Int. J. Human–Computer Studies 211 (2026) 103775

Contents lists available at ScienceDirect

International Journal of Human - Computer Studies

journal homepage: www.elsevier.com/locate/ijhcs

Trust the Explanation or my Expectation? Effects of Output Accuracy and Explanations on Expectation Violations and Trust in AI-Supported Decisions

Tim Hunsicker [a,*], Isabel Duhl [a], Pascal Haubert [a], Linda Onnasch [b], Markus Langer [c]

[a] Fachrichtung Psychologie, Universität des Saarlandes, Germany
[b] Institut für Psychologie und Arbeitswissenschaft, Technische Universität Berlin, Germany
[c] Institut für Psychologie, Universität Freiburg, Germany

ARTICLE INFO

Keywords:
Artificial intelligence
Algorithmic decision-making
Explanations
Trust
System accuracy
Expectation violation

ABSTRACT

Systems based on Artificial Intelligence (AI) increasingly support decision-making, but their outputs may be inaccurate. Prior research has suggested that explanations might help detect inaccuracies, aiding successful human-AI interaction. This study investigates how the accuracy of system outputs influences users' trust, trusting behavior, and trustworthiness perceptions, the role of expectation violations in this process, and how explanations for the system outputs influence these effects. In an online study with a 2(explanation vs. no explanation) × 2(accurate vs. inaccurate outputs) between-within design, 218 participants evaluated six job applicants. They received CVs and algorithmic evaluations of applicants' suitability. For three applicants, outputs were accurate; for the other three, outputs reflected a 40% lower suitability than their true suitability. Half of the participants received explanations. Accurate outputs led to higher trustworthiness, trust, and trusting behavior than inaccurate outputs. Expectation violation fully mediated how accuracy affected trust and trustworthiness, and partially how accuracy influenced trusting behavior. Moreover, there was a significant interaction between explanations and output accuracy concerning trusting behavior: when outputs were accurate, explanations had little effect on trusting behavior; however, when outputs were inaccurate, explanations led to stronger trusting behavior, as participants less strongly deviated from the inaccurate outputs. We conclude that users are able to deviate from inaccurate outputs, and we highlight the importance of expectation violations in this regard. However, our findings also show possible detrimental effects of explanations as they can increase the decisional weight of inaccurate outputs instead of facilitating the detection of inaccuracies.

1. Introduction

Artificial intelligence (AI) is reshaping decision-making processes across numerous high-stakes domains, leveraging a diverse array of methodologies from classical algorithms to neural networks and sophisticated deep learning models (LeCun et al., 2015). In contexts where decisions have far-reaching consequences (e.g., medicine, personnel selection; Sterz et al., 2024), due to ethical, legal, and safety reasons, AI-based systems are mainly used as decision support. The final decision then remains with a human decision-maker (Enarsson et al., 2022), who must weigh how much they trust and follow the system for their final decision (e.g., Schemmer et al., 2022). This would be easy in a world with perfectly accurate AI-based systems. In reality, however, it is challenging for decision-makers to distinguish between accurate and inaccurate outputs (Green, 2022).

In line with theoretical foundations in the trust literature, higher system accuracy should lead to higher perceived trustworthiness, trust, and trusting behavior, whereas inaccuracies should decrease these trust-related outcomes (e.g., Hoff & Bashir, 2015; Schlicker et al., 2025). Expectations should play a crucial role in this process as decision-makers expect accurate outputs from systems (Lee & See, 2004; Mayer et al., 1995). We propose that expectation violations drive this process: when decision-makers suspect inaccuracies, this should violate their expectations, which in turn will lead them to reject system recommendations or deviate from system outputs. In this regard, explanations for system outputs may help decision-makers as they could facilitate realizing that outputs are inaccurate (Rader et al., 2018; Springer & Whittaker, 2020). In particular, local explanations that offer insights into individual predictions may facilitate expectation violations by highlighting inaccurate system outputs (Ribeiro et al., 2016). However, whether and why local

* Corresponding author at: Universität des Saarlandes, Campus A1 3, 66123, Saarbrücken, Germany.
E-mail address: tim.hunsicker@uni-saarland.de (T. Hunsicker).

https://doi.org/10.1016/j.ijhcs.2026.103775
Received 13 December 2024; Received in revised form 12 November 2025; Accepted 22 January 2026
Available online 25 February 2026

explanations can help to distinguish accurate from inaccurate outputs has not been sufficiently investigated. Whereas some studies show that they contribute to the detection of inaccurate outputs (e.g., Ribeiro et al., 2016; Wang & Yin, 2022), others indicate that they may lead to overreliance (e.g., Bansal et al., 2021; Buçinca et al., 2021).

In order to enhance our understanding of effective human-AI interaction and to explore why explanations can have both beneficial and detrimental effects on user trust and decision accuracy (Lai et al., 2023), the present study has three primary objectives. (1) To determine the effects of AI system output accuracy on users' perceived system competence, their trust in the system, and their trusting behavior. (2) To investigate the role of expectation violation as a key mediating mechanism in the relationship between AI output accuracy and the aforementioned trust-related outcomes. (3) To explore how the provision of local explanations for AI outputs interacts with output accuracy to influence these processes.

To achieve these objectives, we employed a 2×2 between-within design where participants judged the suitability of several job applicants and received decision support from an AI-based system. We employed a decision paradigm that allowed us to assess user behavior from two critical angles: how strongly decision-makers deviate from a ground truth and how strongly they deviate from the system's outputs. While much research in human-AI decision-making has focused on dichotomous choices (e.g., users fully accepting or rejecting system outputs; Lai et al., 2023), real-world scenarios often involve users making partial adjustments to system outputs (Bonaccio & Dalal, 2006). Understanding under what conditions users might partly integrate or modify AI recommendations is crucial and reflects a large but understudied share of AI-assisted decision-making practices. We investigated how the accuracy of system outputs and explanations for these outputs affect participants' perceived trustworthiness of the system (i.e., competence of the system), as well as trust and trusting behavior toward the system (operationalized as the absolute deviation of the numeric evaluation of the applicant by the decision-maker from the numeric value provided by the system). Furthermore, we examined the concept of expectation violation as a possible underlying psychological process explaining the intertwined effects of output accuracy and explanations on trustworthiness, trust, and trusting behavior.

Overall, we see the following contributions of this paper. First, we delineate the role of expectation violations as a crucial psychological process in determining trust-related outcomes when decision-makers face inaccurate outputs of AI-based systems. Second, our findings indicate that the explanations in our study did not foster expectation violations and instead increased the decisional weight of inaccurate outputs, providing additional support for the potentially detrimental effects that explanations can have (e.g., Bansal et al., 2021; Cecil et al., 2024). Consequently, while fostering expectation violations can be a key process for decision-makers to detect inaccurate outputs, and explanations in theory could foster such expectation violations, explanations can also convince decision-makers of the quality of system outputs, lead to subjective understanding, and may thus make overtrusting inaccurate outputs more likely.

2. Background and Hypotheses Development

2.1. The Role of Accuracy of Outputs for Trust in Systems

When an AI-based system is used as a decision support tool, the final decision remains with the human (Köchling et al., 2023; Lai et al., 2023). The users, in this case, the decision-makers, must decide how much weight they put on the system's output when they incorporate it into their final decision. This decision can be influenced by many factors, such as the accuracy of outputs, cognitive processes, and attitudes toward the system (Hoff & Bashir, 2015; Schaefer et al., 2016).

The accuracy of system outputs may vary. Poor accuracy may, for example, stem from being trained on unsuitable data or a criterion for prediction not clearly defined (Adomavicius & Zhang, 2012). No matter the cause for imperfect accuracy, without a human decision-maker who effectively oversees the system, those inaccuracies may proliferate to affect decision-making processes negatively. Inaccuracies mean that outputs are inconsistent with the true state of the world (e.g., a candidate is suitable for a job, but a system presents a low suitability score). To realize inaccuracies, decision-makers need to have access to information other than the system output (e.g., raw data) and need to search for evidence of output (in)accuracy (Parasuraman & Manzey, 2010).

To understand the possible consequences of inaccurate system outputs, it is important to investigate the effects of inaccuracies on the behavior of decision-makers. In this regard, trust in the system and its outputs is crucial. Trustworthiness, trust, and trusting behavior are key concepts of trust processes. These concepts are part of both theoretical models of interpersonal trust (Mayer et al., 1995) and trust in automation (Lee & See, 2004). Trustworthiness reflects the trustor's perception of the trustee's characteristics and performance in a task (Lee & See, 2004; Mayer et al., 1995). Trustworthiness usually consists of different facets, such as the trustees' ability or competence, integrity, and benevolence (Mayer et al., 1995; Schlicker & Langer, 2021). In the context of automated systems, the ability of the system is particularly important. For example, in the case of an AI-based system for personnel selection purposes, its ability would involve whether decision-makers perceive that a system is able to produce an accurate evaluation of applicants, that is, the perceived competence. Trust can be considered an overall intention toward a system and its outputs. It is assumed to be fueled by the assessment of the system's trustworthiness as well as decision-makers' dispositions to trust systems (Hoff & Bashir, 2015; Mayer et al., 1995). Trusting behavior is the behavioral trust-related outcome (Kohn et al., 2021). For example, trusting behavior could be that a decision-maker actually delegates a decision to a system. It is also reflected in the weight individuals assign to system outputs when making decisions.

The goal of interaction between decision-makers and AI-based systems is neither a generally high nor a generally low level of trust but an appropriate level of trust (de Visser et al., 2020; F. Yang et al., 2020; Zhang et al., 2020). Although there is an ongoing discussion about the exact nature of the concept of appropriate, calibrated, or adequate trust, one common theme is that decision-makers follow accurate outputs and detect and deviate from inaccurate outputs (Schlicker et al., 2025; Wischnewski et al., 2023).

2.2. Expectations and Trust

Most conceptualizations of trust include notions that refer to people's expectations toward a trustee (McKnight et al., 2011; Rousseau et al., 1998). In fact, expectations are an integral part of one of the most common definitions of trust by Mayer et al. (1995): trust is "the willingness of a party to be vulnerable to the actions of another party based on *the expectation* [emphasis added] that the other will perform a particular action important to the trustor, irrespective of the ability to monitor or control that other party" (p. 712). For example, in the context of personnel selection, a decision-maker expects a system to produce accurate evaluations of applicants. This expectation may be fueled by the fact that a decision-maker expects a system to have a high ability to produce accurate outputs.

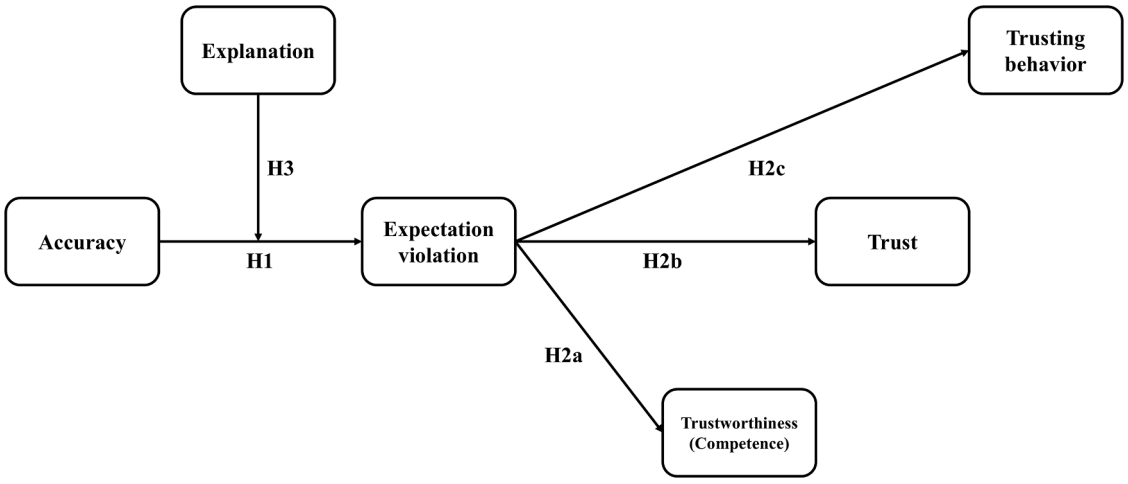


Figure 1. Overview of the Framework and the Hypotheses of the Study.

Although expectations play a central role for trust, and examining (the violation of) expectations may offer insight into the psychological processes involved in trust development, empirical research has paid little attention to the concept of expectations. This study aims to address and investigate this gap, as shown in Figure 1.

We hypothesize that inaccurate system outputs will lead to expectation violations in decision-makers.[1] To make a judgment, decision-makers must integrate different evidence relevant to the task (e.g., evidence found in CVs of applicants; Fischhoff & Broomell, 2020). Suppose we assume that there is a sufficiently able decision-maker who has access to relevant information beyond the system output, and they have at least partly or initially analyzed this available information for themselves. In that case, accurate outputs may be approximately consistent with the judgment of the decision-maker. In this case, there should be an expectation fulfillment because the decision-maker may assume that the system output (i.e., the system's evaluation of an applicant's suitability) and the state of the world as the decision-maker has judged it (i.e., the decision-maker's evaluation of the applicant's suitability) align. Likewise, expectations may also be fulfilled if the decision-maker first sees an output of the system and then assesses its validity to see how well the system works (e.g., by analyzing additional available information). In this case, initial expectations of the system's capabilities are met after assessing its validity. In contrast, inaccurate outputs may deviate from the decision-makers' assessments. This should lead to expectation violations because the outputs do not match the decision-maker's judgment of the state of the world. Specifically, this could mean that the decision-maker compares available raw data with the system's interpretation of this data as reflected in its outputs and then wonders why there is a mismatch. As a consequence of an expectation violation, decision-makers may notice the inaccuracy of an output. Accordingly, we hypothesize that:

Hypothesis 1. Inaccurate outputs lead to higher expectation violations than accurate outputs.

