2007). We thus experimentally manipulate these expectations by emphasizing that the automated system or the human trustee might not always produce error-free outputs, thus highlighting their potential imperfection (Bahner et al., 2008). If the expectation of perfection causes differences in trust processes between human and automated systems as trustees, making potential imperfection salient might affect perceptions of automated systems in a way that is more similar to reactions to human trustees (de Visser et al., 2016). Thus, we propose:

> Hypothesis 3: Trustworthiness, trust, and trust behavior in the automated system with information about imperfection condition will initially be lower compared to the condition without such information.
> Hypothesis 4: Following a trust violation, trustworthiness, trust and trust behavior in the automated system with information about imperfection condition will decrease to a lesser extent compared to the condition without such information.
> RQ3: Will there be interaction effects between the trust repair intervention and the information regarding imperfection for the automated system for trustworthiness, trust, and trust behavior?

Method

Sample

We determined the required sample size with G*Power (Faul et al., 2009). In an ANOVA with a within-between interaction effect of $\eta^2_p = 0.01$, $N = 108$ participants would be necessary for a power of $1 - \beta = 0.80$. Assuming a small to medium effect for a between-groups effect of $\eta^2_p = 0.04$, $N = 148$ participants would be necessary to achieve a power of $1 - \beta = 0.80$. Therefore, we wanted to recruit between 108 and 148 participants. This study was advertised to people interested in human resource management. We posted the advertisement on different social media groups, around the

campus of a German university, and in the downtown area of a German city.

We anticipated issues during data collection (e.g., technical issues), so we continued data collection until our sample consisted of $N = 211$ participants. We excluded 10 participants because they did not follow the instructions in the experimental procedure and 3 because they reported technical issues. Furthermore, we removed 3 participants because they did not recall that there were task rounds where male applicants had been predominantly selected. Additionally, we removed 30 participants because they indicated that they received advice from a human when they actually were in the automated system condition and vice versa. Finally, we removed 44 participants because they indicated they were not told that the trustee can make errors when they were actually told so and vice versa. The final sample consisted of N[5] = 121 German participants (79% female), of which 93% studied psychology. The mean age was 23.56 years ($SD = 5.47$), participants took a mean of 32 min ($SD = 10$) to complete the study, had experienced a median of three personnel selection processes as applicants, and 24% of participants indicated experience in applicant selection.

Procedure

In a 2 (human vs. automated trustee) × 2 (no information regarding imperfection vs. information regarding imperfection) online experiment with 12 task rounds, participants were randomly assigned to one of the four groups. Participants were instructed to imagine that they were responsible for personnel selection in an HR department at a large insurance company operating in Germany (all instructions and items in this study were presented in German). They were informed that their company is recruiting trainees for its subsidiaries. We chose trainees at an insurance company because of the equal gender distribution in insurance jobs in Germany (Rudnicka, 2020). Furthermore, participants were informed that they will receive support in selecting candidates through a preselection of applicants from a larger applicant pool across 12 situations (representing consecutive days where there are new applications). Participants were informed that their task was to evaluate the quality of the applicant preselection regarding three organizational goals. These goals were customer satisfaction, innovation, and diversity. For each goal, participants read a passage outlining how the company defines these goals (see Supplemental Material A). The purpose of this was to introduce diversity explicitly so that not providing a diverse preselection would indicate a trust violation. Note that we chose two more goals to (a) provide participants with quality criteria for the preselection, (b) make the task more realistic, and (c) make the focus on diversity less obvious.

After being introduced to their task and the company's goals, participants received a job description including the desired qualifications of applicants. Then, participants received a tutorial round introducing them to the general process of each upcoming task round. In the tutorial (and in each task round) participants received between six and eight application photo-like pictures of white male and female applicants (see Supplemental Material B). We chose to only include white applicants as balancing for racial diversity would have greatly increased study complexity. The pictures were taken from https://generated.photos, a website that uses AI to produce realistic pictures of human faces.

For each applicant, there was additional information accompanying the picture (i.e., family name, age, years of job experience, final high school grade, strengths). During the tutorial, participants were informed that they should analyze the preselection against the company goals and the job description. They were also told that they will receive the question: "Do you want to see the statistics of the applicant pool?" If participants responded with "No" they were directed to the next page in the online tool. If participants responded with "Yes," they were directed to a page showing them the underlying statistics of the applicant pool for a given task round. In these statistics, participants were informed about the number of applicants, the percentage of male and female applicants as well as provided with a list of means and distribution information for further information (e.g., "the mean number of years of job experience of today's applicant pool was 1.3 with a deviation of 0.3"). In the tutorial, participants were instructed to respond with "Yes" so that every participant would see what happens if they request the applicant pool statistics. We included the option to view the applicant pool to increase psychological fidelity, as well as giving participants the option to check whether there was gender diversity in the applicant pool.

At the beginning of the actual experiment, participants received information that there was an increasing number of applicants making it necessary to have a preselection stage where they as hiring manager receive support. They were told that this support is a colleague (an automated system) who analyzes and preselects applicants. In the imperfection condition, participants were additionally informed that "the colleague (automated system) usually produces good work outcomes but that there are always possibilities for errors." Participants were then informed that they had the opportunity to evaluate the preselection and accept or reject

[5] If we include participants who did not remember whether there were rounds where there were predominantly male applicants, who did not remember the nature of the trustee, and who did not remember that there was information regarding imperfection (resulting in $N = 198$), results remained mostly stable with one exception: The results for the three-way interactions were mostly not significant any more. We thus emphasize that the results of the three-way interaction should be interpreted cautiously.

it. Participants then saw the information of the preselected applicants.

Afterwards, participants responded to the fairness item, to three items assessing whether the trustee adhered to the company's goals (customer satisfaction, innovation, diversity), to the trustworthiness (i.e., ability, integrity, benevolence; only in situations 1, 3, 5, 7, 9, and 12) and trust items. Participants then indicated whether they wanted to see the applicant pool statistics. After clicking "No" or after seeing those statistics, participants were asked: "Do you accept this preselection?," which was used as a measure of trust behavior. This process was repeated for twelve task rounds.

