User Judgment of an AI Model is Biased by its Description: A Study in a Job Interview Training Context

Sharon Lynn Chu, Marcin Karcz, Amal Hashky, Neha Rani, Theodora Chaspari, Winfred Arthur Jr., Eric D. Ragan

PII: S1071-5819(25)00248-4
DOI: https://doi.org/10.1016/j.ijhcs.2025.103691
Reference: YIJHC 103691

To appear in: *International Journal of Human - Computer Studies*

Received Date: 13 February 2025
Accepted Date: 18 November 2025

Please cite this article as: Chu, SL., Karcz, M., Hashky, A., Rani, N., Chaspari, T., Arthur Jr., W., Ragan, ED., User Judgment of an AI Model is Biased by its Description: A Study in a Job Interview Training Context, *International Journal of Human - Computer Studies* (2025), doi: https://doi.org/10.1016/j.ijhcs.2025.103691.

Highlights

User Judgment of an AI Model is Biased by its Description: A Study in a Job Interview Training Context

Sharon Lynn Chu, Marcin Karcz, Amal Hashky, Neha Rani, Theodora Chaspari, Winfred Arthur Jr., Eric D. Ragan

- Users' impressions of the model's origin shape how they perceive its abilities.
- Sophisticated model descriptions result in higher agreement with the model outputs.
- Users were more willing to align ratings with AI decisions in sophisticated model descriptions.
- Users reported a greater willingness to apply system outputs in advanced model descriptions.

User Judgment of an AI Model is Biased by its Description: A Study in a Job Interview Training Context

Sharon Lynn Chu[a], Marcin Karcz[a], Amal Hashky[a], Neha Rani[a], Theodora Chaspari[b], Winfred Arthur Jr.[c], Eric D. Ragan[a,*]

[a] *University of Florida, Gainesville, 32611, Florida, United States*
[b] *University of Colorado, Boulder, 80309, Colorado, United States*
[c] *Texas A&M University, College Station, 77840, Texas, United States*

Abstract

The growth of artificial intelligence (AI) has introduced new AI-powered systems in many aspects of everyday life. Although many such systems may be embedded in contexts where long-term use is justified, there are also many cases where usage of such systems can be brief or within a single session. In those cases, initial information given to the user about the AI model is important since users may not have enough engagement with the system to develop a mental model over time through use. Instead, users may simply rely on first impressions. However, little is known about how given information about the AI model in a system affects user judgment of the system. This work investigates this question within the context of job interview training. We conducted a controlled experiment where the description of the AI model within a simulated job interview training system was manipulated to describe the model as being either basic or more advanced. Participants in the condition where the AI model was described as more sophisticated and advanced reported significantly higher levels of agreement with the model outputs, more favorable ratings, and a greater willingness to use the system output.

Keywords:
Human-centered AI, Explainable AI, Human Biases, First Impressions, Job

[*] Corresponding author
Email address: eragan@ufl.edu (Eric D. Ragan)

Interview Training Systems (JITS)

1. Introduction

Artificial intelligence (AI) is now being used to assist in human decision making in a broad range of situations—from mushroom picking [1] to deciding on medical treatment [2]. However, AI is far from perfect; errors are inevitable. A serious danger with the use of AI stems from the fact that most end users are often unaware of the limitations of AI models and a system's actual performance. Lack of knowledge of model capabilities awareness of limitations can lead to problems with user trust, whether it is over trust (relying too much on the AI) [1, 3, 4] or under trust (not leveraging enough the benefits of the AI) [5, 6, 7].

While users may develop an improved understanding of model capabilities through continued use over time [2, 8], many practical use cases for AI applications would not involve sufficiently long exposure to the system to allow users to develop an accurate understanding of the model. When end users first begin using an AI system, they will make assumptions of model capabilities based on a combination of their existing knowledge and the information presented by the system [3, 9, 10]. First impressions based on initially available information are crucial. For example, prior studies have found that if users notice system errors early during their interactions with an AI system, they are likely to quickly distrust or underestimate the model's accuracy [7, 8]. Beyond judgment of AI, a large body of research has demonstrated how numerous different factors can influence first impression assessments, with a broad range of examples including the quality of visual aesthetics [11] or the availability of educational briefings [12]. Our research is motivated by the need to similarly add deeper knowledge of what aspects of AI applications contribute to users' judgments. We hypothesize that many other design factors of AI systems may also contribute to first impressions and potentially influence user perceptions of AI.

In particular, we study how non-expert users might be influenced by provided technical information about the type and origin of an AI system. That is, it is unknown whether novice end-users are concerned with technical details—for example, the type of AI model, the reputation of the developer, or the perceived technical complexity—as they form their impressions of the system. This paper presents an experiment of how the presentation of initial introductory information about the AI model affects users' first impressions or judgment, including their agreement with and trust in the system, and their willingness to use the system.

As the application area for the study, our research is conducted in the context of AI-supported job interview training systems. Training systems for job interviews analyze people's performance during practice job interviews and provide feedback for them to improve. These kinds of systems have a significant influence on how people behave during real job interviews that may lead to them securing a job or not [13, 14]. Investigating the impact of initial introductory model description in such a job interview training system context provides an authentic scenario in which users have a personal stake in trusting the AI system or not.

2. Background and Related Work

2.1. Trust and Reliance on AI Assistance

How users judge AI and choose whether to rely on automated support depends on many factors such user experience, mental workload, temporal demand, or fatigue [15, 16]. While support systems are implemented with the intention of providing assistance to improve human decision making, users are not always comfortable relying on systems they do not understand or do not trust. That is, if a user distrusts the outputs of a computational model when correct, it may result in underreliance on the support, which could reduce overall efficiency or lead to incorrect decision making (e.g., [17, 15]). Conversely, *overreliance* on automation (also called *automation bias*) can occur when users accept AI recommendations too readily, even in problem cases where the AI is incorrect or when outside information contradicts the AI [16, 18]. Classic examples of overreliance include cases where vehicle operators (e.g., airline pilots or ship mariners) have had critical accidents due to too much trust on faulty or misinterpreted system outputs [16, 19]. Numerous lab studies have also observed overreliance. For instance, in a study with a simulated aircraft monitoring task by Skitka et al. [20], participants demonstrated higher error rates with automation support than participants without it—as the given support was imperfect, and participants did not appropriately distinguish between good or bad outputs. Another example by Will [21] involved a guidance system for an engineering task requiring identification of undesirable pressure buildup in a well. In this case, overreliance on the AI guidance caused problems for both novice and experienced engineers for cases where the system outputs were wrong.