Expectations (and whether they are fulfilled or violated) are a central prerequisite for trust. If expectations are violated (i.e., due to poor accuracy), this should affect trust-related outcomes. Specifically, an expectation violation shows decision-makers that a system output is inconsistent with their own assessment of a situation, which may reduce the perceived trustworthiness, trust, and trusting behavior toward the system.[2] To date, little research can be interpreted as empirical support for this assumption. However, in the context of peer assessments (Kizilcec, 2016) or credit scoring (de Zoeten et al., 2023), violated expectations (i.e., lower grades than expected or credit not granted) led to lower trust in the system. To shed light on the role of expectations in the trust process, we manipulate the accuracy of the system and measure expectation violations. In the context of our study, we propose that expectation violations mediate the relationship between accuracy and trust-related outcomes.

Hypothesis 2. Expectation violation mediates the relationship between accuracy and a) perceived competence, b) trust in the system, and c) deviation from system output.

2.3. *Explanations May Foster Expectation Violations and Facilitate Identifying Inaccuracies*

The way AI systems work is often not transparent (Adadi & Berrada, 2018; Miller, 2019). Especially with models based on Machine Learning, it can be difficult for decision-makers to understand the relation between inputs and outputs (Burkart & Huber, 2021). This challenge is amplified by the increasing complexity of modern AI, such as deep learning architectures and Large Language Models (LLMs), whose internal logic is inherently opaque. This makes it difficult for decision-makers to evaluate the accuracy of a system's outputs (Zerilli et al., 2019). Eventually, this makes expectation violations less likely because decision-makers have little information to realize that a system output may be inaccurate. This is especially true when they receive no further information or explanations regarding a system's outputs.

While some argue for the inherent value of using interpretable models rather than explainability approaches in high-stakes contexts

[1] Initially, we proposed (and preregistered under https://aspredicted.org/VFS_TTK) that perceived competence mediates the effect of accuracy on trust and trusting behavior, explanations moderate the path from accuracy to perceived competence, and accuracy has an effect on expectation violations. However, during writing this paper, we realized that expectation violation theoretically lies between accuracy and the trust-related outcomes (such as perceived competence). We thus concluded that what is now depicted in Figure 1 seems to be the conceptually more valid line of reasoning, particularly also in terms of how and where explanations may act as a moderator. This is why we partly deviate from the preregistration in a way that we now put expectation violations as the central mediator between accuracy and the trust-related variables.

[2] In the context of this study, we refer to negative expectation violations (i.e., a system performs worse than expected). Positive expectation violations (e.g., a system detects a suitable applicant that a decision-maker would not have detected) are not the subject of this study, but we acknowledge that they are possible and could lead to higher perceived trustworthiness, trust, and trusting behavior.

(Rudin, 2019), the prevalence of opaque systems necessitates methods to explain their behavior. Local explanations aim to provide decision-makers with information about the rationale behind a system's output for a specific instance or prediction (Langer et al., 2021; Ribeiro et al., 2016). If local explanations help decision-makers to understand the rationale behind a system's prediction, they may more likely be able to assess the accuracy of that output. Whereas this may sound intuitive, there is limited research regarding the psychological processes that help understand why explanations may have beneficial or detrimental effects in human-AI decision-making (Lai et al., 2023). Initial research indicates that explanations may lead to a higher subjective understanding of algorithmic functioning (Rader et al., 2018; Wang & Yin, 2022) and may help people to recognize whether system outputs are accurate (Rader et al., 2018).

We claim that one key process by which explanations could affect decision-makers is by fostering expectation violations for inaccurate outputs. For instance, if an explanation helps users identify inconsistencies between raw data (e.g., a CV) and the system's output, this might trigger an expectation violation. Such a violation, in turn, could motivate users to more closely scrutinize the AI-based output rather than accepting it without deeper cognitive engagement. Well-designed explanations may then help decision-makers to determine what additional information to scrutinize when checking the accuracy of the outputs (i.e., to know which raw data is worth looking at) because the explanations may provide insight into what evidence was particularly important for the output. For inaccurate outputs, checking the respective additional information and finding that the system may have incorrectly used this information can make decision-makers realize that system outputs are inaccurate, thus reducing trust-related outcomes.

Hypothesis 3. Explanations alter the strength of the mediated relationship between output accuracy, expectation violations, and the trust-related outcomes. Specifically, we propose that explanations will moderate the indirect effect of output accuracy via expectation violation on a) perceived competence, b) trust, and c) deviation from system output. Explanations will make expectation violations more salient when output accuracy is low, thus strengthening the link between accuracy and expectation violation.

3. Method

3.1. Sample

Our a priori power analysis, conducted using G*Power3 (Faul et al., 2007), indicated that $N = 162$ participants were needed to detect a small effect (partial $\eta^2 = .01$) for a within-between interaction in a repeated measures Analysis of Variance (ANOVA) framework (two groups, two measurement points, power = .80, $\alpha = .05$). Regarding the expected mediation effects, according to Fritz and MacKinnon (2007), a sample of $N = 162$ is also needed for percentile bootstrapping with a power of .80, an alpha level of .05, and a small to medium effect on the alpha and beta paths of a mediation. Because of possible technical problems in online studies and given that we also hypothesized a moderated mediation effect, we wanted to gather data from at least 162 participants, but were open to collecting data from as many participants as possible during two months.

We collected the data for this study online via SoSci Survey (Leiner, 2022) in August and September 2022. The study was targeted at people with an interest in human resource management and personnel selection. The requirements for participation were a minimum age of 18 years and good German language skills. Psychology students received course credit for their participation. The mean completion time was 12.86 minutes. In total, $N = 237$ participants completed the study. In line with our preregistered exclusion criteria, three participants were excluded because they completed the study in less than four minutes, indicating inattentive responding.[3] Six participants were excluded because they stated that their data should not be used, and ten participants were excluded because they failed the attention check regarding the explanation. Thus, the final sample was $N = 218$.

The sample included 75 (34.40%) male, 139 (63.76%) female, one (0.46%) participant who indicated their gender as "diverse", and four (1.38%) who did not select their gender. Participants' age ranged from 18 to 67 years ($M = 32.97$, $SD = 13.35$). In total, $n = 43$ participants (19.72%) were psychology students, and $n = 59$ participants (27.06%) indicated they were studying another field. Most participants, $n = 167$ (76.61%), were employed at the time of the survey. Regarding their previous experience in personnel selection, $n = 183$ participants (83.94%) had previous experience as applicants, and $n = 88$ (40.37%) stated experience as recruiters.[4]

3.2. Procedure

The study has been considered a low-risk study and was thus exempt from ethical approval according to the regulation of the main author's local ethical review board. The study also complies with the provisions of the European General Data Protection Regulation. The present study followed a 2 × 2 mixed design with a between (explanation: no explanation vs. local explanation) and a within factor (output accuracy: accurate vs. inaccurate output). For the within factor, all accurate recommendations were presented in one block, and all inaccurate recommendations were presented in one block. To prevent order effects, the order of the blocks was randomized between participants. We decided on this block-wise presentation of accurate and inaccurate recommendations because we also wanted to capture participants' perceptions of the accurate vs. inaccurate version of the system comprehensively after the respective blocks. Figure 2 shows a flow chart summarizing the study procedure.

Participants were randomly assigned to their experimental conditions. The welcome page of the survey showed the data privacy statement and further information about the study. After giving their informed consent, participants were directed to the instructions. They were informed that they would evaluate six applicants regarding their suitability for a position as an insurance and finance clerk, and an algorithm would assist them. In the condition with local explanations, participants were also informed that they would see an explanation for each system output. On the next page, participants were instructed to download the job ad.

Participants were then randomly presented with the first block of the accuracy condition (accurate vs. inaccurate). This block consisted of three consecutive pages, each containing a CV of an applicant. At the top of each page was a fold-out tab under which participants could also access the job ad (see Figure A1 in Appendix A). Below the tab, we presented the resume of the respective application, including fictitious contact details and information on education, school career, professional career, IT and language skills, as well as honorary posts and personal interests (see Figure B1-Figure B6 in Appendix B). On the same page, in the explanation condition, the respective graphical explanation followed (see Figure C1-Figure C6 in Appendix C). Then, the system's output regarding the suitability of the candidate for the job was presented (from 0%–100% suitable; this assessment was accurate or

[3] See the preregistration under https://aspredicted.org/VFS_TTK.

[4] We conducted exploratory analyses controlling for participants' recruiter experience. Our main mediation and two-way interaction results remained robust. Further 3-way moderation tests (Accuracy × Explanation × Experience), corrected for multiple comparisons, were not significant, confirming our findings remain robust even when controlling for participants' recruiter experience.

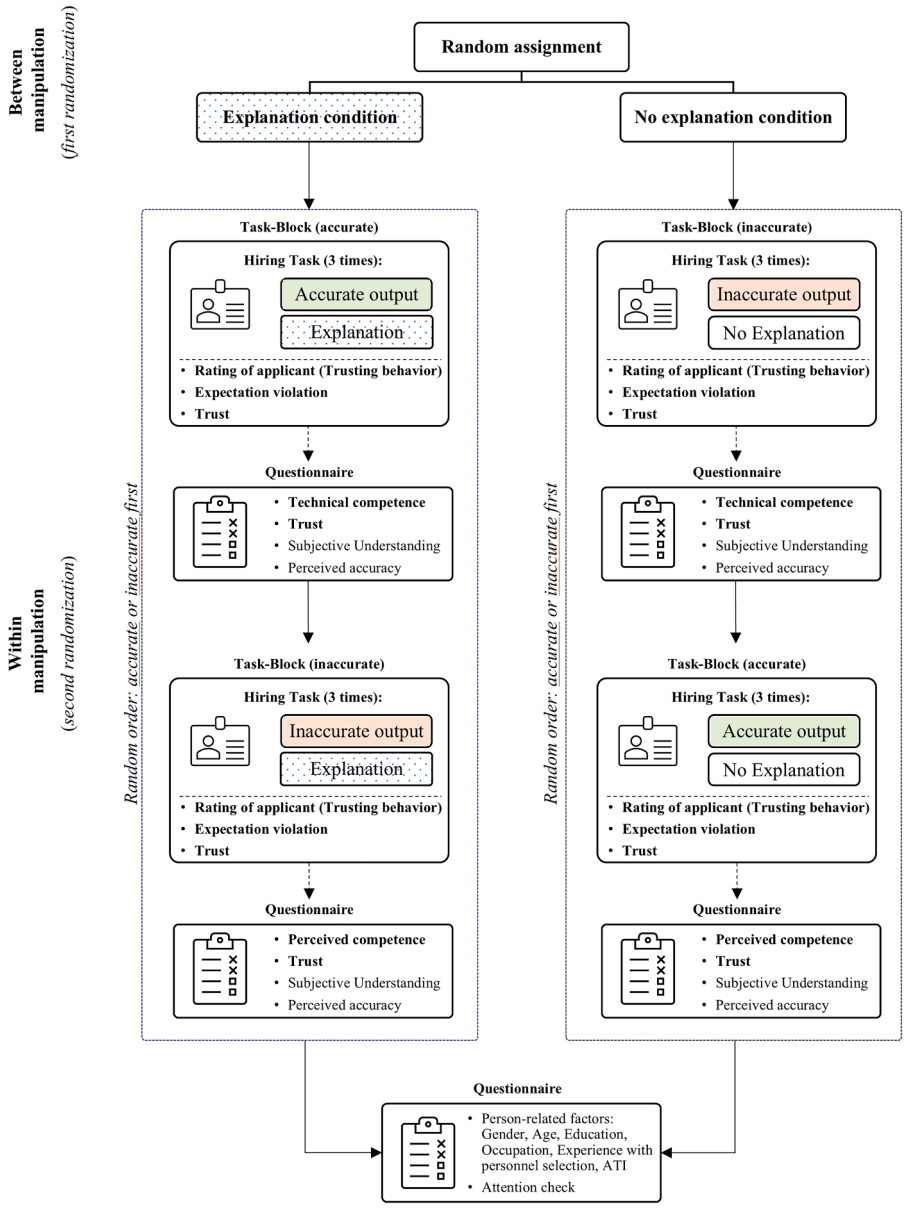


Figure 2. Flow Chart of the Study Procedure.

Note. The flowchart illustrates the experimental design. Participants were first randomly assigned to an explanation condition. Subsequently, all participants completed two task blocks (accurate vs. inaccurate outputs) in a randomized, counterbalanced order, represented by the two parallel paths. Each block consisted of three hiring tasks and a subsequent questionnaire. After completing both blocks, all participants proceeded to a final questionnaire. Colors are a visual aid for the conditions described in the text. Bullet points indicate the measured variables; measures relevant to the hypotheses are **bolded**.

inaccurate depending on the condition). Participants were then asked to provide their own rating of the applicant. After doing this, they responded to items on expectation violation as well as trust in the system output.

After the first block, participants received several scales with items on perceived accuracy (as a manipulation check), perceived competence, trust in the system, and further exploratory variables (see measures section). After the scales, the second block started, including the second accuracy condition. This block was analogous to the first block; participants received CVs of three applicants and provided their assessment of these applicants. After block two, participants responded to the same scales as after block one. On the following page, there was an attention check: based on an exemplary screenshot, participants had to state whether they had received graphical explanations during the study.

Then, we captured participants' Affinity for Technology Interaction (ATI; Franke et al., 2019) and further exploratory variables (see measures section). The demographic questionnaire asked for information on gender, age, studies, occupation, and experience with personnel selection. Participants were then asked whether they had conscientiously completed the questionnaire and whether their data could be used. Finally, participants were debriefed, received information on how to obtain their compensation, and had the opportunity to give feedback on the study.

3.3. Development of the Materials

3.3.1. Pilot Studies for the Ground Truth and the Manipulation

We conducted a pilot study to define a ground truth of the fit between the job and the respective applications. We presented twelve applications and the advertised position to $N = 6$ graduate students specializing in industrial and organizational psychology. The students

had to assess the applications and determine their suitability for the position. To achieve average suitability while maintaining a variance in the numbers between candidates, we selected six applicants who were closest to an average value and had the lowest standard deviation, which led to the following six applicants for the final study (see Appendix B): one application with a fit of 50%, four applications with a fit of 60% and one application with a fit of 70%. Afterward, we conducted a second pilot study ($N = 16$) regarding the extent of inaccuracy that may be noticeable to participants. Specifically, participants saw three applicants with accurate system outputs (control group) or inaccurate system outputs (20%, 30%, and 40% downward bias) and responded to the dependent variable expectation violation. In line with the findings of this pilot, we decided on a downward bias of the system in the inaccurate condition by 40%.[5] This substantial and consistent downward bias was selected to simulate real-world scenarios where systemic issues, such as errors in data parsing or biases in the training data, can result in applicants being assessed incorrectly and significantly lower than their true suitability.