In task rounds five through eight, there were predominantly male applicants included in the preselection (see Supplemental Material B, see also Supplemental Table 1 for the numbers of male and female applicants included in the preselection per round). Thus, rounds five through eight constitute the trust violation phase. Before round nine, participants received the following information: "Dear colleague, in previous preselection outcomes, there were more male than female applicants, although the applicant statistics indicated that there were about as much women as there were men in the applicant pool. We were made aware of this issue and it has been solved. We apologize for this. Applicant preselection in the future should follow the goals of the organization again. Thank you for your understanding." This constituted the trust repair intervention. After the last task round, participants responded to exploratory items asking them whether they were satisfied with the advisor and if they would want to work together with this advisor in future. Finally, we measured propensity to trust in humans and in technology and collected demographic information.

Measures

Since participants in the main study needed to report their evaluation of trustees several times throughout the study, we wanted to optimize the use of items to reduce the burden for participants. To determine which items to keep, we conducted a pre-study with 54 student participants. In this pre-study, participants were instructed to imagine that they work at an HR department. They were told that they will receive support for a personnel selection task and were randomly assigned to either the human or the automated trustee condition. Participants then responded to twelve items assessing perceptions of ability, integrity, benevolence, and trust. We removed five items because dropping them had the least negative influence on *Cronbach's α* reliability of the scales.

Unless otherwise stated, participants responded to the items on a scale from 1 (strongly disagree) to 7 (strongly agree). To keep the gender of the human trustee undefined, the items did not mention the colleague's gender (using the inclusive "Kolleg/in" in German).

Ability was measured with two items taken from Höddinghaus et al., (2020). A sample item was "I believe the colleague/the automated system has the competency to consider all important information for the decision." Integrity was measured with two items[6] taken from Höddinghaus et al., (2020) who adapted the items from Wang and Benbasat (2005). A sample item was "I believe that the colleague/the automated system makes unbiased decisions." Benevolence was measured with two items taken from Wang and Benbasat (2005). A sample item was "The colleague/the automated system would consider my interests." Trustworthiness was calculated as the mean of the items for ability, integrity, and benevolence. Trust was measured with two items taken from Thielsch et al., (2018). A sample item was "I would strongly rely on the colleague/the automated system." Trust behavior was measured with the item "I accept this preselection" with the response options "Yes" or "No". Accepting the preselection reflects higher trust behavior.

Manipulation Checks

To assess if the trust violation and trust repair intervention worked as intended, we used two manipulation check measures. First, after each preselection they received, participants responded to the item "In your opinion, did the preselection adhere to the company's goals?" on a scale from 1 (disagree) to 5 (agree). Participants responded to this item three times, once for each of the three company goals (customer satisfaction, technology and innovation, diversity). This allowed us to assess if participants realized that the biased preselection violated the organization's diversity goal. Second, participants responded to the item "I perceived the preselection to be fair," assessing whether participants actually perceived a predominantly male preselection to be unfair.

Results

Table 1 displays correlations, descriptive information, and reliabilities for the means of all variables across all tasks. Figures 2 and 3 show line-graphs for the continuous dependent variables for each task. We used ANOVAs and

[6] For integrity, exclusion of items in the pre-study left us with one item but we wanted to include two items per facet of trustworthiness. Therefore, we decided to make two items out of the item "I believe that X makes unbiased and objective decisions". Consequently, we included the items "I believe that X makes unbiased decisions" and "I believe that X makes objective decisions".

Table 1 Means, standard deviations, and correlations over all measurement points

Variable	*M*	*SD*	1	2	3	4	5	6	7	8
1. Ability	4.36	0.93	*.90 (.86–.94)*							
2. Integrity	4.18	1.11	.21*	*.88 (.82–.91)*						
3. Benevolence	3.81	0.96	.29**	.28**	*.77 (.68–.83)*					
4. Trustworthiness	4.11	0.71	.68**	.74**	.72**	*.77 (.65–.85)*				
5. Trust	3.66	0.98	.70**	.17	.32**	.54**	*.95 (.91–.96)*			
6. Trust behavior	0.57	0.21	.41**	.18*	.23*	.38**	.41**	-		
7. Imperfection	-	-	.02	.04	−.06	.01	.06	−.06	-	
8. Human vs. system	-	-	−.34**	.43**	.00	.08	−.41**	.05	.04	-

Mean reliability (calculated as the mean of reliability for all task rounds) of the measures is presented in italics in the diagonal and the range of reliabilities throughout the task rounds is presented in brackets. We report *Cronbach's α* for measures with more than two items and for measures with two items the Spearman-Brown correlation as suggested by Eisinga et al. (2013). Coding of trust behavior: 0 = rejecting preselection (low trust behavior), 1 = accepting preselection (high trust behavior). Coding of human vs. System: 0 = human, 1 = system. Coding of imperfection: 0 = no information regarding imperfection, 1 = information regarding imperfection. $N = 121$