As behaviors of both underreliance and overreliance are suboptimal, the goal is to enable *appropriate reliance* so that users can judge when to trust and when to discount AI outputs [22, 23]. Researchers have explored whether *explanations* for AI logic might provide the necessary information and context behind AI decisions to facilitate better overall decision making [24, 25, 26]. For example, through studies using an object detection task in images an a simulated AI aid, Dzindolet et al.[22] found that study participants quickly lost trust in the reliability of the AI aid when they observed it making mistakes. Importantly, the research demonstrated the loss of trust could be prevented if participants were provided with explanations for the AI mistakes. While this effect might be interpreted as evidence of the intended result of appropriate reliance, the authors note that the effect of explanations increasing trust may have been unwarranted, meaning the inclusion of explanations could risk leading to overreliance. Evidence of similar concerns of biasing effects due to explanations was discussed by Bansal et al. [27]. Their studies with a sentiment classification task found that including explanations increased the frequency that users acted in agreement with AI recommendations.

Research by Bucinca et al. [28] also studied problems with overreliance on AI with the inclusion of explanations. Their study showed how differences in availability of algorithmic explanations (e.g., cost based on additional interaction to access information or added temporal delay to reveal information) can influence the extent to which users adopted overreliance behaviors. The authors concluded that promoting deeper thinking about algorithmic processing and the reasons for different model outputs can reduce overreliance, though users preferred more simplistic and less mentally demanding explanations—even though the latter were found to lead to worse overall task performance. Research Vasconcelos et al. [29] also addressed the cost-benefit tradeoffs for the effort needed to achieve appropriate reliance compared to the effort of manually doing the task without the AI. That is, if trying to interpret AI explanations to judge the validity AI outputs is more taxing than the task itself, users might as well ignore the AI, possibly leading to underreliance. The authors discuss how overreliance may be more of a concern for more difficult tasks or for cases when interpreting AI explanations requires greater effort. The effort of interpreting explanatory information is relevant to our research, as we study whether a simple presentation of up-front framing information has the potential advantage of serving as a low-effort, one-time form of explanation. However, providing information up-front does not encourage continued critical thinking about the implications of that information over time, and it does not serve as meaningful explanations on post-hoc basis. Thus, this complexity motivates investigation of how up-front framing information might influence user judgment or impressions of AI.

2.2. Algorithm Explanations

The goal for explanations is to bridge the gap in users' understanding of the system. Explanations have been widely explored as a way to present information about the AI model in a system to users. Prior literature includes a broad array of examples of different explanation strategies (e.g., [30, 31, 25, 26]). For example, explanations can be at a global level, explaining the overall structure of a model, or at a local level, explaining a specific decision made by the AI [10, 30, 32]. Factual explanations focus on facts, data, and actual features of the model that resulted in the given output [33, 34, 35], while counterfactual explanations focus on how changes to certain features or inputs would affect the model's outcomes [33, 34, 35, 36]. Post-hoc explanations are a common approach that approximate the workings of existing models by helping users identify relevant information that contributed to specific outputs. [3, 37, 38]. Other explanation types include differentiating reasons *why* and *why not* certain outcomes are observed as model outputs for a given input [39].

To aid in understanding the wide range of existing explanation techniques, Arrieta et al. [40] presents a descriptive taxonomy of explanation examples organized by common model types and properties, with examples including distinction between inherently interpretable models compared to black box methods requiring post-hoc techniques, and also grouping deeper levels of model-specific and model-agnostic explanation techniques. Separately, a survey by Dwivedi et al. [41] presents a taxonomy with attention to common phases of machine learning development processes along with discussion of trade-offs among different explanation options balanced with project priorities. Taking a different approach from the perspective of developers making explanation choices for specific cases, Bennetot et al. [26] provide a guide for selecting an explanation approach based on specific needs. For example, their guide recommends different explanation options based on details of the data type (e.g., textual, image, tabular) or explanation properties (e.g., interactive explanations, counterfactual explanations). As another example, Mohseni et al. [30] offers recommendations based on the needs of different types of user needs, suggesting an integrated design and evaluation process with consideration of human-centric metrics (e.g., user trust, human-AI task performance, cognitive load) and model-centered metrics (e.g., explanation fidelity, model performance).

While algorithmic transparency has been a topic of interest for a several decades, trends in the literature indicate a sharp increase in interest in explanation after 2015 [42]. Despite the rapid advancement of explanation techniques and designs, scientific knowledge of how such different types of explanations influences users has not kept pace [43, 44]. In a large review in 2025, Suh et al. [43] estimated that approximately only 0.7% of a sample of 18,254 papers about explainable AI included human-subjects validation or studies of their effects on user behaviors. As discussed in a review by Miller [44], a wide range of factors influence how users give attention to and interpret algorithmic capabilities. Users can be quick to discount benefits of algorithmic capabilities if the system logic is not considered to align with human logic [45, 46], and first impressions with a system's effectiveness carry a heavy influence that many users are reluctant to adjust [7, 47, 48, 6].

A review by Raees et al. [31] highlighted the importance of considering the level of user interaction with explanatory information when studying different types of explainable systems. That is, cases where users are passive receivers of post-hoc explanatory output are fundamentally different from cases where users meaningfully and intentionally engage with explanations. Even for more interactive *human-in-the-loop* approaches and *mixed-initiative* systems that aim to increase user interaction and incorporate user feedback to the algorithms [49], increased interaction also has the risk of drawing disproportional attention to model problems and harming the accuracy of the user's mental model of the system [50]. In our study, we investigate whether even differences in explanations of basic technical information—presented independent of observed model outputs and without interactive or meaningful engagement while using the system—might still be enough to influence user perception of the system.

Explanations have been shown to affect a range of human-centered measures with respect to an AI system, including the user's mental model, task performance, satisfaction, and trust and confidence in the system [51]. Explanations have also been shown to help increase system acceptance by preparing users for AI imperfections [35, 52]. Generally-speaking, the effects of explanations on user understanding, trust, or system usage will depend on a number of factors, such as the type of explanation, (e.g., [36, 53, 54]), the amount of explanatory information given (e.g., [55, 10]) and users' domain expertise (e.g., [8, 56, 9]). Our study investigates whether a-priori presentation of technical properties of a model can also influence and has yet to be investigated. In many layperson usage scenarios, users of common AI-powered applications might be expected to have awareness of the reputation of the providing company or the caliber of the resources behind the application.

2.3. The Impact of First Impressions

Psychologists have found that first impressions are formed rapidly, often within the first few minutes of interaction, and these early impressions can have a lasting impact on how humans perceive and engage with new information [57, 58, 59, 60]. In particular, the earliest pieces of information often have a much more prominent influence on our judgments than information received later on, a phenomenon known as first impression bias [61].

Researchers in human-centered AI have employed various experimental manipulations to investigate the effects of first impressions on human attitudes and perceptions of AI systems. These manipulations often involve altering a system component to create different initial perceptions, ranging from perceiving the system as more or less intelligent, reliable, or efficient. Among these manipulations is controlling the order of information presented, such as introducing system errors during the first moments of interaction, potentially leading users to form a negative impression of the system early on and vice versa (e.g., [62, 6]).