3.3.2. Development of the Local Explanations

For the between factor, one group received no explanations, and the other received local explanations of the system output. We chose a graphical presentation method based on the local explanations by Wang and Yin (2022). In their study, respective graphical explanations led to an increased subjective understanding of system outputs. The present study used bar graphs that included the positive and negative influences of each feature for a single system output. The explanations included information on five features highlighted as relevant in the job description for a good fit: work experience, education type, final grade, MS Office skills, and English skills. All this information was reflected in applicants' CVs that included additional information as distractors. The graphical explanations plot the relevant features on a scale from -2 to +2, based on the ground truth for the respective CV. Figure 3 shows an example of an inaccurate output with an explanation and Figure 4 an example of an accurate output with an explanation. The local explanations for the inaccurate condition were downward-biased, just like the system outputs: the bars of each feature were biased toward the negative range by one to two scale points. For inaccurate outputs, the explanations were intended to help participants realize that the output of the system cannot be correct: Comparing the graphical explanation with the original information from the CV should lead to the insight that the applicant deserves better job suitability values.

3.4. Measures

Unless otherwise stated, the original versions of the items were in German, or we used translated adaptations of the scales. The full list of items for the key scales can be found in Table D1 in Appendix D. Unless otherwise stated, participants responded to all items measuring the dependent variable on a scale from 1 (*strongly disagree*) to 5 (*strongly agree*).

3.4.1. Trust

We used one item by Rieger et al. (2022) to capture participants' trust in the algorithm for each decision which was presented after each assessment of an applicant (i.e., 6 times): "How much do you trust the algorithm?" on a five-point Likert scale (1 = *completely*, 5 = *not at all*), which we recoded for the further analysis such that higher value indicate higher trust. In addition, we used five items by Madsen & Gregor (2000) to assess trust (sample item: "When I am uncertain about a decision I believe the system rather than myself.") once after each block. For our analyses, we decided to only report results for the one-item measure captured after each decision, given that the results for the blockwise scale measure did not differ from the one-item measure.

3.4.2. Trusting Behavior

The participants had to rate each applicant in terms of suitability for the position. Participants received the following statement "This applicant fits ... percent to the advertised position." and could select a value from 0 to 100 using a slider. We decided to use the absolute deviation from the system output as the value for the strength of the trusting behavior (a lower deviation indicates more trusting behavior):

$$\text{Deviation} = |(\text{system's output value}) - (\text{participant's rating})| \quad (1)$$

3.4.3. Perceived Competence

To capture perceived competence, we used Madsen & Gregor's (2000) "Perceived Technical Competence" scale (sample item: "The system uses appropriate methods to reach decisions."). We note that the original scale items use the term decisions, which in the context of our study corresponds to the system's outputs or predictions. Participants responded to this four-item scale once after each of the two experimental blocks (i.e., twice in total), as this measure, along with perceived accuracy and subjective understanding, reflects a more global assessment of the system's performance within each condition.

3.4.4. Expectation Violation

We captured the expectation violation of the presented system output with one item adapted from Burgoon and Walther's (1990) "Expectedness" scale. Participants received the statement "The algorithm returned a value that I expected." They responded to this on a five-point Likert scale (1 = *strongly agree*, 5 = *strongly disagree*), thus higher values indicate more expectation violation. We collected expectation violation together with trust and trusting behavior after each applicant assessment (i.e., six times per participant) in order to capture participants' immediate, instance-specific reactions to single system outputs and explanations (in the explanation condition).

3.4.5. Affinity for Technology Interaction (ATI)

We used the short version (ATI-S) of the scale, which consists of four items (sample item: "I like to occupy myself in greater detail with technical systems."; Franke et al., 2019). We utilized the original scale ranging from 1 (*not at all true*) to 6 (*completely true*).

3.4.6. Experience in Personnel Selection

As part of the demographic data, we collected the participants' experience in personnel selection. They were asked about their experience in applying for jobs ("I have experience as an applicant (have sent out CVs, gone through an application process).") and in selecting applicants ("I have experience in selecting applicants.").[6]

3.5. Attention Checks and Manipulation Check

To ensure that participants saw the explanations, we included an attention check. For this purpose, we presented all participants with an exemplary screenshot of an explanation. There, participants had to answer whether they had received such explanations (response options: *no* vs. *yes*).

[5] Additional data and results of the pilot studies are available upon request.

[6] For exploratory purposes, in the end of the study we captured *Perceived general competence* ("How competent do you rate the algorithm overall in this task?" and "How competent do you rate yourself overall in this task?"; 0 = *Not at all,* 11 = *Fully*), *Responsibility* ("Who is responsible for performance in the jointly processed task?" and "Who is responsible for consequences resulting from the jointly processed task?"; 0 = *Me*, 11 = *The algorithm*) and *Role perception.* ("[…] How involved were you in the decision-making process?" and "[…] How involved were you in the information gathering/analysis?"; 0 = *Not at all involved*, 11 = *Fully involved*), by Meussling et al. (2022).

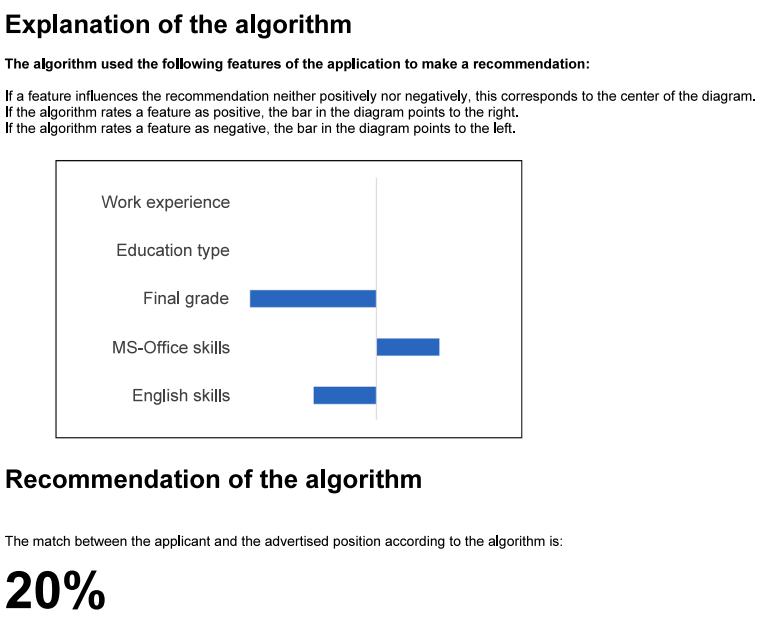


Figure 3. Sample Explanation With Output (Inaccurate) for Applicant Number 4.
Note. All participants received the output (accurate or inaccurate) of the system, showing the suitability of the candidate for the job, and the participants in the explanation condition were also presented with respective local graphical explanations indicating the importance of different features. In this example, participants see that the final grade had a strong negative influence, the English skills a negative influence, and the MS-Office skills a slight positive influence on the recommendation. However, if the participants compared this with the CV of the applicant, they would realize that the final grade is above average, that the applicant is fluent in English, an expert in MS Office, and already has 2 years of professional experience. All of this should actually be positive (or at least not strongly negative) for the evaluation of the applicant. Materials translated from German.

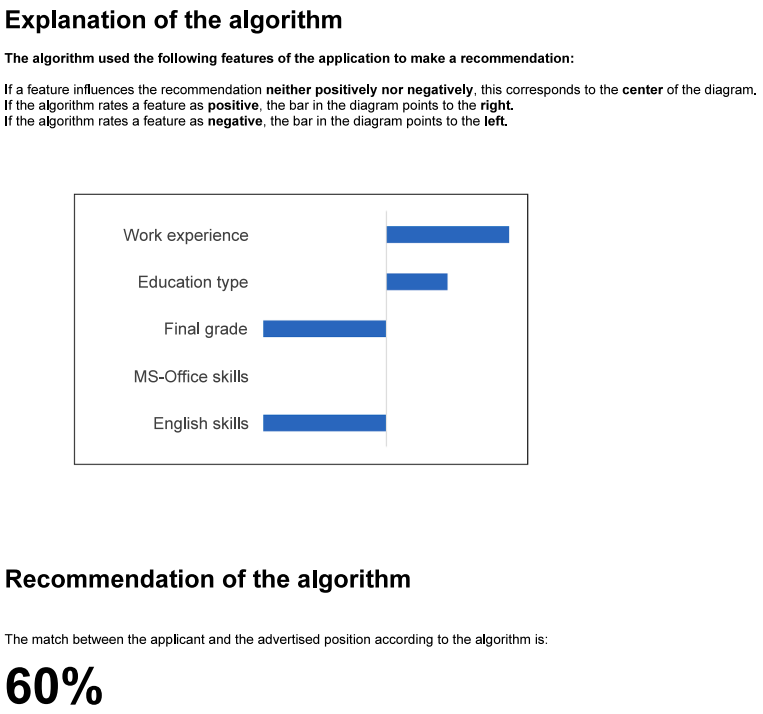


Figure 4. Sample Explanation With Output (Accurate) for Applicant Number 2.
Note. All participants received the output (accurate or inaccurate) of the system, showing the suitability of the candidate for the job. The participants in the explanation condition also received graphical local explanations indicating the importance of different features. Materials translated from German.

3.5.1. Perceived Accuracy

To check that the manipulation of the accuracy of the outputs worked as intended, perceived accuracy was measured using an item from the accuracy scale of Shin et al. (2020). Participants rated the statement "I believe the outputs made by the algorithm are accurate." We collected this item once after each block (i.e., twice in total).

3.5.2. Subjective Understanding

Because explanations should lead to higher perceived understanding, we captured subjective understanding with two items as a manipulation check for the explanation condition. These were translated and adapted from Wang and Yin (2022), with a sample item being "I understand how the algorithm works." Participants responded to these items after each block (i.e., twice in total).

3.6. Data Analysis Strategy

All data were analyzed using R (Version 4.3.3) with a significance level of $\alpha = .05$. As a preliminary step, we conducted manipulation checks to confirm the effectiveness of our experimental manipulations. We used two-way mixed analyses of variance (ANOVAs) for this purpose, as this test is well-suited for assessing how a between-participant factor (explanation) and a within-participant factor (accuracy) influence our manipulation check variables (subjective understanding and perceived accuracy).

To test our primary hypotheses, we employed structural equation modeling (SEM) using the lavaan package in R (Rosseel, 2012). This approach is ideal for testing path models involving direct and indirect effects. For Hypotheses 1 and 2, we specified a mediation model. This analysis allows us to test whether the effect of our predictor (output accuracy) on an outcome (e.g., trust) is mediated by expectation violation. The path from output accuracy to expectation violation tests Hypothesis 1, while the significance of the overall indirect effect, assessed via 5,000 bootstrapped samples, tests the mediation proposed in Hypothesis 2 for each trust-related outcome. To test Hypothesis 3, we specified a moderated mediation model. This approach allowed us to examine whether the strength of the aforementioned indirect effect was conditional on our moderator, the presence of an explanation.

Finally, to complement our hypothesis testing and explore the overall pattern of results, we conducted exploratory two-way mixed ANOVAs on our main dependent variables. This allowed us to examine the main effects of both accuracy and explanation, as well as their interaction effect. We followed up significant interactions with simple main effects analyses and pairwise comparisons with Bonferroni correction to identify the source of the effect.

4. Results

We first report descriptive statistics for our key variables and the results of our manipulation checks. Figure 5 shows the mean values and standard errors for expectation violation and the trust-related outcomes (perceived competence, trust, and deviation from system output), broken down by the experimental conditions of system output accuracy and explanation. Table 1 provides descriptive statistics and intercorrelations for all study variables, with values also presented separately for the within-subjects accuracy conditions. We confirmed that all necessary statistical assumptions for the analyses reported herein were met.

The manipulation checks were analyzed with two-way analyses of variance (ANOVAs) with the conditions explanation and accuracy. First, local explanations should lead to increased subjective understanding. There was a significant difference in subjective understanding between the explanation conditions, $F(1, 216) = 16.01$, $p < .001$, $\eta_p^2 = .07$. The local explanations led to a higher subjective understanding ($M = 3.04$, $SD = 0.77$) than no explanations ($M = 2.58$, $SD = 0.91$). Second, we captured perceived accuracy to check whether the participants perceived the manipulated accuracy. There was a significant difference regarding the perceived accuracy between the accuracy conditions, $F(1, 216) = 7.04$, $p = .009$, $\eta_p^2 = .03$. Participants perceived the AI-based system as more accurate for the accurate cases ($M = 2.72$, $SD = 0.89$) than for the inaccurate cases ($M = 2.51$, $SD = 0.98$). Thus, both manipulations seem to have worked as intended.