*$p < .05$, ** $p < .01$

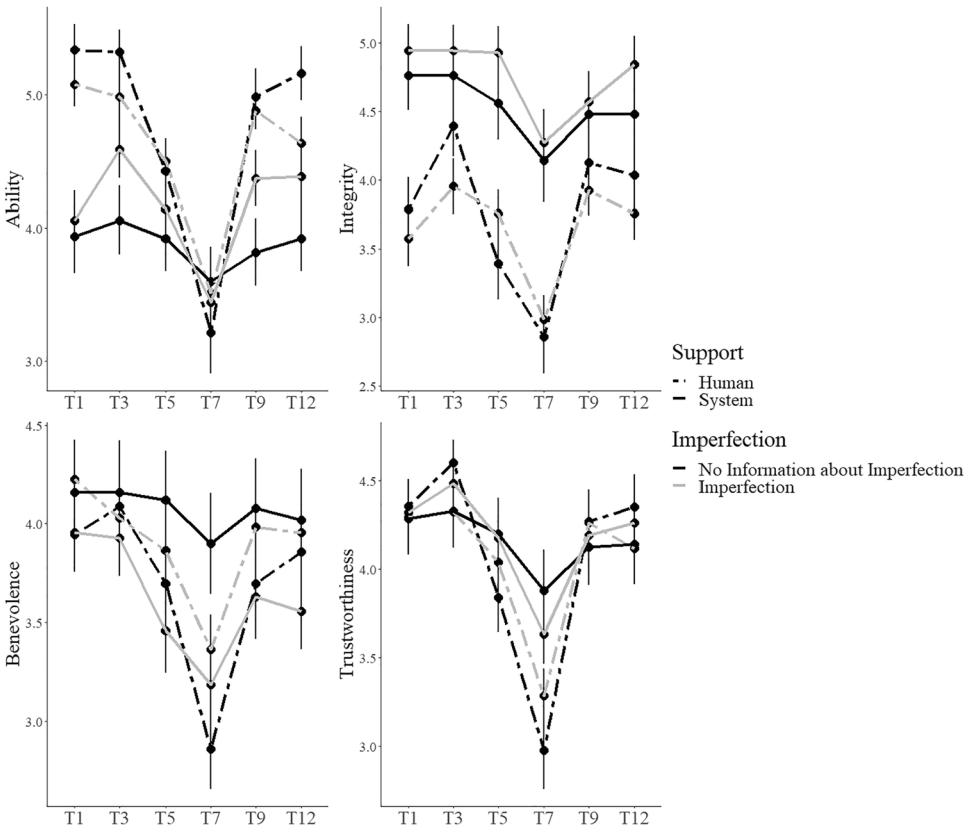

Fig. 2 Line graphs for the mean values of the dependent variables measured at Tasks 1, 3, 5, 7, 9, and 12. T = task. Error bars indicate standard errors

within-between interaction contrasts to analyze the data. Unless otherwise stated, we included human versus automated system as well as no information about imperfection versus information about imperfection as between-participant independent variables. As within-participant independent variable, we included the task rounds. Within-between contrasts analyze whether the difference between two task rounds differs depending on between-participant independent variables. In addition to round-by-round between-within contrasts, we present results of between-within interaction contrasts where we combined all measures in one phase into one measure for each phase: the initial phase (task rounds one through four), trust violation phase (task rounds five through eight), and the trust repair phase (task rounds nine through twelve). Since Hypotheses 1, 2, and 4 and RQ3 proposed or asked about differences

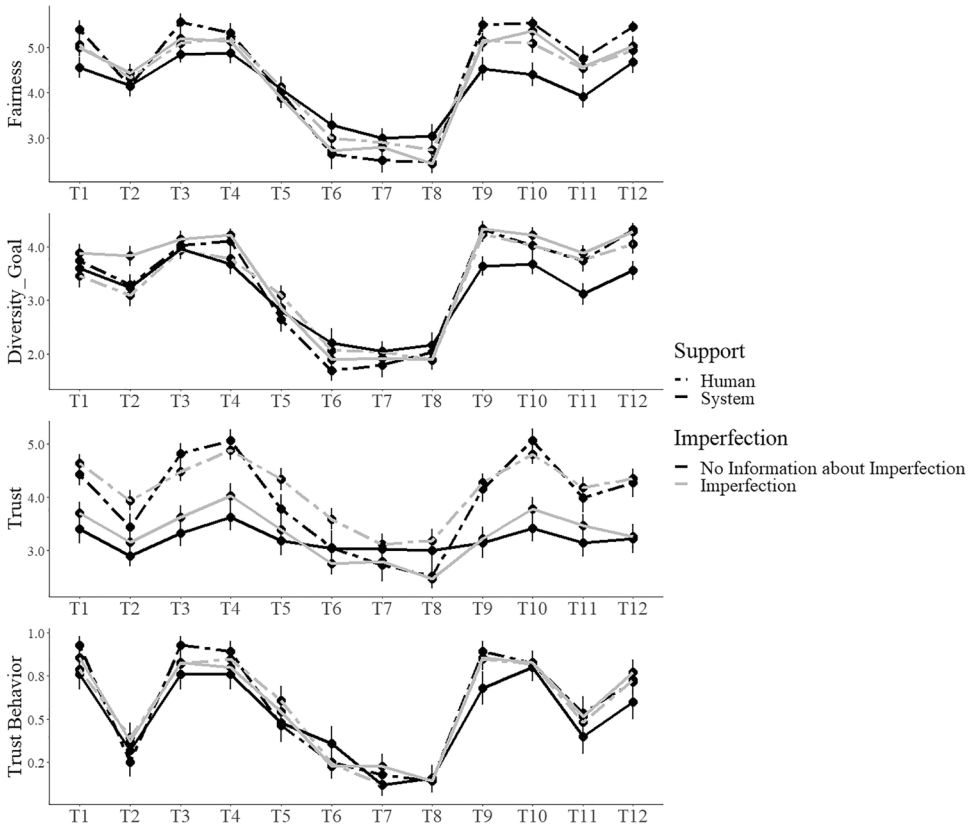

Fig. 3 Line graphs for the mean values of the dependent variables measured for all tasks. T = task. Error bars indicate standard errors

between the phases, these analyses were used as basis for the analysis for these Hypotheses and RQ3.

Manipulation Checks, Research Questions, and Hypotheses[7]

Figure 3 shows that both perceptions of fairness and evaluations of the fulfillment of the diversity goal decreased under the trust violation and recovered through the trust repair intervention. This indicates that those manipulations had their intended effects. Tables 2 and 3 show the results of the regression analyses for the evaluation of trustworthiness, trust, and facets of trustworthiness over time.