For instance, a study involving a humanoid robot [62] demonstrated that a positive first impression has positively affected trust formation even when the user doubts the robot's decision; however, trust in the robot gradually diminishes as the robot makes mistakes. Interestingly, participants were willing to trust the robot in time-sensitive scenarios even when observing robot mistakes beforehand. Nourani et al. [8] investigated how domain expertise affected first impression formation. In an arthropod classification task, researchers varied the order of errors made by the AI system; their findings suggested that only those with domain experience formed first impressions, significantly affecting their trust formation over time. Furthermore, Tolmeijer et al. [63] examined trust development through an intelligent housing recommendation system while manipulating task difficulty. Participants received a variation of accurate or inaccurate advice throughout the task. Their results on the impact of first impressions on trust formation also aligned with previous research.

Extending this research to investigate the interplay between first impressions and user reliance, Nourani et al. [6] conducted an experiment that manipulated two variables, the presentation of strengths or weaknesses and the presence of explanations. Participants experienced either the strengths or the weaknesses of an AI video recognition system early on, with or without explanations. The study found that users who saw system strengths first had a more positive impression of the system, while those who saw weaknesses first, relied more on themselves than on the system. Further, that study also showed that users can become overconfident of their understanding of the system when explanations are given. Another related study [64] showed that the timing of error occurrence significantly affects user reliance. Experiencing system errors early had a lasting negative impact on participants' reliance on the system, whereas later error occurrences affected their reliance momentarily.

2.4. Use of AI in Job Interview Training Systems

Our study of the effects of AI model descriptions was conducted in the context of job interview training. Job interview training systems (JITS) help job seekers practice and improve on the important skill of interviewing. Two types of JITS can be identified. The first type enables the user to practice engaging in a job interview with a simulated recruiter. The second type analyzes aspects of a user's practice or real interview, and delivers feedback to the user. Both types can use AI to support its functioning. The first type of JITS has used technologies such as chatbots [13], androids [65], virtual agents [66], and full virtual reality environments [67]. The second type of JITS rely on techniques such as natural language processing to analyze the user's speech (e.g., [14, 68]) or interview video (e.g., [68, 69]). Our research focuses on the use of AI in the second type of JITS, where a user interview is analyzed and feedback is given.

In much prior work, e.g., [67, 68, 69, 70, 71, 72, 73], JITS have used AI in multi-modal feature extraction from video recording with audio or live camera feed to analyze both verbal and non-verbal cues. AI has been used to analyze interviewees' facial expressions and movement, gaze, voice tone, volume, speed, pauses, body language, posture, emotions, and head position [68, 70, 72], as well as word semantics, pronunciation, vocabulary, grammar, and sentence structure [68, 74]. The AI analysis then typically either results in a score for an interviewee [68, 69], or textual feedback and actionable comments [70, 72, 74].

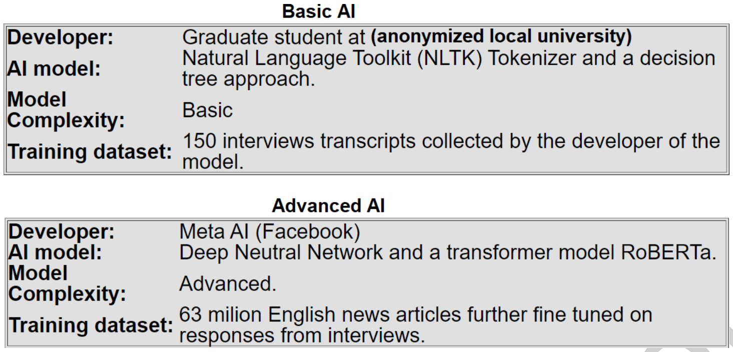
3. Experiment

Our experiment investigates whether first impressions formed from the presentation of introductory information about the AI model affect users' judgment of the system. The experiment compared the presentation of introductory information about AI models in an interview training application that gave feedback about how to improve answering questions during a job interview. Our hypothesis was that users would judge the system more positively if they believed the AI model to be more technologically advanced—even if there were no differences.

3.1. Experimental Design

The study used a between-subjects design with a single independent variable: description of the AI model prior to viewing system output. The study compared two conditions giving different model information—one describing a *basic AI* model, and the other a more *advanced AI* model. The study was conducted as a deception study using a Wizard-of-Oz approach where participants were under the impression that the system was using a real model, but the functionality was actually manually provided by the researchers. The simulated system performance did **not** change based on experimental condition; the *basic* and *advanced* conditions differed only in the given framing of model descriptions.

The information about the two hypothetical AI models varied based on multiple details: developer expertise, algorithm complexity, source of training dataset, and size of training dataset. The actual AI model descriptions used is presented in Figure 1. The *basic AI* condition described the model as having been created by a graduate student developer using a Natural Language Toolkit (NLTK) tokenizer, a decision tree algorithm, and training set built from 150 interview transcripts collected by the graduate student at their university. In the other case (*advanced AI*), the different aspects were presented as a more sophisticated model developed by Meta AI as a Deep Neural Network (DNN) with a transformer model RoBERTa, and trained on 63 million English news articles and further fine tuned on over 100 interview transcripts. In case participants were unfamiliar with terminology, the model descriptions also included an explicit summative indicator of "Model



Basic AI

Developer:	Graduate student at **(anonymized local university)**
AI model:	Natural Language Toolkit (NLTK) Tokenizer and a decision tree approach.
Model Complexity:	Basic
Training dataset:	150 interviews transcripts collected by the developer of the model.

Advanced AI

Developer:	Meta AI (Facebook)
AI model:	Deep Neutral Network and a transformer model RoBERTa.
Model Complexity:	Advanced.
Training dataset:	63 milion English news articles further fine tuned on responses from interviews.

Figure 1: The two experimental conditions presented different information (basic or advanced) about the hypothetical AI model prior to participants viewing system output. While the given information was different, all participants experienced the same level of simulated output provided by an experimenter.

complexity" that was presented as either "Basic" or "Advanced" for the appropriate corresponding condition.

Dependent variables assessed were participants' trust ratings of the system, level of agreement with the system outputs, and willingness to use or apply the given system feedback in future. To obtain a profile of participants, other demographic and psychographic variables were also measured, including self-report ratings of general interview experience, their personality using the Big-5 personality inventory [75], their general propensity to trust machines using the scale from [76], and their general attitudes towards AI using the scale from [77].

3.2. Participants

Study participants were all students from a large university in the United States. The study was completed by 43 participants, though one participant's data was excluded due to a technical error in data capture. This remaining participants with complete data (n = 42) were included as the final sample for analysis, with an equal split of 21 participants in each of the two study conditions. Participant ages ranged from 18 to 39 with a median age of 21.5 years and standard deviation of 5.2. Fifty-one percent of the participants self-reported as men, 44% as women, 3% as other, and 2% preferred not to provide this information.