4.1. Mediation Effects of Accuracy on Trust-Related Outcomes Over Expectation Violations

We expected that inaccurate system outputs lead to expectation violations in decision-makers (H1) and that these expectation violations mediated the relationship between accuracy and the trust-related

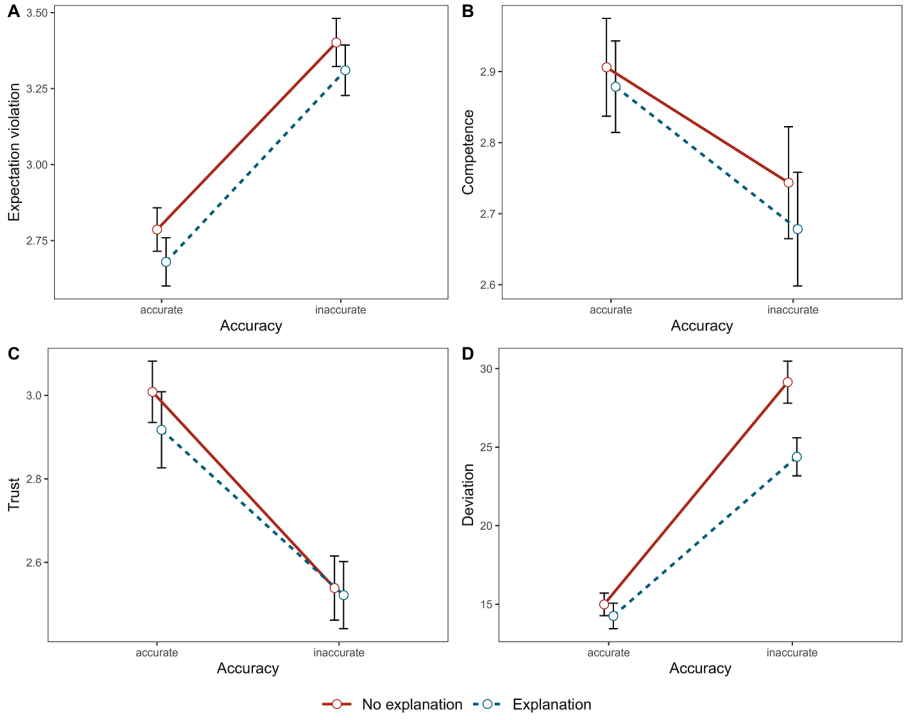


Figure 5. Overview of the Results Regarding Expectation Violation and the Trust-Related Outcomes Depending on the Accuracy and the Explanation Condition.
Note. Error bars represent the standard error of the mean. $N = 218$.

Table 1
Descriptive Statistics, Reliabilities and Correlations of the Study Variables.

	M	*SD*	1	2	3	4	5	6	7	8	9	10	11	12
1. Explanation	-	-												
2. Expectation violation	3.05	0.50	-.08											
3. accurate	2.74	0.62	-.07	.74**										
4. inaccurate	3.36	0.78	-.05	.78**	.15**									
5. Competence	2.80	0.85	-.04	-.56**	-.46**	-.39**								
6. accurate	2.89	0.65	-.02	-.41**	-.38**	-.24**	.82**							
7. inaccurate	2.71	0.70	-.04	-.53**	-.40**	-.40**	.87**	.44**						
8. Trust	2.75	0.83	-.04	-.67**	-.54**	-.48**	.64**	.49**	.59**					
9. accurate	2.97	0.68	-.05	-.48**	-.68**	-.07	.55**	.47**	.47**	.83**				
10. inaccurate	2.53	0.85	-.01	-.62**	-.20**	-.72**	.50**	.34**	.49**	.81**	.34**			
11. Deviation	20.79	0.82	-.16*	.61**	.37**	.55**	-.44**	-.37**	-.37**	-.52**	-.32**	-.54**		
12. accurate	14.65	8.68	-.05	.41**	.59**	.05	-.34**	-.30**	-.28**	-.43**	-.49**	-.20**	.65**	
13. inaccurate	26.94	7.93	-.17*	.54**	.12	.67**	-.36**	-.29**	-.31**	-.42**	-.12	-.57**	.90**	.24**

Note. Unless otherwise indicated, this table shows Pearson product-moment correlations.
Deviation = absolute deviation from output value; lower value indicated higher trusting behavior. The Explanation variable (first row and column in the correlation matrix) was dummy-coded as 0 = no explanation, 1 = explanation for the purpose of calculating its point-biserial correlations with the other variables. $N = 218$. *$p <$ .05. **$p <$ .01.

outcomes competence (H2a), trust (H2b), and deviation (H2c). To analyze the mediation hypotheses, we performed mediations using lavaan (Rosseel, 2012).

Supporting hypothesis 1, inaccurate outputs led to higher expectation violations than accurate outputs in all models.

Hypothesis 2a proposed that expectation violations mediate the relationship between accuracy and perceived competence. There was a significant positive direct effect of accuracy on competence, $B = 0.18$, $p = .015$. After including the mediator (expectation violation) in the model, accuracy was significantly negatively related to expectation violation, $B = -0.62$, $p < .001$, expectation violation was significantly negatively related to competence, $B = -0.37$, $p < .001$, and the direct path of accuracy on competence was not significant anymore, $B = -0.05$, $p = .502$. The indirect effect over expectations violations was significant, $B = 0.23$, $p < .001$. Thus, expectation violations fully mediated the relationship between accuracy and competence, see Figure 6. Therefore, hypothesis 2a was supported.

Hypothesis 2b proposed that expectation violations mediate the relationship between accuracy and trust in the system. There was a significant positive direct effect of accuracy on trust, $B = 0.44$, $p < .001$. After including expectation violation in the model, accuracy was significantly negatively related to expectation violation, $B = -0.62$, $p < .001$, expectation violation was significantly negatively related to trust, $B = -0.72$, $p < .001$, and the direct path of accuracy on trust was not significant anymore, $B = -0.01$, $p = .876$. The indirect effect over expectation violation was statistically significant, $B = 0.45$, $p < .001$. Thus, expectation violations fully mediated the relationship between accuracy and trust, see Figure 7. Therefore, hypothesis 2b was supported.

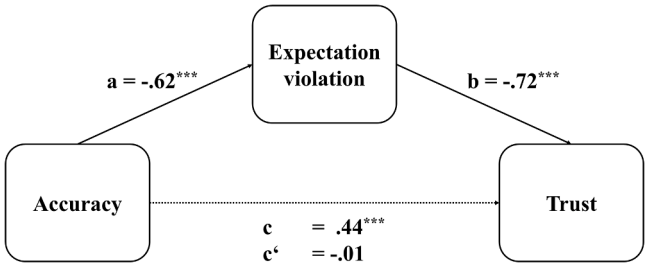


Figure 7. Mediation Model With Trust as Dependent Variable.
Note. a = regression coefficient for the path accuracy on expectation violation, b = regression coefficient for the path expectation violation on trust, c = total effect, c'= direct effect. $N = 218$. *$p < .05$; **$p < .01$.

Hypothesis 2c proposed that expectation violations mediate the relationship between accuracy and deviation from system outputs. There was a significant effect of accuracy on the deviation from system outputs, $B = -12.29$, $p < .001$. After including expectation violation in the model, accuracy was significantly negatively related to expectation violation, $B = -0.62$, $p < .001$, and expectation violation was significantly positively related to deviation from system outputs, $B = 8.54$, $p < .001$. The direct path of accuracy on the deviation was still significant, $B = -6.98$, $p < .001$. The indirect effect over expectation violation also was statistically significant, $B = -5.32$, $p < .001$. Thus, expectation violations partially mediated the relationship between accuracy and deviation from the system outputs, see Figure 8. Therefore, hypothesis 2c was partially supported.

In summary, expectation violations fully mediated the effect of

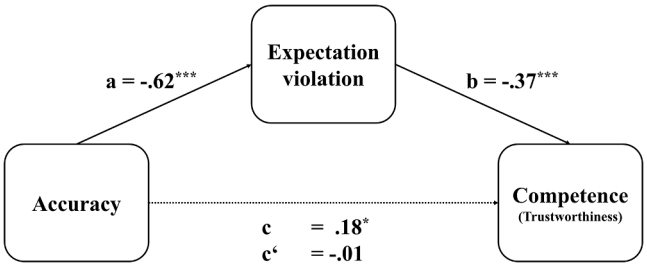


Figure 6. Mediation Model With Perceived Competence (Trustworthiness) as Dependent Variable.
Note. a = regression coefficient for the path accuracy on expectation violation, b = regression coefficient for the path expectation violation on competence (trustworthiness), c = total effect, c' = direct effect. $N = 218$.*$p < .05$; **$p < .01$.

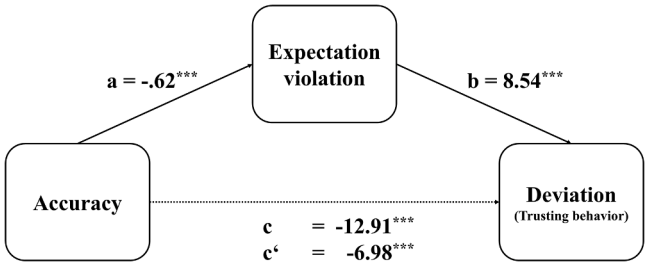


Figure 8. Mediation Model With Deviation (Trusting Behavior) as Dependent Variable.
Note. a = regression coefficient for the path accuracy on expectation violation, b = regression coefficient for the path expectation violation on deviation (trusting behavior), c = total effect, c'= direct effect. $N = 218$. *$p < .05$; **$p < .01$.

accuracy on trust and perceived competence. Expectation violations partially mediated the effect of accuracy on deviation from system outputs.

4.2. Moderating Effects of Explanations

We proposed that the mediation of accuracy via expectation violation on a) perceived competence, b) trust, and c) deviation is moderated by the explanation (H3). There was no moderated mediation for any model, thus hypothesis 3 was not supported. For the sake of conciseness, we decided not to report results regarding the moderated mediation models. Figure 9 summarizes the results regarding all hypotheses.

4.3. Exploratory: Direct Effects of Explanations and Accuracy on Trusting Behavior

We were aware of possible heterogeneous effects of explanations (e.g., Bansal et al., 2021; Eiband et al., 2019) before conducting this study, and we stated in our preregistration that explanations may also have convincing rather than understanding-increasing effects. There is even initial evidence showing that explanations may merely give the impression of a technically competent system (e.g., Bansal et al., 2021; Buçinca et al., 2021).

Given that expectation violations only partially mediated the effect of accuracy on trusting behavior, and given that explanations did affect trusting behavior but not in the way we expected (see Figure 5D), we decided to conduct further exploratory analyses. We computed mixed ANOVAs on the trust-related outcomes. Both explanations, $F(1, 216) = 6.40$, $p < .001$, $\eta_p^2 = .03$, and accuracy, $F(1, 216) = 239.35$, $p < .001$, $\eta_p^2 = .03$, significantly affected trusting behavior. Furthermore, there was a significant two-way interaction between explanations and accuracy on trusting behavior, $F(1, 216) = 7.40$, $p < .001$. See Figure 5D for a graphical representation of the interaction effect. Considering the Bonferroni-adjusted p-value, the simple main effect of explanation was significant for inaccurate cases ($p = .006$) but not for accurate cases ($p = 1$). Pairwise comparisons showed that the mean deviation was significantly different when there were explanations versus when there were no explanations for inaccurate cases ($p = .003$): Participants deviated *less strongly* from inaccurate system outputs when there were explanations available.

5. Discussion

The goal of this study was to examine how the accuracy of system outputs influences users' trust, perceived competence of the system, and trusting behavior, the role of expectations in this process, and how explanations for the system outputs influence these effects. The results demonstrated that inaccurate outputs negatively affected all trust-related variables. Most importantly, participants deviated in their final rating of applicants more strongly from inaccurate outputs. Expectation violation emerged as a full mediator in the relationship between system accuracy and both trust and perceived competence and as a partial mediator between system accuracy and deviation from system output. Moreover, whereas explanations had little impact for accurate outputs, explanations contributed to participants more closely following inaccurate outputs. This highlights the complexity of designing effective explanations.

5.1. Theoretical Implications

5.1.1. The Role of Expectations in Trust-Related Processes

The fact that inaccurate outputs led to more expectation violations than accurate outputs indicates that the participants were able to detect inaccuracies. This is a promising result for human oversight of AI-based systems, as detecting inaccuracies is a crucial precondition for adapting and overwriting system outputs (Langer et al., 2025). Our study thus provides empirical evidence that decision-makers are partially able to detect these inaccuracies.

The results from the mediation analyses provide insights into the dynamics between system accuracy, expectation violations, and trust-related outcomes and help to shed light on understudied psychological processes, the understanding of which is essential for designing effective and user-centered AI systems (Lai et al., 2023). By finding that expectation violations act as a mediator in the trust process of AI-based systems, we empirically support the notion that expectations play a crucial role in the trust process, just as reflected in trust theories and models (Mayer et al., 1995). Empirically supporting the initial evidence that expectation violations may reduce trust in systems (e.g., de Zoeten et al., 2023; Kizilcec, 2016), our study helps to understand how decision-makers perceive and assess the accuracy of system outputs and ultimately how this affects trust outcomes: Confronted with inaccurate outputs, the expectations of the decision-makers concerning system outputs (i.e., they should accurately reflect the state of the world) may

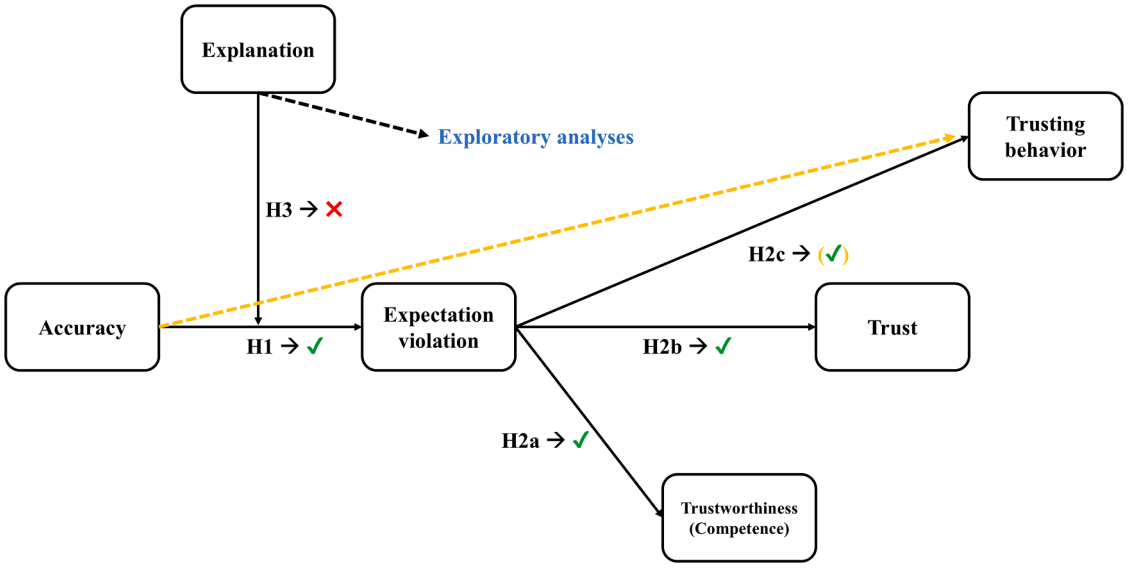


Figure 9. Overview of the Results Regarding the Hypotheses of the Study.

not be fulfilled. Thus, expectation violations may result from comparing the inaccurate system outputs with the (perceived) true state of the world based on available raw data. Expectation violations may then affect trust-related outcomes for specific outputs and for the entire system: Decision-makers perceive the system to be less competent, trust it less, and follow its recommendations less closely. This is positive news for the possible effectiveness of human oversight of AI in high-stakes tasks (Sterz et al., 2024), as decision-makers in our study not only seemed to be able to recognize output inaccuracies but also to build on this insight and deviate from inaccurate outputs.