Trust worthiness, Trust and Trust Behavior

RQ1 asked whether there are initial differences for trustworthiness assessments, trust, and trust behavior between the automated system and the human trustee. For this RQ, we analyzed differences in trustworthiness and trust for only the first task as it reflects an appropriate time for initial evaluations of trustworthiness, trust, and trust behavior. There were no significant differences regarding initial trustworthiness $F(1,117)=0.02$, $p=0.92$, $\eta_p^2=0.00$, but participants had more trust in the human compared to the automated trustee $F(1,117)=20.64$, $p<0.01$, $\eta_p^2=0.15$. Regarding trust behavior, the percentage of participants accepting the preselection of applicants provided by the human (85% of participants accepted) or automated system (82% of participants accepted) showed no substantial difference (see Supplemental Material C for the results for trust behavior per task round). As such, we did not conduct further analyses. In sum, the answer to RQ1 is: Trust was initially higher for human trustees but there was no difference for trustworthiness and trust behavior.

Hypothesis 1 proposed that after a trust violation, trustworthiness, trust, and trust behavior in the automated system condition will decrease more compared to the human trustee condition. In contrast to the hypothesis, Figs. 2 and 3 and Tables 2 and 3 indicate that the trust violation more strongly reduced trustworthiness and trust for the human trustee. For trust behavior, we initially analyzed the percentage of people who accepted the preselection depending on the conditions and the phases. For the initial phase and in the human condition, participants accepted the preselection in 73% of cases, and in the case of the automated system, 69% of cases were accepted. For the trust violation phase and in the human condition, participants accepted the preselection in 27% of cases compared to 28% in the case of the automated system indicating no differences in the decline of trust behavior. Overall, results did not support Hypothesis 1.

[7] Including propensity to trust in the analyses did not change the results in a way that would have changed the interpretation of the results.

Table 2 Results of the within-between contrast analyses for trustworthiness and the facets of trustworthiness

		Trustworthiness			Ability			Integrity			Benevolence		
		$F(1,117)$	p	η_p^2	$F(1,117)$	p	η_p^2	$F(1,117)$	p	η_p^2	$F(1,117)$	p	η_p^2
Results per phase													
Task round*trustee	Trust violation	12.85	< .01	.10	18.23	< .01	.14	8.12	< .01	.07	2.56	.11	.02
	Trust repair	11.42	< .01	.09	11.87	< .01	.09	15.23	< .01	.12	2.75	.10	.02
Task round*trustee*imperfection	Trust violation	6.16	.01	.05	3.52	.06	.03	5.04	.03	.04	5.30	.02	.04
	Trust repair	4.07	.046	.03	6.64	.01	.05	2.62	.11	.02	1.31	.26	.01
Results per round													
Task round*trustee	Round 1 vs. Round 3	0.07	.79	.00	3.40	.07	.03	5.73	.02	.05	0.01	.94	.00
	Round 3 vs. Round 5	4.71	.03	.04	3.23	.08	.03	9.24	< .01	.07	0.02	.89	.00
	Round 5 vs. Round 7	6.43	.01	.05	6.75	.01	.06	2.09	.15	.02	11.96	< .01	.09
	Round 7 vs. Round 9	15.29	< .01	.12	15.86	< .01	.12	17.25	< .01	.13	4.70	.03	.04
	Round 9 vs. Round 12	0.34	.56	.00	0.21	.65	.00	2.92	.09	.02	0.97	.33	.01
Task Round*trustee*imperfection	Round 1 vs. Round 3	1.73	.19	.02	1.38	.24	.01	0.63	.43	.01	0.81	.37	.01
	Round 3 vs. Round 5	5.57	.02	.05	2.58	.11	.02	6.37	.01	.05	3.96	.049	.03
	Round 5 vs. Round 7	1.26	.26	.01	1.80	.18	.02	0.00	.99	.00	2.54	.11	.02
	Round 7 vs. Round 9	2.81	.10	.02	5.00	.03	.04	0.71	.40	.01	1.55	.22	.01
	Round 9 vs. Round 12	1.28	.26	.01	0.75	.39	.01	1.40	.24	.01	0.45	.50	.00

Notes. Trust Violation denotes the comparison of the mean value of the respective dependent variable of the trust violationphase compared to the initial phase. Trust Repair denotes the comparison of the mean value of the respective dependentvariable of the trust repair phase compared to the trust violation phase. $N = 121$. $n_{no\_information\_human} = 33$, $n_{no\_information\_system} = 35$, $n_{imperfection\_human} = 28$, $n_{imperfection\_system} = 25$

Table 3 Results of the within-between contrast analyses for trust

		Trust		
		$F(1,117)$	p	η_p^2
Results per phase				
Task round*trustee	Trust violation	14.70	< .01	.11
	Trust repair	16.79	< .01	.13
Task round*trustee*imperfection	Trust violation	8.80	< .01	.07
	Trust repair	6.78	.01	.06
Results per round				
Task round*trustee	Round 1 vs. Round 2	2.17	.14	.02
	Round 2 vs. Round 3	5.02	.03	.04
	Round 3 vs. Round 4	0.03	.87	.00
	Round 4 vs. Round 5	2.52	.12	.02
	Round 5 vs. Round 6	2.70	.10	.02
	Round 6 vs. Round 7	3.85	.052	.03
	Round 7 vs. Round 8	0.21	.65	.00
	Round 8 vs. Round 9	14.92	< .01	.11
	Round 9 vs. Round 10	2.44	.12	.02
	Round 10 vs. Round 11	6.92	.01	.06
	Round 11 vs. Round 12	1.98	.16	.02
Task round*trustee*imperfection	Round 1 vs. Round 2	0.58	.45	.01
	Round 2 vs. Round 3	3.68	.06	.03
	Round 3 vs. Round 4	0.02	.90	.00
	Round 4 vs. Round 5	4.01	.048	.03
	Round 5 vs. Round 6	1.24	.27	.01
	Round 6 vs. Round 7	0.23	.63	.00
	Round 7 vs. Round 8	2.10	.15	.02
	Round 8 vs. Round 9	5.81	.02	.05
	Round 9 vs. Round 10	3.16	.08	.03
	Round 10 vs. Round 11	1.40	.24	.01
	Round 11 vs. Round 12	0.18	.68	.00