Participants were recruited from computer science courses over one semester and were offered the option of partial course credit or extra credit for select

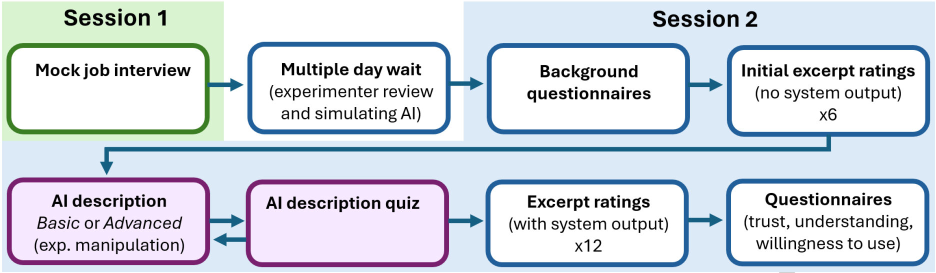

Session 1
Mock job interview
Multiple day wait (experimenter review and simulating AI)
Session 2
Background questionnaires
Initial excerpt ratings (no system output) x6
AI description *Basic* or *Advanced* (exp. manipulation)
AI description quiz
Excerpt ratings (with system output) x12
Questionnaires (trust, understanding, willingness to use)

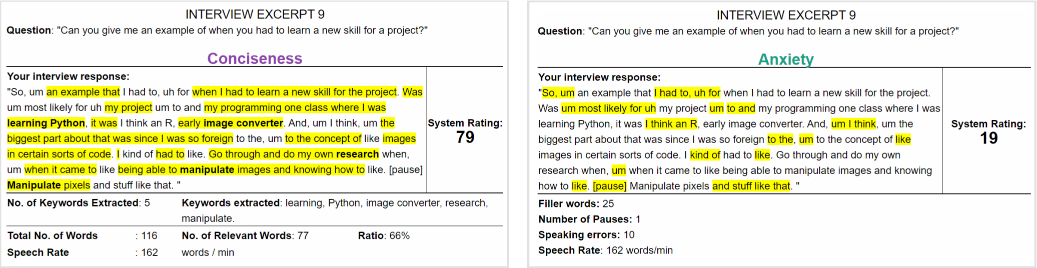
Figure 2: Summary of study procedure. The only experimental manipulation was the presented description of the AI model (purple boxes) in study Session II.

courses (approved by our organizations ethics review board). While all participants were students in technically-oriented courses, all are expected to have some familiarity with AI capabilities, though most would still be considered as novices in AI expertise. From the pre-study questionnaire about experience with AI, 25% reported using AI for less than 1 year, 23% between 1 and 2 years, 16% for 2-3 years, 12% for 3-4 years, 5% for 4 and 5 years, 12% for over 5 years, and 7% have never used AI before. Types of AI experience ranged from simply interacting with AI algorithms or applications on a day-to-day basis to working on implementing an AI or working on deep neural networks.

Participant experience with prior job interviews varied considerably in our sample, with mean of 4.44 and standard deviation of 6.45 for the number of previous times participants have been interviewed. Participants also self-reported their personal estimation of their general interviewing confidence based on five criteria (covering overall confidence, interview skills, behaviors, knowledge of mistakes, and techniques), each using a five-point Likert scale. Confidence results yielded a mean rating of 3.58 and standard deviation of 0.88.

3.3. Procedure and Task

The study consisted of two online sessions and was approved by our organization's ethics board. Participants first conducted a practice job interview in Session I, and then took part in Session II several days later, where they reviewed outcomes and feedback generated by what they believed to be the AI model. Figure 2 summarizes the main stages of the study procedure.



INTERVIEW EXCERPT 9

Question: "Can you give me an example of when you had to learn a new skill for a project?"

Conciseness

Your interview response:
"So, um an example that I had to, uh for when I had to learn a new skill for the project. Was um most likely for uh my project um to and my programming one class where I was **learning Python**, it was I think an R, early **image converter**. And, um I think, um the biggest part about that was since I was so foreign to the, um to the concept of like images in certain sorts of code. I kind of had to like. Go through and do my own **research** when, um when it came to like being able to **manipulate** images and knowing how to like. [pause] **Manipulate** pixels and stuff like that. "

System Rating:
79

No. of Keywords Extracted: 5
Keywords extracted: learning, Python, image converter, research, manipulate.

Total No. of Words : 116
No. of Relevant Words: 77
Ratio: 66%
Speech Rate : 162 words / min

INTERVIEW EXCERPT 9

Question: "Can you give me an example of when you had to learn a new skill for a project?"

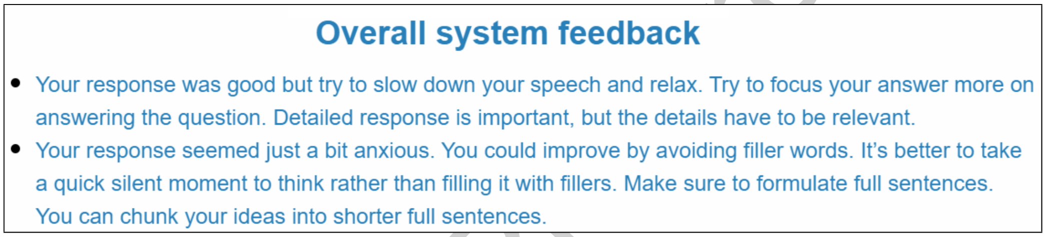
Anxiety

Your interview response:
"So, um an example that I had to, uh for when I had to learn a new skill for the project. Was um most likely for uh my project um to and my programming one class where I was learning Python, it was I think an R, early image converter. And, um I think, um the biggest part about that was since I was so foreign to the, um to the concept of like images in certain sorts of code. I kind of had to like. Go through and do my own research when, um when it came to like being able to manipulate images and knowing how to like. [pause] Manipulate pixels and stuff like that. "

System Rating:
19

Filler words: 25
Number of Pauses: 1
Speaking errors: 10
Speech Rate: 162 words/min

Figure 3: Example of system output presented for an excerpt from the participant's own mock interview. Left: System rating for response conciseness; Right: System rating for anxiety. Participants were also asked to provide their own ratings for their response for both dimensions.



Overall system feedback

- Your response was good but try to slow down your speech and relax. Try to focus your answer more on answering the question. Detailed response is important, but the details have to be relevant.
- Your response seemed just a bit anxious. You could improve by avoiding filler words. It's better to take a quick silent moment to think rather than filling it with fillers. Make sure to formulate full sentences. You can chunk your ideas into shorter full sentences.

Figure 4: An example of the system's output of constructive feedback for interview improvement.