However, although expectation violations significantly influenced how strongly participants deviated from outputs, a direct and significant effect of the output accuracy manipulation still persisted on participants' deviation from outputs. First, the effects are therefore different for trusting behavior than for trust, which shows that these two aspects of the trust process do not necessarily go hand in hand (see also Papenmeier et al., 2022; Rieger et al., 2022; Schmitt et al., 2021). Second, the fact that expectation violations did not fully mediate the deviation from the output may be associated with an anchoring effect induced by the output value (Furnham & Boo, 2011). Our participants seemed to have been influenced by the numeric value of the output, meaning that for lower numeric values, participants also ended up with lower final ratings of applicants compared to outputs with higher numeric values. Staying closer to available numerical values may not involve any violation or fulfillment of expectations. It may simply imply that users of AI that provides numeric outputs will be affected by the numeric value they see. Specifically, decision-makers may fail to adequately adjust away from anchor values (i.e., the numerical outputs that act as the initial starting point; Tversky & Kahneman, 1974), search for anchor-consistent information based on these anchor values (e.g., in the raw data), or interpret any information in an anchor-consistent way (Strack & Mussweiler, 1997).

It is important to note that while expectation violation in our study was a response to inaccurate AI outputs (as per our experimental design), the broader role of expectation violation in human-AI interaction is nuanced. An AI might correctly challenge users' flawed expectations, possibly improving human decisions. Our study, however, focused on the scenario in which the AI provides erroneous output and whether expectation violations can serve as a signal for users to detect such errors.

5.1.2. *Explanations may be Convincing Rather Than Inaccuracy-Revealing*

The unexpected absence of moderation effects of our explanation manipulation suggests that local explanations did not help decision-makers detect inaccuracies. We also did not find that explanations foster expectation violations. We were therefore unable to support the proposition that explanations make it easier to detect inaccuracies, for instance, by leading decision-makers to scrutinize system outputs and check whether available raw data are consistent with system outputs (Rader et al., 2018; Wang & Yin, 2022). Instead, explanations contributed to decision-makers more closely following inaccurate outputs. This can be interpreted in a way that the explanations' content was less influential than the mere availability of the explanations. In other words, explanations can be convincing, regardless of their content; the mere availability of local explanations can make people more likely to follow system outputs and scrutinize them less strongly (Eiband et al., 2019; Petty & Cacioppo, 1986; Rozenblit & Keil, 2002). Other studies also showed that explanations can increase the tendency to accept inaccurate recommendations (Bansal et al., 2021; Jacobs et al., 2021; Lai & Tan, 2019; Zhang et al., 2020). Accordingly, explanations could lead to a faster and more heuristic processing of system outputs (Buçinca et al., 2021).

From a theoretical point of view, this convincing effect of explanations may imply that (a) the mere availability of explanations works as an evidence of system trustworthiness and (b) explanations lead to a higher subjective understanding, and this is sufficient that people more closely follow system outputs. Regarding (a), explanations or the availability of explanations could be perceived as evidence of system trustworthiness. In the trust literature, facets of trustworthiness for human trustees are ability, benevolence, and integrity (Mayer et al., 1995), facets that have also been discussed with respect to the trustworthiness of automated and AI-based systems (Lee & See, 2004; Schlicker et al., 2025). Regardless of the exact nature of these facets, research proposes that if a trustee shows evidence for their ability, benevolence, and integrity, this contributes to a higher perceived trustworthiness, which should also lead to higher levels of trust and trusting behavior toward that trustee. A trustee's transparency has been discussed as either a sub-facet of perceived integrity or an additional facet of trustworthiness (Schlicker et al., 2025). Providing explanations for system outputs may be perceived as an attempt to make predictions more transparent (Shin, 2021). Regardless of the content of the explanation, the fact that the system is designed in a way that tries to make predictions more transparent could be interpreted as evidence for the perceived trustworthiness of the system, which, in our study, may have led to less strongly deviating from inaccurate cases. Note that we did not measure "perceived transparency" as a facet of system trustworthiness but only the perceived competence of the system. This may explain why we did not find an effect of the explanation manipulation on perceived trustworthiness. Future studies could further explore the role of available explanations as evidence of system trustworthiness.

Regarding (b), explanations may also be convincing because they increase the subjective but not necessarily the actual understanding of system outputs (Eiband et al., 2019; Speith et al., 2024). This may have been the case in our study, where explanations led to higher perceived understanding but did not help to detect inaccurate outputs, possibly a sign that they did not lead to actual understanding (Chromik et al., 2021; Eiband et al., 2019). Accordingly, if explanations look plausible, this may contribute to a subjective understanding of system outputs, which can further complicate trust dynamics. In such cases, decision-makers do not only have to evaluate the accuracy of system outputs but also the accuracy of explanations for outputs. If both or just one of these things are perceived to be plausible, this may contribute to following system outputs more closely (Bansal et al., 2021). For instance, research has shown that people can misplace trust in inaccurate systems when provided with an explanation, even if the explanations are nonsensical or overly simplistic (Lockey et al., 2021). This is problematic given that current approaches to explain AI-based outputs may not necessarily provide accurate explanations but only approximations of rationales for system outputs (Lakkaraju et al., 2019; W. Yang et al., 2023). Explanations that are easier to understand may seem more intelligible but may simultaneously be more likely to be misinterpreted (Xuan et al., 2025). Our study does not allow for more than hypotheses about the role of the perceived understandability of explanations and their perceived (and actual) accuracy, but future research could try to discern the contribution of the perceived accuracy of explanations and the perceived accuracy of outputs to the extent to which decision-makers follow system outputs.

5.2. *Main Practical Implications*

It is crucial for effective human oversight of AI that decision-makers can detect inaccurate outputs (Sterz et al., 2024). Our study shows that this is possible, at least if the degree of inaccuracy is strong enough to induce expectation violations. These results underscore the critical role of managing expectations in designing and deploying AI-based systems. Before the actual use of systems, it is necessary to consider how to communicate to users the capabilities and limitations of systems (e.g., through system manuals or model cards; Mitchell et al., 2019).

During system use, it seems important to consider how to make expectation violations more likely and encourage decision-makers to question their expectations. Although our study, together with other research, shows that explanations may even be detrimental in this

regard, there is also research supporting that explanations are a design option that can be considered to help decision-makers better understand and scrutinize system outputs (Rader et al., 2018; Wang & Yin, 2022). However, explanations must be carefully designed, and their effects must be tested and monitored during system use. Irrespective of why explanations in our study led to decision-makers deviating less strongly from inaccurate outputs, our study, together with other research (Bansal et al., 2021; Cecil et al., 2024), shows that explanations may foster overtrust, which can be detrimental when using AI-based systems in high-risk contexts. Decision-makers might make critical decisions based on inaccurate information and even more so in cases where the system presents seemingly plausible explanations for its outputs (a task that current large language models are very good at; e.g., Kim et al., 2024).

Consistent with these concerns, AI developers and policymakers must consider the implications of explanations on user trust and system transparency. AI systems should be designed to provide explanations that are not only accurate and easy to understand but also clearly indicate the reliability of the AI-based system's outputs (Bansal et al., 2021; Nannini et al., 2023). Taking into account the results of our study, explanations could be designed to evoke expectation violations if the output is inaccurate. Furthermore, rather than focusing primarily on explanations, we need to think about the way people and systems work together: this means getting people to think more about outputs, the task, or implications of their decisions to engage critically with systems (Buçinca et al., 2021). All of this could be directed toward making expectation violations more likely.

Critically, our results also show that even when the system is close to perfect (i.e., resembles the ground truth), as was the case for the accurate outputs in our study, decision-makers deviated on average 10 percentage points from these outputs (see Figure 5D). In our case, this would have reduced the overall accuracy of decisions (see also Neumann et al., 2023). This is evidence that questions the effectiveness of human oversight of AI-based systems: our participants showed overtrust for inaccurate outputs when explanations were available, but also undertrust for accurate outputs. For practice, this means that we need not only to think about how to facilitate the detection of inaccurate outputs but also to facilitate relying on accurate outputs (see Neumann et al., 2023, who propose increasing human autonomy in this regard). This is particularly important in contexts where AI-based systems surpass human capabilities.

5.3. Limitations and Future Work

There are five key limitations to our study. First, our study investigated only one type of local, graphical explanation. The potentially convincing effects we observed, where explanations led to increased reliance on inaccurate outputs, might not generalize to other forms of explanations (e.g., textual, example-based, or interactive explanations), which could have different impacts on user understanding and reliance. Future research needs to explore a wider array of explanation types.

Second, we decided to make output inaccuracies substantial (i.e., 40% downward-biased) and one-dimensional (i.e., the system's evaluation of applicants was consistently worse than their actual suitability). Although real-world output inaccuracies may be less strong and deviate from the ground truth in any direction, our deliberate design choices allow for an unambiguous examination of the effects of such inaccuracies in this study. Future work could investigate more nuanced scenarios with minor inaccuracies and deviations in any direction to gain a broader understanding. Notably, even with the heightened awareness of the system's inaccuracies reflected in the observed expectation violations in our study, strongly inaccurate outputs still affected our participants' decisions.

Third, our study measures trust dynamics *during and after* interaction, but it does not account for participants' initial perceptions of the system *before* the first interaction. As literature in technology adoption highlights (e.g., Distler et al., 2018; Martin et al., 2016), these pre-use expectations can significantly shape post-use perceptions. Future work could measure baseline expectations to track the evolution of trust from pre-use to post-use.

Fourth, the phrasing of our primary single-item trust measure ("How much do you trust the algorithm?") introduces a potential ambiguity. This item, captured after each instance, does not distinguish between trust in the general "system" as an entity versus trust in the specific output just presented. While our multi-item scale (Madsen & Gregor, 2000), measured at the block level, showed a consistent pattern of results, future research could assess whether mentioning different objects of trust (i.e., system vs. single outputs) affects the study outcomes.

Fifth, although we advertised the study to people interested in personnel selection, and most of our participants had experience in personnel selection as applicants, only about 40% of our participants had experience as decision-makers (e.g., hiring managers) in personnel selection. While exploratory analyses confirmed that our main findings are robust and not moderated by this recruiter experience, future work may conduct this study focusing entirely on experts in the field of personnel selection. In principle, expert decision-makers may be better at knowing what supporting or contradicting evidence to look for and may thus be better at distinguishing accurate from inaccurate outputs. However, research supporting that experts are better than lay people in detecting system inaccuracies is limited, and most studies we are aware of show that expert decision-makers also have a hard time distinguishing accurate from inaccurate outputs (Cecil et al., 2024; Green, 2022).

5.4. Conclusion

We have demonstrated a link between system accuracy, expectation violations, and trust-related outcomes and have shown that decision-makers are partly capable of deviating from inaccurate system outputs. Our findings also revealed an interaction between the provision of explanations and the accuracy of AI-based systems. Although explanations are intended to enhance transparency, facilitate user understanding, and help detect inaccuracies, they can also increase the decisional weight of inaccurate outputs. This highlights the double-edged nature of explanations in AI-based systems. On the one hand, they are essential for fostering understanding (Rader et al., 2018; Wang & Yin, 2022); on the other hand, they can mislead users about the capabilities of AI-based systems (Bansal et al., 2021; Cecil et al., 2024; Papenmeier et al., 2022; Xuan et al., 2025). Addressing this challenge requires a careful balance in the design of human-centered AI. It is crucial to provide explanations that accurately represent the system's limitations – explanations that may evoke expectation violations if system outputs are inaccurate. This balance is key to developing AI-based systems that effectively support human decision-making.

CRediT authorship contribution statement

Tim Hunsicker: Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Isabel Duhl:** Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Pascal Haubert:** Methodology, Investigation, Formal analysis, Data curation, Conceptualization. **Linda Onnasch:** Writing – review & editing, Methodology, Conceptualization. **Markus Langer:** Writing – review & editing, Writing – original draft, Visualization, Validation, Supervision, Software, Resources, Project administration, Methodology, Funding acquisition, Data curation, Conceptualization.