Trust violation denotes the comparison of the mean value of the respective dependent variable of the trust violation phase compared to the initial phase. Trust repair denotes the comparison of the mean value of the respective dependent variable of the trust repair phase compared to the trust violation phase. $N = 121$. $n_{\text{no\_information\_human}} = 33$, $n_{\text{no\_information\_system}} = 35$, $n_{\text{imperfection\_human}} = 28$, $n_{\text{imperfection\_system}} = 25$

Hypothesis 2 suggested that after the trust repair intervention, trustworthiness, trust, and trust behavior in the automated system condition will increase less compared to the human trustee condition. Figures 2 and 3 and Tables 2 and 3 show that for trustworthiness and trust, this was the case. For trust behavior, the percentage of cases where participants accepted the human trustee's preselection in the trust repair phase was 73%, and for the automated system's preselection, 69% of cases were accepted showing no substantial differences. This supports Hypothesis 2 for trustworthiness and trust, but not for trust behavior.

Facets of Trustworthiness

RQ2 asked whether there is an initial difference and different effects for trust violations and trust repair interventions for the facets of trustworthiness regarding human and automated systems as trustees. Regarding initial assessments, ability was lower $F(1,117) = 31.35$, $p < 0.01$, $\eta_p^2 = 0.21$, and integrity was higher for the automated system $F(1,117) = 12.98$, $p < 0.01$, $\eta_p^2 = 0.10$, but there was no difference in benevolence $F(1,117) = 0.21$, $p = 0.89$, $\eta_p^2 = 0.00$. Regarding trust violations and trust repair interventions, effects were weaker for the automated system for ability and integrity (see Table 2; Fig. 2).

Effects of Information Regarding Imperfection

Hypothesis 3 proposed that trustworthiness, trust, and trust behavior in the automated system with information about imperfection condition will initially be lower compared to the condition without such information; however, for the first

task round, we found no effects of information regarding imperfection for trustworthiness, $F(1,117) = 0.11$, $p = 0.92$, $\eta_p^2 = 0.00$ or trust $F(1,117) = 1.38$, $p = 0.24$, $\eta_p^2 = 0.01$ and no interaction between the independent variables for trustworthiness, $F(1,117) = 0.10$, $p = 0.75$, $\eta_p^2 = 0.00$ or trust $F(1,117) = 0.05$, $p = 0.83$, $\eta_p^2 = 0.00$. For trust behavior, 93% of participants accepted the preselection from the human with no additional information (76% for the system), and 79% the preselection from the human when information about imperfection was present (86% for the system). If anything, trust behavior was thus more pronounced when information about imperfection was presented for the automated systems. Thus, Hypothesis 3 was not supported.

Hypothesis 4 proposed that following a trust violation, trustworthiness, trust, and trust behavior in the automated system with information about imperfection condition will decrease to a lesser extent compared to the condition without such information. This was not supported for trustworthiness and trust, as seen in Tables 2 and 3 (see Task round*trustee*imperfection) and Figs. 2 and 3. Instead, the effects of trust violations were stronger when information regarding imperfection was presented. Regarding trust behavior, in the initial phase, the percentage of cases where participants accepted the preselection of the automated system when information regarding imperfection was presented was 71%, and in the case of the automated system with no such information presented, it was 65%. In the trust violation phase, those acceptance rates were 29% (information regarding imperfection presented) and 28% (no such information presented) respectively. Those results indicated that there was no difference for trust behavior. Overall, these results did not support Hypothesis 4.

RQ3 asked whether there are interaction effects between the trust repair intervention and the information regarding imperfection for the automated system for trustworthiness, trust, and trust behavior. Indeed, trust repair interventions were more effective at restoring trustworthiness and trust for the automated system when information regarding imperfection was presented (see Task round*trustee*imperfection in Tables 2 and 3 as well as Figs. 2 and 3). Regarding trust behavior, the percentage of cases where participants accepted the preselection of the automated system when information regarding imperfection was presented was 74% compared to 62% when no such information was presented. We calculated χ^2-tests for the trust violation and the trust repair phase separately. For the trust violation phase, the difference between automated systems where information about imperfection was presented and where no such information was presented was not significant, $\chi^2(1) = 0.01$, $p = 0.92$, whereas there was a difference for the trust repair phase, $\chi^2(1) = 4.13$, $p < 0.05$. As a benchmark, for the human trustee, the percentages of acceptance in the trust violation phase were 28% (information about imperfection presented) and 26% (no such information presented), and 72% (information about imperfection presented) and 74% (no such information presented) in the trust repair phase which constituted no significant differences. The response to RQ3 therefore is: Trust repair interventions were more effective for automated systems as trustees when information regarding imperfection was presented.

Discussion

With the increasing use of AI-based automated systems in context where they contribute to decisions over individuals' fates, it becomes crucial to understand trust processes in human-AI collaboration for such decisions (Glikson & Woolley, 2020; Raisch & Krakoswki, 2021). Overall, our findings imply that the theoretical assumptions and effects found in classical application areas of automated systems only partly translate to novel application contexts of automated systems such as personnel selection. Specifically, in classical application contexts, where trustworthiness assessments mainly stem from system performance measures related to accuracy, people seem to expect near-to-perfection from automated systems, specifically high and consistent performance. Our results indicate that people do not expect high performance from automated systems in personnel selection as initial ability assessments were lower for automated systems. However, there seems to remain expectations of consistency: trust violations associated with potentially unfair bias and trust repair effects were weaker for automated systems. We may also tentatively conclude that these expectations of consistency were partly reduced by highlighting system imperfection, leading to stronger trust violation and trust repair effects. In sum, our study suggests that research assessing trust in automated systems must be aware of the application context in which systems support decision-making because although expectations of systems as being consistent might generalize, expectations of high performance might not. This seems to affect trust processes associated with automated systems in respective contexts.