3.3.1. *Study Session I: Mock Job Interview*

Session I consisted of the practice job interview, and was conducted through a synchronous Zoom video call and took approximately 20 minutes. The interview was conducted by a researcher who played the role of a recruiter for the practice interview. Participants were informed that the interview would be recorded. The mock interview was designed to simulate a generic human resource interview; that is, it did not focus on any specific field of study, but rather focused on how a job candidate thinks and responds in different kinds of relevant situations. The interview consisted of 13 questions. Example questions included "Tell me about your educational background and work experience", "Tell me about one time you failed on a project or when something went terribly wrong. How did you handle it or what did you learn from it?", and "How would you define leadership, and do you often take on a leading role in your team projects?". After the interview, participants were informed that they would be sent an email with information for the second study session in a few days.

3.3.2. Simulating AI Feedback

After the mock interview in Session I and before the participant's second session, manual qualitative coding was used to generate system outcomes and feedback to be reported as a Wizard-of-Oz proxy for an AI system. A researcher reviewed the video and transcript, and manually corrected any errors in Zoom-generated transcriptions (including pauses and filler words). Coding was done by a set of three coders (none of whom conducted the interview). Who were specifically trained for the task, and achieved an approximately 82% inter-rater agreement. Participants' responses to each interview question were coded on two dimensions, conciseness and verbal anxiety, and a score was given for each of these dimensions for each interview response.

A conciseness rating assessed the extent to which a participant's response was relevant to the interview question asked. Conciseness scores were measured as the ratio of the number of words used in the response that the researchers considered relevant to the question and the number of total words [78] used in the response. Conciseness scores ranged between 0 and 100 for each particular response, where 0 was not concise at all and 100 was very concise.

In addition, the ratings assessed verbal anxiety as the extent to which a certain response exhibits cues that may be taken as indicative that the respondent is experiencing anxiety. This was operationalized as the ratio of the number of anxiety cues identified in a response and the total number of words used in the response. Anxiety cues consisted of filler words, longer-than-typical pauses and speaking errors [79, 80, 81]. Errors were for example interrupted or unfinished sentences, resetting or breaking with new thoughts, repeating words or stutters. The number of each type of error was summed up and divided by the total number or words to result in a ratio that was then averaged with a reversed speech rate ratio. Based on the two ratios, the coder gave a score between 0 to 100 for verbal anxiety for that particular response, where 0 was not anxious at all and 100 was very anxious.

Once obtained, the scores for conciseness and anxiety for a response were used to decide on the system feedback that would be given to the participant. Participant's scores were mapped with a feedback dictionary (containing predefined comments) to generate actionable advice on how to improve responses or address mistakes during that interview question.

3.3.3. Study Session II: Reviewing System Feedback

The second session was completed online through a web application at the participant's convenience and without researcher supervision. Participants completed the second session 2–12 days after the first session. The delay between sessions gave time for coding the interview responses and generating the simulated AI feedback, and additional time depended on participants' decision of when to complete the study.

Through the study application, participants first provided demographic and psychographic information. Next, instructions were presented to explain to participants what their tasks would be in this session. Participants were asked to assist the system by providing ratings on a subset of textual interview excerpts from their own mock interview. They were instructed that they will see their own answers and have to rate it in terms of consciousness and anxiety on a scale of 0 to 100. Participants reviewed 6 randomly chosen interview questions, presented one at a time, along with the transcript of their corresponding response.

After rating their 6 interview responses, participants were introduced to the AI system, where the description depended on the assigned experimental condition for *basic AI* or *advanced AI* (see Figure 1). To ensure participants gave sufficient attention to the description, participants were required to answer a short quiz on the model details immediately after viewing the description. If any of the quiz questions were answered incorrectly, participants were redirected back to the description, and they had to reattempt the quiz. This repeated until all quiz questions were answered correctly.

In the next phase of the study session, participants reviewed again excerpts of interview questions and their own responses from the mock interview, but this time the system also included the (simulated) AI ratings for conciseness and anxiety. Figure 3 shows examples of the system output for the dimensions of *conciseness* (left) and *anxiety* (right). Both system ratings presented an overall numerical rating (0–100) along with a set of additional explanatory factors (e.g., speech rate, pauses in speech, relevant words, detected keywords). In addition, the output included textual highlighting of relevant tokens in each interview response contributing to the rating.

For each excerpt and system output, participants were also asked to again rate their interview responses on a scale of 0 to 100 in terms of conciseness and anxiety, presented one at a time. Participants reviewed 12 excerpts in this stage of the study, where 6 out of the 12 were the same as the 6 they previously rated prior to viewing the AI output. Having participants rate some of the same responses both before and after seeing the AI description and reviewing the system output made it possible to check whether participants changed their ratings to better align with the AI. The other 6 of the 12 responses rated with system output were included to obscure the fact that participants were repeating the same items while also providing additional data for general human-AI agreement.

After participants rated an interview response for both conciseness and anxiety, the system offered additional suggestions for improving interview responses. This feedback was shown as bullet points with actionable comments for both conciseness and anxiety. Figure 4 shows an example of given feedback. Participants were asked to rate the feedback output, and also optionally to enter free-text comments about the feedback given.

After reviewing and rating all 12 interview excerpts, participants were asked to complete, using a 5-point Likert scale, a measure assessing system trust, capabilities, understanding of the system and AI model description, day-to-day usage of AI, and acceptance of AI in daily life.

3.4. Measures

3.4.1. Human-AI Agreement with Interview Ratings

Because participants both reviewed the systems' output and provided their own rating for each excerpt based on conciseness and anxiety, we can calculate a measure of *human-AI agreement* based on the absolute difference (i.e., disagreement) between the system output and participant rating. For this calculation, the value each system output was the overall numerical 0-100 "system rating" (see Figure 3), and the participant's own 0-100 rating of each interview response. The differences were averaged across all of the 12 excerpts that the participants reviewed and across the two dimensions of *conciseness* and *anxiety* to obtain a single score for each participant. Note that since this measure is based on disagreement, a value of 0 indicates maximum possible agreement, and larger differences indicate lower agreement. *Agreement* provides an estimate of the participant's perception of the quality of the AI.

3.4.2. Agreement with AI's Textual Feedback

As an additional measure of human-AI agreement, we also assessed participants' agreement with the system's textual, constructive feedback for how to improve on interviewing (shown in Figure 4). For this measure, participants were explicitly asked to provide a 0–100 rating ("How much do you agree with the feedback?"). Because the system output for the constructive feedback was only a textual description and did not include a numerical score, agreement could not be calculated as a difference, so the participant's self-reported numerical rating of agreement was used as the measure. Participants provided a rating for each interview excerpt reviewed with AI output in Session II, and ratings were averaged to produce a single numerical measure per participant.