Declaration of competing interest

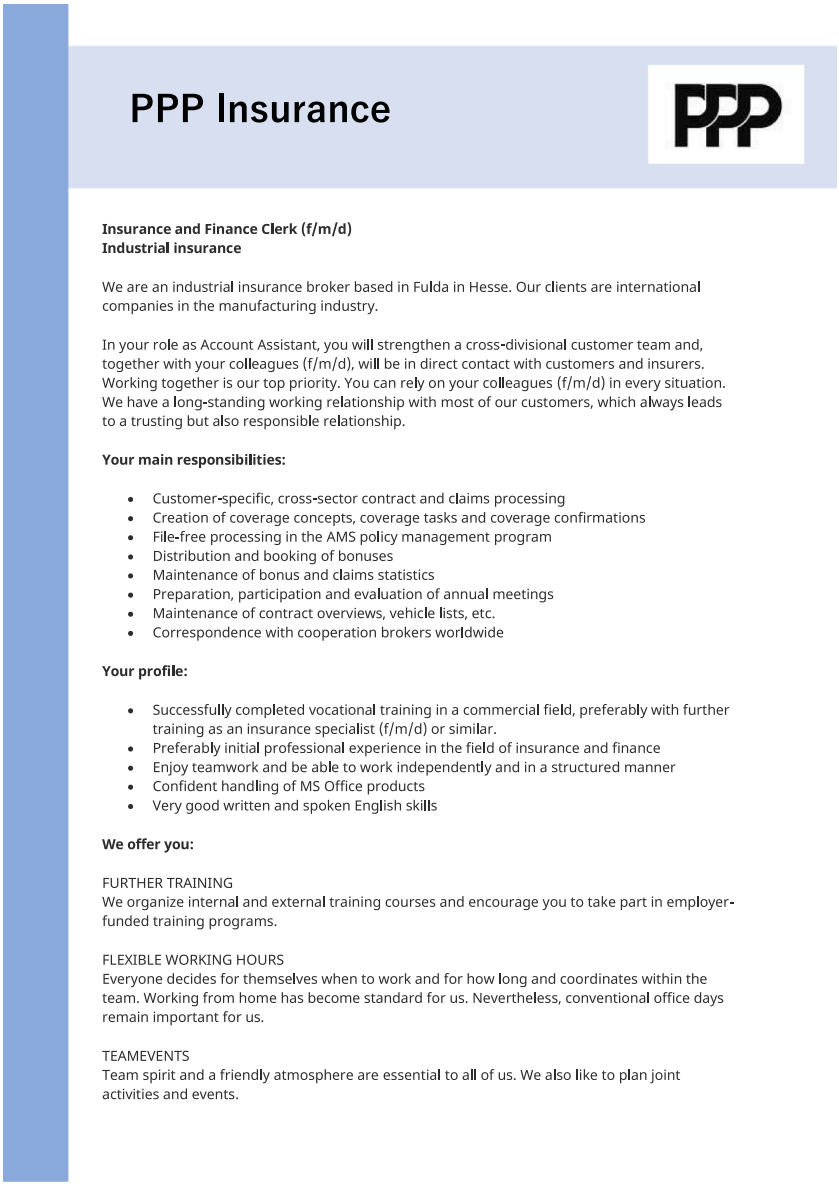
The authors declare that they have no conflict of interest.

Acknowledgments

Work on this article was partially funded by the VolkswagenStiftung in the project "Explainable Intelligent System" (AZ 98513), by the Bundesministerium für Bildung und Forschung (BMBF) in the project "Ophthalmo-AI" (grant 16SV8640), and by the German Research Foundation DFG in the project 389792660 as part of the Transregional Collaborative Research Center TRR 248 "Foundations of Perspicuous Software Systems" project A6.

Appendix A

Study Material (Job Ad)



PPP Insurance

Insurance and Finance Clerk (f/m/d)
Industrial insurance

We are an industrial insurance broker based in Fulda in Hesse. Our clients are international companies in the manufacturing industry.

In your role as Account Assistant, you will strengthen a cross-divisional customer team and, together with your colleagues (f/m/d), will be in direct contact with customers and insurers. Working together is our top priority. You can rely on your colleagues (f/m/d) in every situation. We have a long-standing working relationship with most of our customers, which always leads to a trusting but also responsible relationship.

Your main responsibilities:

- Customer-specific, cross-sector contract and claims processing
- Creation of coverage concepts, coverage tasks and coverage confirmations
- File-free processing in the AMS policy management program
- Distribution and booking of bonuses
- Maintenance of bonus and claims statistics
- Preparation, participation and evaluation of annual meetings
- Maintenance of contract overviews, vehicle lists, etc.
- Correspondence with cooperation brokers worldwide

Your profile:

- Successfully completed vocational training in a commercial field, preferably with further training as an insurance specialist (f/m/d) or similar.
- Preferably initial professional experience in the field of insurance and finance
- Enjoy teamwork and be able to work independently and in a structured manner
- Confident handling of MS Office products
- Very good written and spoken English skills

We offer you:

FURTHER TRAINING
We organize internal and external training courses and encourage you to take part in employer-funded training programs.

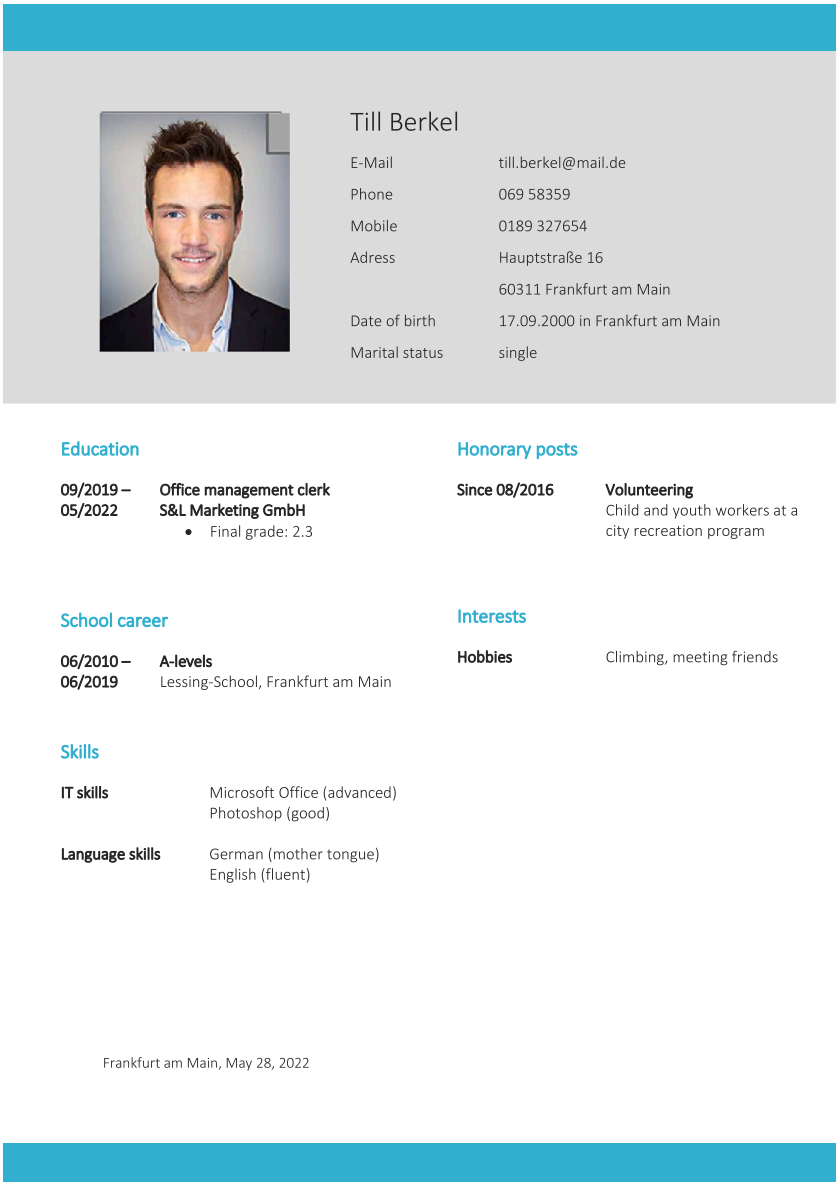
FLEXIBLE WORKING HOURS
Everyone decides for themselves when to work and for how long and coordinates within the team. Working from home has become standard for us. Nevertheless, conventional office days remain important for us.

TEAMEVENTS
Team spirit and a friendly atmosphere are essential to all of us. We also like to plan joint activities and events.

Figure A1. Job Ad (Page 1).
Note. Materials translated from German. A second page of the job advertisement (not shown) provided contact information for applicants.

Appendix B

Study Material (CVs)



Till Berkel

E-Mail	till.berkel@mail.de
Phone	069 58359
Mobile	0189 327654
Adress	Hauptstraße 16
	60311 Frankfurt am Main
Date of birth	17.09.2000 in Frankfurt am Main
Marital status	single

Education

09/2019 – 05/2022 **Office management clerk**
S&L Marketing GmbH
- Final grade: 2.3

School career

06/2010 – 06/2019 **A-levels**
Lessing-School, Frankfurt am Main

Skills

IT skills Microsoft Office (advanced)
Photoshop (good)

Language skills German (mother tongue)
English (fluent)

Honorary posts

Since 08/2016 **Volunteering**
Child and youth workers at a city recreation program

Interests

Hobbies Climbing, meeting friends

Frankfurt am Main, May 28, 2022

Figure B1. CV of Applicant 1 (60% Suitable, Accurately Presented as 60% in System Output).
Note. Materials translated from German.

Milena Sauer

E-Mail	mil.sauer@mail.de
Phone	0641 1019
Mobile	0158 160961
Adress	Kirschweg 13
	35390 Gießen
Date of birth	17.05.1988 in Gießen
Marital status	single

Professional career

Since 09/2010 **Bank clerk**
Volksbank Rhein-Main eG
- Customer service
- Support for online banking
- Insurance sales

Education

09/2007 – 09/2010 **Bank clerk**
Volksbank Rhein-Main eG
- Training
- Final grade: 3.9

Further training

09/2009 – 11/2009 **Further training as an insurance agent**
Insurance sales and customer service

School career

08/1998 – 06/2006 **A-levels**
Liebig-School Gießen, Gießen

Skills

IT skills Microsoft Office (good)

Language skills German (mother tongue)
English (basic knowledge)

Voluntary social year

07/2006 – 07/2007 **Voluntary social year**
Elementary school secretary

Interests

Hobbies Cycling, gardening

Gießen, May 28, 2022

Figure B2. CV of Applicant 2 (60% Suitable, Accurately Presented as 60% in System Output).
Note. Materials translated from German.

Lukas Deckert

E-Mail	m.deckert@mail.de
Phone	0661 38456
Mobile	0170 210548
Adress	Schlossberg 37
	36037 Fulda
Date of birth	03.12.1997 in Fulda
Marital status	single

Professional career

Since 04/2019 **Bank clerk**
Sparkasse Groß-Gerau
- Preparation of credit offers
- Customer service

Skills

IT skills Microsoft Office (good)

Language skills German (mother tongue)
English (basic knowledge)

Education

09/2016 – 04/2019 **Bankkaufmann**
Sparkasse Groß-Gerau
- Training
- Final grade: 2.3

Honorary posts

Since 2014 **Volunteering**
Everyday activities for senior citizens in the Kreuzberg retirement home

School career

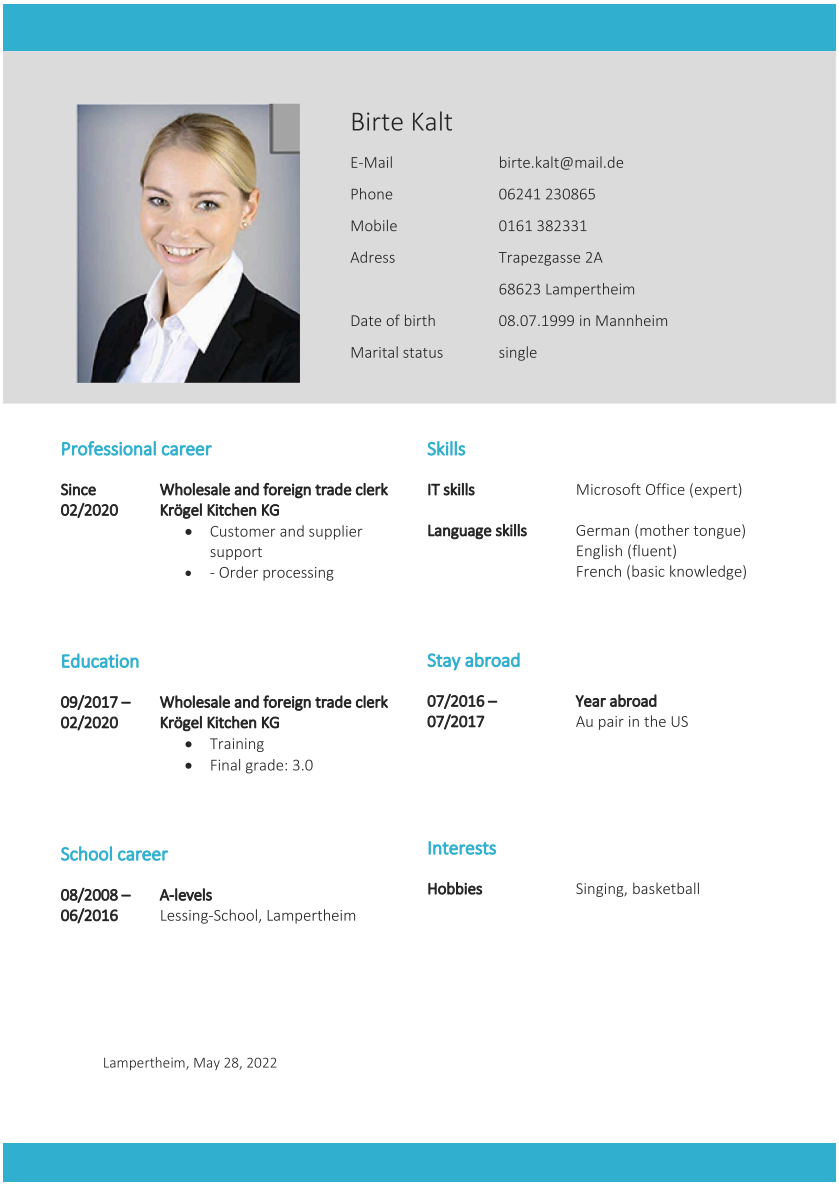
08/2008 – 06/2016 **A-levels**
Winfried School Fulda, Fulda

Interests

Hobbies Cooking, volleyball

Fulda, May 28, 2022

Figure B3. CV of Applicant 3 (50% Suitable, Accurately Presented as 50% in System Output).
Note. Materials translated from German.