Theoretical Implications

Trust in Automation Depends on the Use Context

Whereas in classical application contexts people expect systems to perform near-to-perfection (Madhavan & Wiegmann, 2007), our findings do not support this for personnel selection. Classical tasks for automated systems usually involve mechanical skills such as combining large amounts of data and performing repetitive tasks (M. K. Lee, 2018). In contrast, personnel selection requires individual and flexible decision-making capabilities, as well as ethical considerations — capabilities that people more likely ascribe to humans (Bigman & Gray, 2018; M. K.

Lee, 2018; Newman et al., 2020). In other words, people might believe that humans are better able to complete tasks where ethical issues and individuals' unique characteristics need to be considered (Longoni et al., 2019; Newman et al., 2020). In line with this and with further research (Höddinghaus et al., 2020; M. K. Lee, 2018; Rieger et al., 2022), our participants perceived the automated system to be less able but also perceived the integrity (associated with systems being less biased) of systems to be comparably stronger. Overall, these findings indicate that people have specific expectations of automated systems that affect their evaluations of trustworthiness of automated systems for different tasks (Elsbach & Stigliani, 2019). Those expectations seem to lead to high overall levels of ability assessments for classical automation tasks (e.g., monitoring) where mechanical skills are important and where primary performance measures are effectivity and efficiency. In contrast, those expectations may prompt comparably low initial ability assessments for tasks that involve consideration of ethical issues and decisions about individuals (e.g., personnel selection, performance evaluation; see also Nagtegaal, 2021).

In sum, this resulted in participants evaluating initial levels of trust to be stronger for human trustees. However, higher trust in the human trustee did not translate into effects on trust behavior. An explanation for this might be that differences in trust were not strong enough. Albeit we may conclude that trustors found trustees' ability to be more important than their integrity for their overall trust assessments, the different direction of effects for human and automated trustees for the facets of trustworthiness may have rendered trust differences too small to affect trust behavior.

Weaker Effects of Trust Violations and Repair Interventions for Automated Systems

Contrary to what could have been expected based on classical trust in automation research (Dzindolet et al., 2003; Madhavan & Wiegmann, 2007), our study showed comparably weaker negative reactions to trust violations by the automated system. This could be due to the fact that initial ability and trust perceptions of automated systems were comparably lower; thus, they could not suffer as much from the trust violation. However, after the trust violation, trustworthiness assessments for the human trustee fell below the level of the automated system as trustee which might imply stronger negative reactions to the ethical trust violation for human trustees.

Our findings regarding high integrity assessments for automated systems additionally support research indicating that people believe that systems are more consistent and less biased than humans (Langer & Landers, 2021). Beyond that, our findings imply that high integrity, consistency, or lack of bias that people ascribe to automated systems even holds when system outputs repeatedly indicate unfair biases. If ethical violations less strongly affect integrity assessments of automated systems, this suggests that people are less likely to realize such violations (Bonezzi & Ostinelli, 2021). Maybe people are less aware that system outputs can reflect unfair bias. Alternatively, people might interpret reasons for ethical violations differently compared to when they result from a human trustee. For instance, automated systems producing biased outputs might result in less negative perceptions because people ascribe more discrimination intention to humans (Bigman & Gray, 2018; Bigman et al., 2022). Beyond personnel selection, this could also apply to the use of automated systems in other management task, or to tasks in medicine and jurisdiction (Jago & Laurin, 2022). However, this interpretation and the fact that there were no significant effects on trust behavior highlight a need for future research because we can only tentatively conclude that expectations of consistency might mitigate negative reactions to ethical trust violations by automated systems (but see Bonezzi & Ostinelli, 2021; Jago & Laurin, 2022 who present similar results).

Although such trust violations caused stronger negative effects for human trustees, our participants were also more forgiving for human trustees. Plausibly, stronger effects of trust repair interventions could result from stronger trust violation effects for humans — there was simply more to gain by trust repair interventions. However, people may also expect humans to learn from mistakes (de Visser et al., 2018), whereas automated systems are deployed with attributes that cannot be easily changed (Madhavan & Wiegmann, 2007). Yet, although the trust repair intervention significantly affected reactions, thus having the intended effect, the results of the trust repair intervention should be interpreted cautiously. More precisely, it remains unclear what participants concluded when they received the information that "the error has been solved." For the human trustee, they might have interpreted that the human trustee received additional training. For the automated system, they might have interpreted that developers have solved the issue. Research investigating more specific trust repair interventions is needed, especially research that highlights causes for trust violations (Tomlinson & Mayer, 2009).

Highlighting Imperfection Reduces Expectations of Consistency?

When information regarding imperfection was presented for the automated system, trust violations effects and trust repair intervention effects were partly stronger. Stronger trust repair effects are in line with prior research (de Visser et al., 2016). However, instead of buffering trust violation effects, emphasizing imperfection led to stronger negative reactions. Apparently, making the imperfection of automated systems more

salient can influence what people expect from automated systems and how they perceive trust violations and trust repair interventions (Bahner et al., 2008; Dzindolet et al., 2003). In our case, highlighting imperfection might have reduced assumptions of consistency; thus, trust violations but also repair interventions had stronger effects because people were more likely to believe that system performance can vary. This tentative interpretation could stimulate future research that uncovers basic human attitudes and expectations regarding automated systems (Elsbach & Stigliani, 2019).