3.4.3. Pre-Post Rating Change

A limitation to the described agreement measures based on the difference between participant rating and system output is the possibility that the participant could ignore the system entirely yet still giving a rating that happens to align with the system result—thus incorrectly indicating high agreement. To account for this limitation, the experiment also captured pre-ratings from the participant. As previously noted, participants first provided ratings for a set of 6 excerpts early in Session II before the model was introduced (see Figure 2). Later in the same session, after the explanation of the model, participants reviewed the system outputs and provided their own ratings for 12 excerpts—half of which were repeats of the same 6 excerpts they rated previously.

For the excerpts that had repeat ratings, the pre-post measure was calculated from responses where the first rating was at least 5 units away from the system score (both ratings on a 0–100 scale). For example, if a pre-rating was 90 and the system score was 94, that case would be considered invalid and excluded from future analyses. This was done because a pre-rating that is too close to the system score could suggest that participants coincidentally already had a notion of their interview responses that matched with the system outputs, and thus any change in their post-rating might not have been necessarily based on the system rating that they saw. Of the valid cases, we calculated the percentage of cases where the participant's post-rating (i.e., the later rating while reviewing the system output) changed from the first rating and moved closer to the system rating.

We also calculated the average difference between post- and pre-scores relative to the system rating across all excerpts that were valid and where the post-rating moved closer to the system rating. The difference scores for conciseness and anxiety were averaged to obtain a single score.

3.4.4. Attitudes

User attitudes in our study encompassed their sense of trust towards the system and their acceptance of the system. To measure self-reported *trust* in the system, we adapted 2 items from the Körber Trust in Automation scale [82] ("I trust the system"; "I can rely on the system"), and another 10 items from Madsen & Gregor's Human-Computer Trust scale [83]: 5 of those items were related to *faith*, 4 related to *perceived technical competence*, and 1 related to *perceived understandability.* All the items were averaged to obtain one trust score for each participant.

The user's *willingness towards the system* consisted of users' willingness to apply the system feedback and willingness to use the system in future. For *willingness to apply system feedback*, participants were asked to answer the question "To what extent will you apply this feedback in future interviews?" using a 0 to 100 response scale. This question was presented after each system feedback. This question presented after each system feedback. The ratings for this question were averaged across all excerpts where feedback was given. For *willingness to use the system*, two items were used: "Using the system would improve my performance in job interviewing" and "Assuming I have access to the system, I would use the system during my interviewing process". These were presented at the end of the study. Ratings for these two items were averaged for each participant.

4. Data Analysis and Results

Based on the bounded scales for ratings and finding that not all distributions were normal, we opted for non-parametric testing for statistical analyses. The Mann Whitney U test was used to compare the *basic* and *advanced* AI model description study conditions, where a separate test was conducted for each of the described measures. None of the background variables assessed, including participants' general confidence in job interviewing, their personality, or their general propensity to trust AI, had any significant effects on the results. The participants' degree of understanding of the AI model description, as measured using a single question on a scale of 1 to 7, also did not seem to affect the results. However, it must be noted that the mean for that variable was relatively high at 5.2 with a standard deviation of 1.04, most likely because of the study sample was recruited from computer science courses. Test statistics for all tests results are given in Table 1.

Measure	Effect	Test Statistic	p-value	Effect Size	Power
Human-AI Agreement	Basic vs. **Advanced** Basic: $MeanRank = 27.07$ Adv.: $MeanRank = 16.69$	z = -2.71	0.007*	$r = 0.42$	-
Feedback Agreement	Basic vs. **Advanced** Basic: $MeanRank = 17.43$ Adv.: $MeanRank = 25.57$	z = -2.15	0.031*	$r = 0.33$	-
Pre-Post rating change	Basic vs. **Advanced** Basic: $MeanRank = 16.12$ Adv.: $MeanRank = 26.88$	z = -2.84	0.004*	$r = 0.44$	-
User trust	Not significant Basic: $MeanRank = 19.93$ Adv.: $MeanRank = 24.17$	z = -1.09	0.269	-	0.74
Willingness to apply feedback	Basic vs. **Advanced** Basic: $MeanRank = 17.71$ Adv.: $MeanRank = 25.29$	z = -2.01	0.044*	$r = 0.31$	-
Willingness to use system	Not significant Basic: $MeanRank = 18.59$ Adv.: $MeanRank = 25.57$	z = -1.81	0.061	-	0.99

Table 1: Summary of Mann Whitney U analyses for study measures. For all significant differences, the *advanced* condition had more positive outcomes than the *basic* condition. Note that the metric for *human-AI agreement* is calculated based on differences in agreement, meaning that lower values indicate higher agreement. Effect sizes are shown for significant effects, calculated as rank-biserial correlation (r), and statistical power is shown for non-significant effects (following the method for Mann Whitney U two-sample tests from [84]).

Graphical summaries of significant results are shown in Figure 5. The colored box shows the interquartile range (IQR) with a horizontal black line for the median. Each vertical "whisker" line extends to the most extreme value falling within an additional $1.5 \times IQR$ beyond the bounds of the IQR, and black dots represent outliers beyond this range. We note that these outliers were *not* excluded for statistical analysis.

Participants showed significantly more *agreement* with the system when the *advanced* AI description was given rather than the *basic* AI description. This was true for both measures of agreement: human-AI agreement based on ratings (Figure 5-A) and by self-report of agreement with feedback (Figure 5-B).

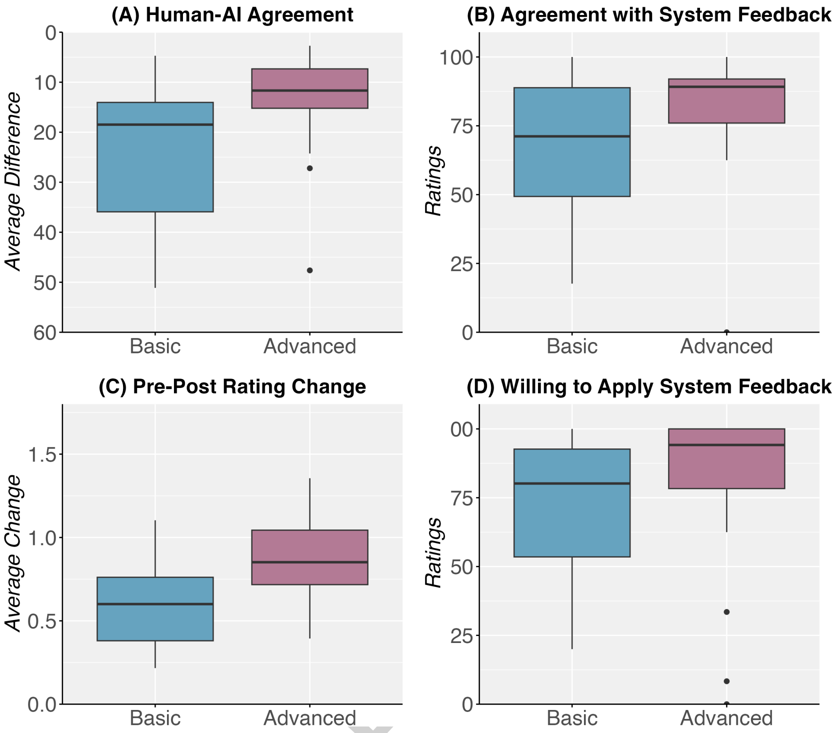


Figure 5: Box-and-whisker plots for significant differences from Table 1. Values higher up on the plot indicate more favorable or more agreement with the system outputs. Note that *Human-AI Agreement* (A) is based on the difference between the system rating and the participant's rating of their own interview responses, while Agreement with System Feedback (B) is based on a self-reported rating of the AI's textual feedback explaining constructive tips for improvement.