Birte Kalt

E-Mail	birte.kalt@mail.de
Phone	06241 230865
Mobile	0161 382331
Adress	Trapezgasse 2A
	68623 Lampertheim
Date of birth	08.07.1999 in Mannheim
Marital status	single

Professional career

Since 02/2020 **Wholesale and foreign trade clerk**
Krögel Kitchen KG
- Customer and supplier support
- - Order processing

Skills

IT skills Microsoft Office (expert)

Language skills German (mother tongue)
English (fluent)
French (basic knowledge)

Education

09/2017 – 02/2020 **Wholesale and foreign trade clerk**
Krögel Kitchen KG
- Training
- Final grade: 3.0

Stay abroad

07/2016 – 07/2017 **Year abroad**
Au pair in the US

School career

08/2008 – 06/2016 **A-levels**
Lessing-School, Lampertheim

Interests

Hobbies Singing, basketball

Lampertheim, May 28, 2022

Figure B4. CV of Applicant 4 (60% Suitable, Inaccurately Presented as 20% in System Output).
Note. Materials translated from German.

Kerstin Leese

E-Mail	kerstin.leese@mail.de
Phone	06106 92867312
Mobile	0178 320875
Adress	Hinter den Birken 13
	63110 Rodgau
Date of birth	02.08.1996 in Rodgau
Marital status	single

Professional career

Since 09/2017 — **Insurance clerk - Claims settlement**
Verso Insurance GmbH

- Administration and support of insurance customers
- Claims settlement

Education

09/2014 – 09/2017 — **Insurance and finance clerk**
Verso Insurance GmbH

- Training with a focus on insurance
- Final grade: 2.7

School career

06/2006 – 06/2014 — **A-levels**
Claus-von-Stauffenberg-School, Rodgau

Skills

IT skills — Microsoft Office (good), DATEV (good)

Language skills — German (mother tongue), English (basic knowledge)

Honorary posts

Since 2012 — **Volunteering**
Youth coach at FSC Rodgau

Interests

Hobbies — Soccer, playing guitar

Rodgau, May 28, 2022

Figure B5. CV of Applicant 5 (70% Suitable, Inaccurately Presented as 30% in System Output).
Note. Materials translated from German.

Marcel Rabenhorst

E-Mail	ma.rabenhorst@mail.de
Phone	06103 4486
Mobile	0171 583793
Adress	Konrad-Schmidt-Weg 3
	63303 Oberursel
Date of birth	26.10.2000 in Frankfurt
Marital status	single

Professional career

Since 09/2021 **Internship - Tax clerk**
Schmidt and Müller tax consultants, Fulda
- Preparation of tax returns
- Maintenance of client data
- Supporting taxpayers with questions

Education

09/2018 – 09/2021 **Tax clerk**
ProFiskus tax consultants, Wiesbaden
- Training
- Final grade: 2.5

School career

06/2010 – 06/2018 **A-levels**
Elly-Heuss-School, Wiesbaden

Skills

IT skills Microsoft Office (good)
DATEV (good)
Elster (good)

Language skills German (mother tongue)
English (basic knowledge)

Honorary posts

Since 2017 **Coaching activity**
Children and youth coach for bouldering

Interests

Hobbies Yoga, bouldering

Oberursel, May 28, 2022

Figure B6. CV of Applicant 6 (60% Suitable, Inaccurately Presented as 20% in System Output).
Note. Materials translated from German.

Appendix C

Study Material (Explanations)

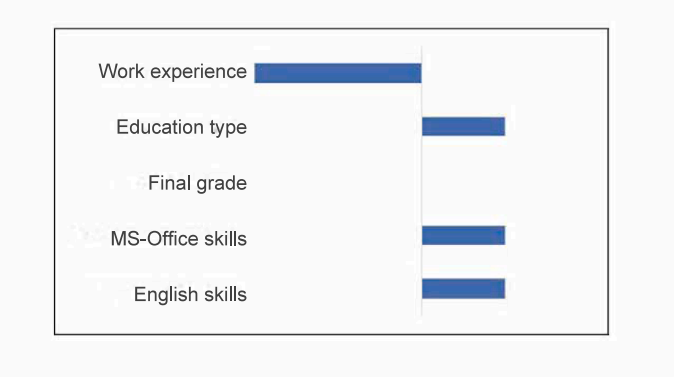


Figure C1. Explanation for Applicant 1 (60% Suitable, Accurately Presented as 60% in System Output).
Note. Materials translated from German.

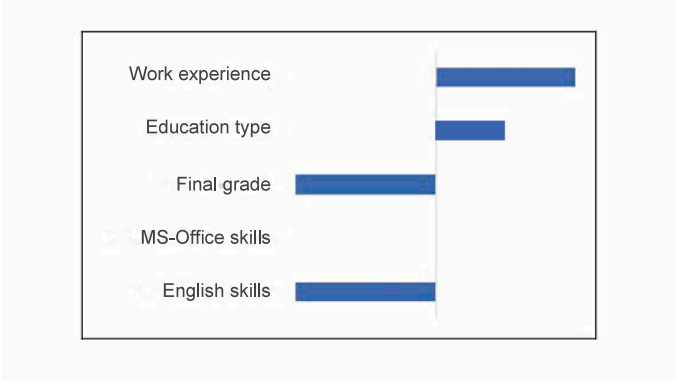


Figure C2. Explanation for Applicant 2 (60% Suitable, Accurately Presented as 60% in System Output).
Note. Materials translated from German.

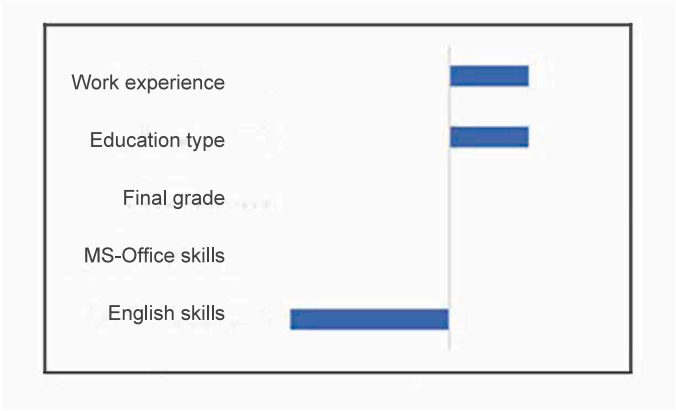


Figure C3. Explanation for Applicant 3 (50% Suitable, Accurately Presented as 50% in System Output).
Note. Materials translated from German.

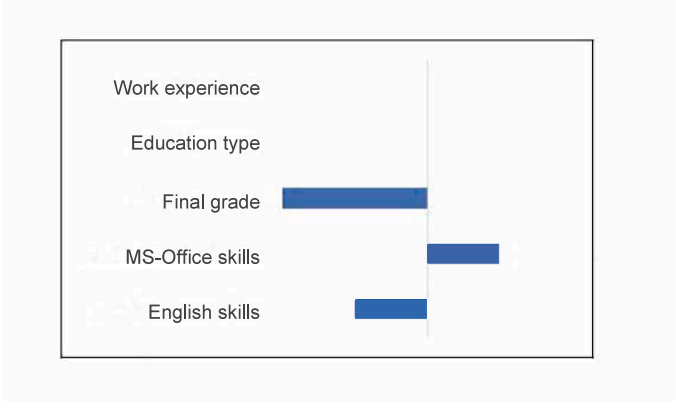


Figure C4. Explanation for Applicant 4 (60% Suitable, Inaccurately Presented as 20% in System Output).
Note. Materials translated from German.

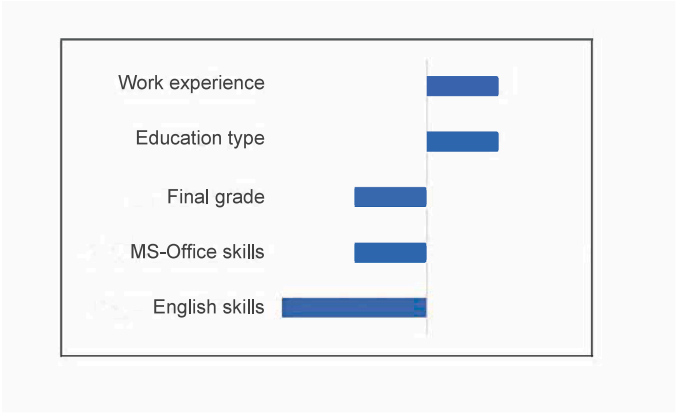


Figure C5. Explanation for Applicant 5 (70% Suitable, Inaccurately Presented as 30% in System Output).
Note. Materials translated from German.

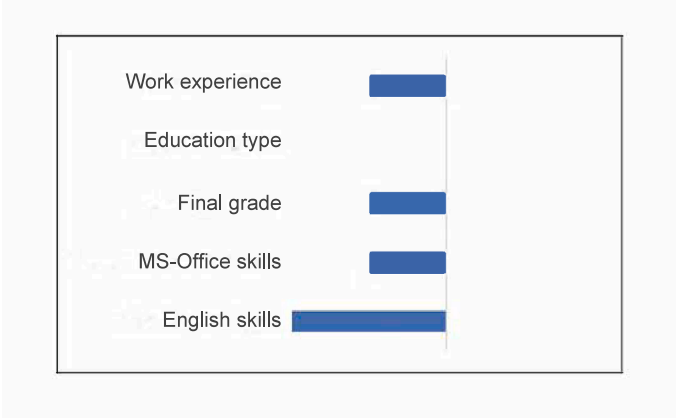


Figure C6. Explanation for Applicant 6 (60% Suitable, Inaccurately Presented as 20% in System Output).
Note. Materials translated from German.

Appendix D

Key Measurement Scales

Table D1
Key Measurement Scales, Items, Response Formats, and Sources.

Scale	Item text	Response format	Source
Trust (One-Item-Measure)	1. How much do you trust the algorithm?	1 (*Completely*) to 5 (*Not at all*)	Rieger et al. (2022)
Trust (Scale Measure)	1. I believe in the recommendation of the algorithm, even if I cannot know for sure if it is correct. 2. If I am unsure about a decision myself, I prefer to believe the algorithm rather than myself. 3. If I am unsure about a decision, I trust that the algorithm will make the best decision regarding the evaluation. 4. If the algorithm makes unusual recommendations, I am confident that the recommendation is correct. 5. I trust the recommendation of the algorithm.	1 (*Strongly disagree*) to 5 (*Strongly agree*)	Madsen & Gregor (2000)
Trusting Behavior	1. This applicant matches the advised position to … percent.	0 to 100	Self-Developed
Perceived Competence	1. The system uses appropriate methods to reach its recommendations. 2. The system has sound knowledge about this type of decision. 3. The system uses the applicants' information correctly. 4. The system uses all available knowledge and information to generate an evaluation.	1 (*Strongly disagree*) to 5 (*Strongly agree*)	Madsen & Gregor (2000)
Expectation Violation	1. The algorithm returned a value that I expected.	1 (*Totally agree*) to 5 (*Totally disagree*)	Burgoon & Walther (1990)

(*continued on next page*)

Table D1 (*continued*)

Scale	Item text	Response format	Source
Affinity for Technology Interaction (ATI-S)	1. I like to occupy myself in greater detail with technical systems. 2. I like to try out the functions of new technical systems. 3. It is enough for me that a technical system works; I don't care how or why. (r) 4. It is enough for me to know the basic functions of a technical system. (r)	1 (*Not at all true*) to 6 (*Completely true*)	Franke et al. (2019)
Perceived Accuracy	1. I believe the outputs made by the algorithm are accurate.	1 (*Strongly disagree*) to 5 (*Strongly agree*)	Shin et al. (2020)
Subjective Understanding	1. I understand how the algorithm works. 2. I can predict what recommendation the algorithm will present for the respective application.	1 (*Strongly disagree*) to 5 (*Strongly agree*)	Wang & Yin (2022)

Note. (r) = reverse-coded item. The response scales for the one-item trust measure were recoded for analysis so that higher values consistently indicated more of the construct.

Data availability

We made all data, analysis code, codebook, and research materials available at https://osf.io/tcyk9.

References

Adadi, A., Berrada, M., 2018. Peeking inside the black-box: A survey on Explainable Artificial Intelligence (XAI). IEEE Access 6, 52138–52160. https://doi.org/10.1109/ACCESS.2018.2870052.

Adomavicius, G., Zhang, J., 2012. Stability of recommendation algorithms. ACM Transactions on Information Systems 30 (4), 23:1–23:31. https://doi.org/10.1145/2382438.2382442.

Bansal, G., Wu, T., Zhou, J., Fok, R., Nushi, B., Kamar, E., Ribeiro, M.T., Weld, D., 2021. Does the whole exceed its parts? The effect of AI explanations on complementary team performance. In: Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems, CHI '21, 81. https://doi.org/10.1145/3411764.3445717.

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.

Buçinca, Z., Malaya, M.B., Gajos, K.Z., 2021. To trust or to think: Cognitive forcing functions can reduce overreliance on AI in AI-assisted decision-making. Proceedings of the ACM on Human-Computer Interaction 5 (CSCW1), 188. https://doi.org/10.1145/3449287.

Burgoon, J.K., Walther, J.B., 1990. Nonverbal expectancies and the evaluative consequences of violations. Human Communication Research 17 (2), 232–265. https://doi.org/10.1111/j.1468-2958.1990.tb00232.x.

Burkart, N., Huber, M.F., 2021. A survey on the explainability of supervised Machine Learning. Journal of Artificial Intelligence Research 70, 245–317. https://doi.org/10.1613/jair.1.12228.

Cecil, J., Lermer, E., Hudecek, M.F.C., Sauer, J., Gaube, S., 2024. Explainability does not mitigate the negative impact of incorrect AI advice in a personnel selection task. Scientific Reports 14 (1), 9736. https://doi.org/10.1038/s41598-024-60220-5.

Chromik, M., Eiband, M., Buchner, F., Krüger, A., Butz, A., 2021. I think I get your point, AI! The illusion of explanatory depth in Explainable AI. In: 26th International Conference on Intelligent User Interfaces, pp. 307–317. https://doi.org/10.1145/3397481.3450644.

de Visser, E.J., Peeters, M.M.M., Jung, M.F., Kohn, S., Shaw, T.H., Pak, R., Neerincx, M. A., 2020. Towards a theory of longitudinal trust calibration in human–robot teams. International Journal of Social Robotics 12 (2), 459–478. https://doi.org/10.1007/s12369-019-00596-x.

de Zoeten, M., Ernst, C.-P., Rothlauf, F., 2023. A matter of trust: How trust in AI-based systems changes during interaction. In: *AMCIS 2023 Proceedings*, Article 15. https://aisel.aisnet.org/amcis2023/sig_odis/sig_odis/15.