Main Practical Implications

Managers might draw different conclusions when systems compared to humans produce potentially biased outputs. Specifically, our results imply that managers may be less likely to believe that there is a problem with the system. In line with this, providers of AI-based personnel selection solutions commonly market their systems in a way that highlights the potential for less bias in personnel selection when organizations use their systems (Raghavan et al., 2020). If people believe that systems are less biased than humans, those marketing campaigns might convince organizations and managers to use such systems. However, evidence that such systems can prevent bias is still needed. It might thus be necessary to train decision-makers on possible challenges when relying on automated systems (Oswald et al., 2020) — this becomes crucial if future human decision-makers are responsible for overseeing AI-based systems in selection as proposed in current drafts for legislation on AI (e.g., the European AI Act[8]). For instance, it might be possible to make decision-makers aware of what they should scrutinize when using an AI-based automated system (Landers & Behrend, 2022): What constitutes the underlying training data? How did developers ensure that the system works as intended? Is there validity evidence and evidence regarding adverse impact?

Limitations and Future Research

In addition to the aforementioned lack of clarity of the trust repair intervention, there are further limitations to our study that may inspire research. First, our sample did not consist of hiring managers. Although hiring managers in organizations might currently not be better trained in working with automated systems than our sample (Oswald et al., 2020), we may expect stronger negative reactions to trust violations because they are more aware of serious consequences of unfair biases (e.g., lawsuits). Second, our sample consisted of predominantly female, German participants. Thus, although our results align with evidence that humans ascribe less bias to automated systems (Bonezzi & Ostinelli, 2021; Jago & Laurin, 2022), they may not generalize to more diverse samples or to samples from another cultural background. Third, we have only investigated reactions to a single automated system. There are many characteristics that differentiate systems (e.g., system performance, interface design) (Landers & Marin, 2021) and that affect trust in automated systems (J. D. Lee & See, 2004). We thus emphasize the opportunity to investigate psychological implications of system design on trust processes. Fourth, we have only investigated a single kind of implementation of systems into decision-making processes: A system that automatically processes information to provide outputs. Systems could also operate under human supervision or could closely collaborate with human decision-makers to provide decisions. The idea to investigate the implications of biased outputs produced by human-system teams may inspire future work. Similarly, future work may examine the attributed reasons for errors produced by humans versus automated systems, for instance, through the lens of attribution theory (Kelley & Michela, 1980). If people expect consistency from automated systems, it is possible that they more likely attribute trust violations by automated systems to stable characteristics of those system (e.g., programming errors) but human trust violations to situational influences (e.g., the person was stressed) (Madhavan & Wiegmann, 2007).

Conclusion

Regardless of whether organizations delegate tasks to automated systems or human beings, trust processes remain central (Glikson & Woolley, 2020). Our study shows that research on interpersonal trust (Mayer et al., 1995) and trust in automation (Hoff & Bashir, 2015; J. D. Lee & See, 2004) is valuable for research on automated systems for managerial purposes because the main concepts and drivers of trust dynamics remain similar. Additionally, similar to previous research (Madhavan & Wiegmann, 2007), having a human or a system as trustee seems to affect initial trust and moderate the effects of trust violations and trust repair interventions — but partly different than was expected. Thus, our results emphasize that we cannot assume that effects from classical application areas of automated systems, where trust depends mainly on effectiveness and efficiency, translate to the use of automated systems for decisions where trust also depends on ethical considerations. We hope that this study inspires future work investigating trust processes in the context of AI-based systems in ethically sensitive domains.

[8] https://eur-lex.europa.eu/legal-content/EN/TXT/HTML/?uri=CELEX:52021PC0206&from=EN

Supplementary Information The online version contains supplementary material available at https://doi.org/10.1007/s10869-022-09829-9.

Funding Open Access funding enabled and organized by Projekt DEAL. This study was pre-registered on AsPredicted (https://aspredicted.org/sj9ud.pdf). Work on this paper was funded by the project "Explainable Intelligent Systems" (Volkswagen Foundation grant number AZ. 98513) and by the project "Foundations of Perspicuous Software Systems" (DFG grant 389792660 as part of TRR 248).

References

Bahner, J. E., Elepfandt, M. F., & Manzey, D. (2008). Misuse of diagnostic aids in process control: The effects of automation misses on complacency and automation bias. *Proceedings of the Human Factors and Ergonomics Society, 52*, 1330–1334. https://doi.org/10.1177/154193120805201906

Bigman, Y. E., Wilson, D., Arnestad, M. N., Waytz, & Gray, K. (2022). Algorithmic discrimination causes less moral outrage than human discrimination. *Journal of Experimental Psychology: General*. Advance online publication. https://doi.org/10.1037/xge0001250

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

Bonaccio, S., & Dalal, R. S. (2006). Advice taking and decision-making: An integrative literature review, and implications for the organizational sciences. *Organizational Behavior and Human Decision Processes, 101*(2), 127–151. https://doi.org/10.1016/j.obhdp.2006.07.001

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

Cheng, M. M., & Hackett, R. D. (2021). A critical review of algorithms in HRM: Definition, theory, and practice. *Human Resource Management Review, 31*(1), 100698. https://doi.org/10.1016/j.hrmr.2019.100698

Dastin, J. (2018). *Amazon scraps secret AI recruiting tool that showed bias against women.* Reuters.Com. https://www.reuters.com/article/us-amazon-com-jobs-automation-insight/amazon-scraps-secret-ai-recruiting-tool-that-showed-bias-against-women-idUSKCN1MK08G

de Visser, E. J., Monfort, S. S., McKendrick, R., Smith, M. A. B., McKnight, P. E., Krueger, F., & Parasuraman, R. (2016). Almost human: Anthropomorphism increases trust resilience in cognitive agents. *Journal of Experimental Psychology: Applied, 22*(10), 331–349. https://doi.org/10.1037/xap0000092

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

Den Hartog, D. N., & De Hoogh, A. H. B. (2009). Empowering behaviour and leader fairness and integrity: Studying perceptions of ethical leader behaviour from a levels-of-analysis perspective. *European Journal of Work and Organizational Psychology, 18*(2), 199–230. https://doi.org/10.1080/13594320802362688