Similarly, the test on *pre-post rating changes* showed participants changed their scores to more closely align with the system outputs significantly more in the *advanced AI* description condition (see Figure 5-C).

In terms of user attitudes, results showed that participants were significantly more *willing to apply the given system feedback* when given the description for the *advanced* AI model than the *basic* description (see Figure 5-D).

Finally, the difference between the two conditions in terms of trust, was not statistically significant.

5. Discussion

Describing the AI model as advanced and more sophisticated resulted in significantly higher agreement with the model outputs, more favorable ratings, and greater reported willingness to apply the system output.

5.1. Implications of Results

The study findings demonstrate that a user's impressions of the nature of the model's origin and creation do matter, as the study procedure only manipulated the given framing information but not the nature of the observed system output or quality. The results suggest that if a user believes the technology is more sophisticated, it can significantly shape how users perceive the model abilities, and subsequently, potentially influence the user's judgments accordingly. It is important to repeat that the different conditions for model descriptions did not include any changes in the actual underlying system or how output was generated; the only difference was in the instructions at the start of the study Session II (see Figure 2). In the *advanced AI model description* condition, participants' ratings aligned more closely with the system's ratings than the basic condition, which suggests that the user's perception of interacting with a more impressive AI model is sufficient to elicit a higher level of agreement with the system's decisions. The *advanced AI model description* condition also showed a significantly higher average difference in pre-post ratings in participants' responses than the basic condition, indicating that the system's decision had a stronger influence on making them more willing to adjust their ratings to align with the system's—even if their earlier ratings prior to viewing model output were substantially different.

These results have important implications for designing AI systems in educational and training contexts. If ethical requirements dictate including details describing the employed AI model, designers should carefully consider the effect of these descriptions on users' behavior and the perceived value of the system. As shown in other studies, deciding when and how to show information to users can significantly impact their first impression formation [64, 85]. Distinguishing from prior studies, our novel findings demonstrate that initial information given about the model is enough to have an impact, and biased impressions can be developed even before seeing any demonstration of system capabilities.

Regarding their willingness to apply the system feedback, participants' responses in the advanced description condition indicated a more positive attitude towards the system. These results can perhaps be attributed to the perceived reliability of an advanced AI model, which likely instilled a perception that the feedback was more valid even though the conditions did not differ in the method or quality of feedback given. The research findings align with existing research on the impact of first impressions on user perceptions and the mental model of the systems, highlighting the importance of initial presentation in shaping user attitudes and acceptance of AI systems [8, 63, 62].

The study did not find significant differences for user trust and participants' willingness to use the system in the future. We note that prior work has sought to distinguish trust from reliance. Scharowski et al. [86], for example, emphasized that trust is an attitude whereas reliance is a behavior. Hoffman et al. [87] also stated that the difference between the two can be understood as trusting a machine as opposed to following its advice. In that sense, prior work see trust and reliance as being "conceptually distinct" and as not sharing a "deterministic but a probabilistic relationship" [86]. Participants' willingness to apply the system feedback, which was significantly different between the conditions in our study, is more aligned with user reliance, and their willingness to use the system in future, which was not significantly different, is more related to trust. Hence it may be sensible that the former is not affected by AI model description, but the latter is.

5.2. Implications for Design and Applications

The results also highlight the awareness of potential ethical concerns about how AI-powered applications are presented in the real world, as the findings suggest the possibility that end users might be inclined to perceive a system as being "better" or more accurate simply due to the origin of the technology. For real applications, developers or companies might have an incentive to overclaim the value of impressive credentials or technical complexity of AI for the purposes of influencing more positive impressions of the technology. Of course, such implications would be counter to preferred practices for transparency. In addition, if users give weight to the origin of a model when judging AI, the findings indicate that large and established companies might have an unfair advantage for acceptance of AI technology over smaller organizations—regardless of the objective capabilities of the AI.

Framing effects also have implications for research studies. In particular, for human-subjects studies adopting low-fidelity prototype, Wizard-of-Oz implementations, or manually-crafted, simulated AI models in place of actual complete AI model, the results of our study suggest that the nature of specific information given to study participants about the AI model might affect study results, and there may be potential confounds with other experimental manipulations in cases where details of the actual model (real or synthetic) does not accurately match presented details. Further, since we have evidence that differences in presenting technical details can influence perception of AI capabilities, this raises ethical concerns about participant deception if details are not presented sufficiently and accurately. Awareness of the importance of presentation differences in research studies further demonstrates the complexity of the design space for possible contextual factors that can have substantial effects on user thinking and behaviors.

From a domain perspective, our work provides guidance for the design of AI-based job interview training systems (JITS). The use of AI in these kinds of systems is becoming increasingly common. The study results indicate that it is important for designers to pay attention to how the AI is described to users prior to their use of the system since the description can affect what they take away from the system eventually, hence the system's usefulness to its users. Oftentimes, JITS and many other similar systems present little to no information about the AI used. However, as noted in the introduction, since it is unlikely that users will have extensive or sustained interactions with systems like JITS, it may perhaps be beneficial to optimize the potential for system output uptake through simple design features like the presentation of AI model information.

5.3. Limitations and Future Work

This research raises important questions about how to present model transparency and explanations. Generally, most people would argue for providing more descriptive information about an AI model, when possible. However, the results of our study also highlight potential risks of biasing user judgment of AI systems regardless of the actual underlying system capabilities or model accuracy. Due to the wizard-of-oz nature of simulating model output for our experiment, we are confident all participants observed reasonably high-quality output from the system. This does leave open questions about whether the effect of model descriptions might change based on differences in model quality or performance.

Other factors such as user demographics and familiarity with the topic or domain receiving AI support could have influenced the results. For example, Nourani et al. [8] found that differences in user expertise made a significant difference on how participants were influenced by early impressions with model errors. Since our experiment was done in the context of evaluating responses in job interviews (with common non-domain specific, human resources questions), the participants likely had the ability to judge the input on their own. Future work could expand the research to include the study of effects of different model descriptions in cases with varying levels of alignment between user knowledge and model decision making. For example, a direct extension for our study design might include interviews involving challenging technical questions.