Distler, V., Lallemand, C., Bellet, T., 2018. Acceptability and acceptance of autonomous mobility on demand: The impact of an immersive experience. Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI ' 18, 1–10. https://doi.org/10.1145/3173574.3174186.

Eiband, M., Buschek, D., Kremer, A., Hussmann, H., 2019. The impact of placebic explanations on trust in intelligent systems. In: Extended Abstracts of the 2019 CHI Conference on Human Factors in Computing Systems, CHI EA '19, pp. 1–6. https://doi.org/10.1145/3290607.3312787.

Enarsson, T., Enqvist, L., Naarttijärvi, M., 2022. Approaching the human in the loop – legal perspectives on hybrid human/algorithmic decision-making in three contexts. Information & Communications Technology Law 31 (1), 123–153. https://doi.org/10.1080/13600834.2021.1958860.

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

Fischhoff, B., Broomell, S.B., 2020. Judgment and decision making. Annual Review of Psychology 71, 331–355. https://doi.org/10.1146/annurev-psych-010419-050747.

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.

Fritz, M.S., MacKinnon, D.P., 2007. Required sample size to detect the mediated effect. Psychological Science 18 (3), 233–239. https://doi.org/10.1111/j.1467-9280.2007.01882.

Furnham, A., Boo, H.C., 2011. A literature review of the anchoring effect. The Journal of Socio-Economics 40 (1), 35–42. https://doi.org/10.1016/j.socec.2010.10.008.

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.

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.

Jacobs, M., Pradier, M.F., McCoy, T.H., Perlis, R.H., Doshi-Velez, F., Gajos, K.Z., 2021. How machine-learning recommendations influence clinician treatment selections: The example of antidepressant selection. Translational Psychiatry 11 (1), 1. https://doi.org/10.1038/s41398-021-01224-x.

Kim, S.S.Y., Liao, Q.V., Vorvoreanu, M., Ballard, S., Vaughan, J.W., 2024. "I'm not sure, but...": Examining the impact of Large Language Models' uncertainty expression on user reliance and trust. ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT '24). https://doi.org/10.1145/3630106.3658941.

Kizilcec, R.F., 2016. How much information? Effects of transparency on trust in an algorithmic interface. In: Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems, CHI '16, pp. 2390–2395. https://doi.org/10.1145/2858036.2858402.

Köchling, A., Wehner, M.C., Warkocz, J., 2023. Can I show my skills? Affective responses to artificial intelligence in the recruitment process. Review of Managerial Science 17 (6), 2109–2138. https://doi.org/10.1007/s11846-021-00514-4.

Kohn, S.C., De Visser, E.J., Wiese, E., Lee, Y.-C., Shaw, T.H, 2021. Measurement of trust in automation: A narrative review and reference guide. Frontiers in Psychology 12, 604977. https://doi.org/10.3389/fpsyg.2021.604977.

Lai, V., Chen, C., Smith-Renner, A., Liao, Q.V., Tan, C., 2023. Towards a science of human-AI decision making: An overview of design space in empirical human-subject studies. In: Proceedings of the 2023 ACM Conference on Fairness, Accountability, and Transparency, FAccT '23, pp. 1369–1385. https://doi.org/10.1145/3593013.3594087.

Lai, V., Tan, C., 2019. On human predictions with explanations and predictions of Machine Learning models: A case study on deception detection. In: Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* '19, pp. 29–38. https://doi.org/10.1145/3287560.3287590.

Lakkaraju, H., Kamar, E., Caruana, R., Leskovec, J., 2019. Faithful and customizable explanations of black box models. In: Proceedings of the 2019 AAAI/ACM Conference on AI, Ethics, and Society, AIES '19, pp. 131–138. https://doi.org/10.1145/3306618.3314229.

Langer, M., Baum, K., Schlicker, N., 2025. Effective human oversight of AI-based systems: A signal detection perspective on the detection of inaccurate and unfair outputs. Minds and Machines 35 (1), 1. https://doi.org/10.1007/s11023-024-09701-0.

Langer, M., Oster, D., Speith, T., Hermanns, H., Kästner, L., Schmidt, E., Sesing, A., Baum, K., 2021. What do we want from Explainable Artificial Intelligence (XAI)? – A stakeholder perspective on XAI and a conceptual model guiding interdisciplinary XAI research. Artificial Intelligence 296, 103473. https://doi.org/10.1016/j.artint.2021.103473.

LeCun, Y., Bengio, Y., Hinton, G., 2015. Deep learning. Nature 521 (7553), 436–444. https://doi.org/10.1038/nature14539.

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.

Leiner, D.J., 2022. *SoSci Survey* (Version 3.3.19) [Computer software]. SoSci Survey GmbH. https://www.soscisurvey.de.

Lockey, S., Gillespie, N., Holm, D., Someh, I.A., 2021. A review of trust in artificial intelligence: Challenges, vulnerabilities and future directions. In: Hawaii International Conference on System Sciences 2021 (HICSS-54). https://aisel.aisnet.org/hicss-54/os/trust/2.

Madsen, M., Gregor, S., 2000. Measuring human-computer trust. In: Proceedings of the 11th Australasian Conference on Information Systems, pp. 6–8.

Martin, N., Jamet, É., Erhel, S., Rouxel, G., 2016. From acceptability to acceptance: Does experience with the product influence user initial representations? In: Stephanidis, C. (Ed.), HCI International 2016 – Posters' Extended Abstracts. Springer International Publishing, pp. 128–133. https://doi.org/10.1007/978-3-319-40548-3_21.

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), 12:1–12:25. https://doi.org/10.1145/1985347.198535.

Meussling, B., Roesler, E., & Rieger, T. (2022). *Explainability and error experience in Human-AI interaction: The influence on trust and dependence*. https://doi.org/10.1016/j.ijhcs.2025.103505.

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

Mitchell, M., Wu, S., Zaldivar, A., Barnes, P., Vasserman, L., Hutchinson, B., Spitzer, E., Raji, I.D., Gebru, T., 2019. Model cards for model reporting. In: Proceedings of the Conference on Fairness, Accountability, and Transparency, FAT* '19, pp. 220–229. https://doi.org/10.1145/3287560.3287596.

Nannini, L., Balayn, A., Smith, A.L., 2023. Explainability in AI policies: A critical review of communications, reports, regulations, and standards in the EU, US, and UK. 2023 ACM Conference on Fairness, Accountability, and Transparency 1198–1212. https://doi.org/10.1145/3593013.3594074.

Neumann, M., Niessen, A.S.M., Linde, M., Tendeiro, J.N., Meijer, R.R., 2023. Adding an egg" in algorithmic decision making: Improving stakeholder and user perceptions, and predictive validity by enhancing autonomy. European Journal of Work and Organizational Psychology 0 (0), 1–18. https://doi.org/10.1080/1359432X.2023.2260540.

Papenmeier, A., Kern, D., Englebienne, G., Seifert, C., 2022. It's complicated: The relationship between user trust, model accuracy and explanations in AI. ACM Transactions on Computer-Human Interaction 29 (4), 35:1–35:33.

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.

Petty, R.E., Cacioppo, J.T., 1986. The elaboration likelihood model of persuasion. In: Berkowitz, L. (Ed.), Advances in Experimental Social Psychology, Vol. 19. Academic Press, pp. 123–205. https://doi.org/10.1016/S0065-2601(08)60214-2.

Rader, E., Cotter, K., Cho, J., 2018. Explanations as mechanisms for supporting algorithmic transparency. In: Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems, CHI '18, pp. 1–13. https://doi.org/10.1145/3173574.3173677.

Ribeiro, M.T., Singh, S., Guestrin, C., 2016. Why should I trust you?": Explaining the predictions of any classifier. In: 22nd ACM SIGKDD International Conference on Knowledge Discovery and Data Mining, KDD '16, pp. 1135–1144. https://doi.org/10.1145/2939672.2939778.

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.

Rosseel, Y., 2012. lavaan: An R package for structural equation modeling. Journal of Statistical Software 48 (2), 1–36. https://doi.org/10.18637/jss.v048.i02.

Rousseau, D.M., Sitkin, S.B., Burt, R.S., Camerer, C., 1998. Not so different after all: A cross-discipline view of trust. Academy of Management Review 23 (3), 393–404. https://doi.org/10.5465/amr.1998.926617.

Rozenblit, L., Keil, F., 2002. The misunderstood limits of folk science: An illusion of explanatory depth. Cognitive Science 26 (5), 521–562. https://doi.org/10.1016/S0364-0213(02)00078-2.

Rudin, C., 2019. Stop explaining black box machine learning models for high stakes decisions and use interpretable models instead. Nature Machine Intelligence 1 (5), 5. https://doi.org/10.1038/s42256-019-0048-x.

Schaefer, K.E., Chen, J.Y.C., Szalma, J.L., Hancock, P.A., 2016. A meta-analysis of factors influencing the development of trust in automation: Implications for understanding autonomy in future systems. Human Factors 58 (3), 377–400. https://doi.org/10.1177/0018720816634228.

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. In: 2022 AAAI/ACM Conference on AI, Ethics, and Society, pp. 617–626. https://doi.org/10.1145/3514094.3534128.

Schlicker, N., Baum, K., Uhde, A., Sterz, S., Hirsch, M.C., Langer, M., 2025. How do we assess the trustworthiness of AI? Introducing the trustworthiness assessment model (TrAM). Computers in Human Behavior 170, 108671. https://doi.org/10.1016/j.chb.2025.108671.

Schlicker, N., Langer, M., 2021. Towards warranted trust: A model on the relation between actual and perceived system trustworthiness. Mensch Und Computer 2021, 325–329. https://doi.org/10.1145/3473856.3474018.

Schmitt, A., Wambsganss, T., Soellner, M., Janson, A., 2021. Towards a trust reliance paradox? Exploring the gap between perceived trust in and reliance on algorithmic advice. In: ICIS 2021 Proceedings. https://aisel.aisnet.org/icis2021/ai_business/ai_business/14.

Shin, D., 2021. The effects of explainability and causability on perception, trust, and acceptance: Implications for explainable AI. International Journal of Human-Computer Studies 146, 102551. https://doi.org/10.1016/j.ijhcs.2020.102551.

Shin, D., Zhong, B., Biocca, F.A., 2020. Beyond user experience: What constitutes algorithmic experiences? International Journal of Information Management 52, 102061. https://doi.org/10.1016/j.ijinfomgt.2019.102061.

Speith, T., Crook, B., Mann, S., Schomäcker, A., Langer, M., 2024. Conceptualizing understanding in Explainable Artificial Intelligence (XAI): An abilities-based approach. Ethics and Information Technology 26 (40). https://doi.org/10.1007/s10676-024-09769-3.

Springer, A., Whittaker, S., 2020. Progressive disclosure: When, why, and how do users want algorithmic transparency information? ACM Transactions on Interactive Intelligent Systems 10 (4), 29:1–29:32. https://doi.org/10.1145/3374218.

Sterz, S., Baum, K., Biewer, S., Hermanns, H., Lauber-Rönsberg, A., Meinel, P., Langer, M., 2024. On the quest for effectiveness in human oversight: Interdisciplinary perspectives. In: ACM Conference on Fairness, Accountability, and Transparency (ACM FAccT '24), pp. 2495–2507. https://doi.org/10.1145/3630106.3659051.

Strack, F., Mussweiler, T., 1997. Explaining the enigmatic anchoring effect: Mechanisms of selective accessibility. Journal of Personality and Social Psychology 73 (3), 437–446. https://doi.org/10.1037/0022-3514.73.3.437.

Tversky, A., Kahneman, D., 1974. Judgment under uncertainty: Heuristics and biases: Biases in judgments reveal some heuristics of thinking under uncertainty. Science 185 (4157), 1124–1131. https://doi.org/10.1126/science.185.4157.1124.

Wang, X., Yin, M., 2022. Effects of explanations in AI-assisted decision making: Principles and comparisons. ACM Transactions on Interactive Intelligent Systems 12 (4), 27. https://doi.org/10.1145/3519266.

Wischnewski, M., Krämer, N., Müller, E., 2023. Measuring and understanding trust calibrations for automated systems: A survey of the state-of-the-art and future directions. In: Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems, CHI '23, 755. https://doi.org/10.1145/3544548.3581197.

Xuan, Y., Small, E., Sokol, K., Hettiachchi, D., Sanderson, M., 2025. Comprehension is a double-edged sword: Over-interpreting unspecified information in intelligible machine learning explanations. International Journal of Human-Computer Studies 193, 103376. https://doi.org/10.1016/j.ijhcs.2024.103376.

Yang, F., Huang, Z., Scholtz, J., Arendt, D.L., 2020. How do visual explanations foster end users' appropriate trust in machine learning?. In: 25th International Conference on Intelligent User Interfaces (IUI '20), IUI '20, pp. 189–201. https://doi.org/10.1145/3377325.3377480.

Yang, W., Wei, Y., Wei, H., Chen, Y., Huang, G., Li, X., Li, R., Yao, N., Wang, X., Gu, X., Amin, M.B., Kang, B., 2023. Survey on explainable AI: From approaches, limitations and applications aspects. Human-Centric Intelligent Systems 3 (3), 161–188. https://doi.org/10.1007/s44230-023-00038-y.

Zerilli, J., Knott, A., Maclaurin, J., Gavaghan, C., 2019. Transparency in algorithmic and human decision-making: Is there a double standard? Philosophy & Technology 32 (4), 661–683. https://doi.org/10.1007/s13347-018-0330-6.

Zhang, Y., Liao, Q.V., Bellamy, R.K.E., 2020. Effect of confidence and explanation on accuracy and trust calibration in AI-assisted decision making. In: Conference on Fairness, Accountability, and Transparency (FAT* '20), FAT* '20, pp. 295–305. https://doi.org/10.1145/3351095.3372852.