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

Dietvorst, B. J., Simmons, J. P., & Massey, C. (2015). Algorithm aversion: People erroneously avoid algorithms after seeing them err. *Journal of Experimental Psychology: General, 144*(1), 114–126. https://doi.org/10.1037/xge0000033

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

Eisinga, R., te Grotenhuis, M., & Pelzer, B. (2013). The reliability of a two-item scale: Pearson, Cronbach, or Spearman-Brown? *International Journal of Public Health, 58*(4), 637–642. https://doi.org/10.1007/s00038-012-0416-3

Elsbach, K. D., & Stigliani, I. (2019). New information technology and implicit bias. *Academy of Management Perspectives, 33*(2), 185–206. https://doi.org/10.5465/amp.2017.0079

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., Buchner, A., & Lang, A. G. (2009). Statistical power analyses using G*Power 3.1: Tests for correlation and regression analyses. *Behavior Research Methods, 41*(4), 1149–1160. https://doi.org/10.3758/brm.41.4.1149

Gilliland, S. W. (1993). The perceived fairness of selection systems: An organizational justice perspective. *Academy of Management Review, 18*(4), 694–734. https://doi.org/10.2307/258595

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

Grgić-Hlača, N., Engel, C., & Gummadi, K. P. (2019). Human decision making with machine assistance: An experiment on bailing and jailing. *Proceedings of the 2019 CSCW Conference on Human-Computer Interaction, 3*, 1–25. https://doi.org/10.1145/3359280

Grove, W. M., & Meehl, P. E. (1996). Comparative efficiency of informal (subjective, impressionistic) and formal (mechanical, algorithmic) prediction procedures: The clinical–statistical controversy. *Psychology, Public Policy, and Law, 2*(2), 293–323. https://doi.org/10.1037/1076-8971.2.2.293

Hickman, L., Bosch, N., Ng, V., Saef, R., Tay, L., & Woo, S. E. (2021). Automated video interview personality assessments: Reliability, validity, and generalizability investigations. *Journal of Applied Psychology*. Advance Online Publication. https://doi.org/10.1037/apl0000695

Höddinghaus, M., Sondern, D., & Hertel, G. (2020). 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, 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

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

Kelley, H. H., & Michela, J. L. (1980). Attribution theory and research. *Annual Review of Psychology, 31*, 457–501. https://doi.org/10.1146/annurev.ps.31.020180.002325

Kim, H., Dirks, K. T., Cooper, C. D., & Ferrin, D. L. (2006). When more blame is better than less: The implications of internal vs. External attributions for the repair of trust after a competence- vs. Integrity-based trust violation. *Organizational Behavior and Human Decision Processes, 99*(1), 49–65. https://doi.org/10.1016/j.obhdp.2005.07.002

Landers, R. N., & Behrend, T. S. (2022). Auditing the AI auditors: A framework for evaluating fairness and bias in high stakes AI predictive models. *American Psychologist.* Advance Online Publication. https://doi.org/10.1037/amp0000972

Landers, R. N., & Marin, S. (2021). Theory and technology in organizational psychology: A review of technology integration paradigms and their effects on the validity of theory. *Annual Review of Organizational Psychology and Organizational Behavior, 8*(1), 235–258. https://doi.org/10.1146/annurev-orgpsych-012420-060843

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

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

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

Lee, J. D., & See, K. A. (2004). Trust in automation: Designing for appropriate reliance. *Human Factors, 46*(1), 50–80. https://doi.org/10.1518/hfes.46.1.50.30392

Li, L., Lassiter, T., Oh, J., & Lee, M. K. (2021). Algorithmic hiring in practice: Recruiter and HR professional's perspectives on AI use in hiring. *Proceedings of the 2021 AIES Conference on AI, Ethics, and Society*, 166–176. https://doi.org/10.1145/3461702.3462531

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

Martin, K. (2019). Ethical implications and accountability of algorithms. *Journal of Business Ethics, 160*(4), 835–850. https://doi.org/10.1007/s10551-018-3921-3

Mayer, R. C., Davis, J. H., & Schoorman, F. D. (1995). An integrative model of organizational trust. *Academy of Management Review, 20*(2), 709–726. https://doi.org/10.2307/258792

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

Oswald, F. L., Behrend, T. S., Putka, D. J., & Sinar, E. (2020). Big data in industrial-organizational psychology and human resource management: Forward progress for organizational research and practice. *Annual Review of Organizational Psychology and Organizational Behavior, 7*(1), 505–533. https://doi.org/10.1146/annurev-orgpsych-032117-104553

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

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

Raghavan, M., Barocas, S., Kleinberg, J., & Levy, K. (2020). Mitigating bias in algorithmic hiring: Evaluating claims and practices. *Proceedings of the 2020 FAT* Conference on Fairness, Accountability, and Transparency*, 469–481. https://doi.org/10.1145/3351095.3372828

Raisch, S., & Krakoswki, S. (2021). Artificial intelligence and management: The automation-augmentation paradox. *Academy of Management Review, 46*(1), 192–210. https://doi.org/10.5465/amr.2018.0072

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

Rudnicka, J. (2020). *Anteil von Frauen und Männern in verschiedenen Berufsgruppen in Deutschland am 30. Juni 2019 [Proportion of women and men in different occupational groups in Germany on the 30th of June 2019.]*. Statista. https://de.statista.com/statistik/daten/studie/167555/umfrage/frauenanteil-in-verschiedenen-berufsgruppen-in-deutschland/

Tambe, P., Cappelli, P., & Yakubovich, V. (2019). Artificial intelligence in human resources management: Challenges and a path forward. *California Management Review, 61*(4), 15–42. https://doi.org/10.1177/0008125619867910

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

Tomlinson, E. C., & Mayer, R. C. (2009). The role of causal attribution dimensions in trust repair. *Academy of Management Review, 34*(1), 85–104. https://doi.org/10.5465/amr.2009.35713291

Wang, W., & Benbasat, I. (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