It would be interesting to consider how important users' technical knowledge is for susceptibility to biasing effects with AI models. Especially for model descriptions, we might expect the effect to only exist when users have enough technical knowledge for the description to matter, as was ensured by participant recruitment from university computing courses for our study. We would encourage expanded research on this topic to consider a broader range of technical knowledge as well as other demographic variability.

Acknowledgement

This work was funded by NSF (grant number 1956087, 2020-2023). Collaborative Research: CHS: Medium: Bio-behavioral data analytics to enable personalized training of veterans for the future workforce. .

References

[1] B. Leichtmann, A. Hinterreiter, C. Humer, M. Streit, M. Mara, Explainable artificial intelligence improves human decision-making: Results from a mushroom picking experiment at a public art festival, International Journal of Human–Computer Interaction (2023) 1–18.

[2] C. J. Cai, S. Winter, D. Steiner, L. Wilcox, M. Terry, " hello ai": uncovering the onboarding needs of medical practitioners for human-ai collaborative decision-making, Proceedings of the ACM on Human-computer Interaction 3 (CSCW) (2019) 1–24.

[3] U. Ehsan, S. Passi, Q. V. Liao, L. Chan, I. Lee, M. Muller, M. O. Riedl, et al., The who in explainable AI: How AI background shapes perceptions of AI explanations, arXiv preprint arXiv:2107.13509 (2021).

[4] F. M. Megahed, Y.-J. Chen, J. A. Ferris, S. Knoth, L. A. Jones-Farmer, How generative AI models such as chatgpt can be (mis) used in spc practice, education, and research? an exploratory study, Quality Engineering (2023) 1–29.

[5] A. Choudhury, H. Shamszare, Investigating the impact of user trust on the adoption and use of chatgpt: Survey analysis, Journal of Medical Internet Research 25 (2023) e47184.

[6] M. Nourani, C. Roy, J. E. Block, D. R. Honeycutt, T. Rahman, E. D. Ragan, V. Gogate, On the importance of user backgrounds and impressions: Lessons learned from interactive ai applications, ACM Transactions on Interactive Intelligent Systems 12 (4) (2022) 1–29.

[7] B. J. Dietvorst, J. P. Simmons, C. Massey, Algorithm aversion: people erroneously avoid algorithms after seeing them err., Journal of experimental psychology: General 144 (1) (2015) 114.

[8] M. Nourani, J. King, E. Ragan, The role of domain expertise in user trust and the impact of first impressions with intelligent systems, in: Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, Vol. 8, 2020, pp. 112–121.

[9] M. Nourani, A. Hashky, E. D. Ragan, User profiling in human-ai design: an empirical case study of anchoring bias, individual differences, and ai attitudes, in: Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, Vol. 12, 2024, pp. 137–146.

[10] R. Linder, S. Mohseni, F. Yang, S. K. Pentyala, E. D. Ragan, X. B. Hu, How level of explanation detail affects human performance in interpretable intelligent systems: A study on explainable fact checking, Applied AI Letters 2 (4) (2021) e49.

[11] K. Reinecke, T. Yeh, L. Miratrix, R. Mardiko, Y. Zhao, J. Liu, K. Z. Gajos, Predicting users' first impressions of website aesthetics with a quantification of perceived visual complexity and colorfulness, in: Proceedings of the SIGCHI conference on human factors in computing systems, 2013, pp. 2049–2058.

[12] S. Madan, K. Savani, G. V. Johar, Over-reliance on aesthetics? the appearance-reveals-character lay theory increases consumers' devaluation of unattractive produce, International Journal of Research in Marketing (2025).

[13] N. Boudjani, V. Colas, C. Joubert, D. B. Amor, Ai chatbot for job interview, in: 2023 46th MIPRO ICT and Electronics Convention (MIPRO), IEEE, 2023, pp. 1155–1160.

[14] R. Verrap, E. Nirjhar, A. Nenkova, T. Chaspari, "am i answering my job interview questions right?": A NLP approach to predict degree of explanation in job interview responses, in: Proceedings of the Second Workshop on NLP for Positive Impact (NLP4PI), 2022, pp. 122–129.

[15] R. Parasuraman, V. Riley, Humans and automation: Use, misuse, disuse, abuse, Human factors 39 (2) (1997) 230–253.

[16] V. Riley, Operator reliance on automation: Theory and data, in: Automation and human performance, CRC Press, 2018, pp. 19–35.

[17] M. Nourani, C. Roy, T. Rahman, E. D. Ragan, N. Ruozzi, V. Gogate, Don't explain without verifying veracity: an evaluation of explainable ai with video activity recognition, arXiv preprint arXiv:2005.02335 (2020).

[18] M. L. Cummings, Designing decision support systems for revolutionary command and control domains, University of Virginia, 2004.

[19] J. D. Lee, T. F. Sanquist, Augmenting the operator function model with cognitive operations: Assessing the cognitive demands of technological innovation in ship navigation, IEEE Transactions on Systems, Man, and Cybernetics-Part A: Systems and Humans 30 (3) (2002) 273–285.

[20] L. J. Skitka, K. L. Mosier, M. Burdick, Does automation bias decision-making?, International Journal of Human-Computer Studies 51 (5) (1999) 991–1006.

[21] R. P. Will, True and false dependence on technology: Evaluation with an expert system, Computers in human behavior 7 (3) (1991) 171–183.

[22] M. T. Dzindolet, S. A. Peterson, R. A. Pomranky, L. G. Pierce, H. P. Beck, The role of trust in automation reliance, International journal of human-computer studies 58 (6) (2003) 697–718.

[23] J. D. Lee, K. A. See, Trust in automation: Designing for appropriate reliance, Human factors 46 (1) (2004) 50–80.

[24] D. Gunning, M. Stefik, J. Choi, T. Miller, S. Stumpf, G.-Z. Yang, Xai—explainable artificial intelligence, Science robotics 4 (37) (2019) eaay7120.

[25] G. Schwalbe, B. Finzel, A comprehensive taxonomy for explainable artificial intelligence: a systematic survey of surveys on methods and concepts, Data Mining and Knowledge Discovery 38 (5) (2024) 3043–3101.

[26] A. Bennetot, I. Donadello, A. El Qadi El Haouari, M. Dragoni, T. Frossard, B. Wagner, A. Sarranti, S. Tulli, M. Trocan, R. Chatila, et al., A practical tutorial on explainable ai techniques, ACM Computing Surveys 57 (2) (2024) 1–44.

[27] G. Bansal, T. Wu, J. Zhou, R. Fok, B. Nushi, E. Kamar, M. T. Ribeiro, D. Weld, 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, 2021, pp. 1–16.

[28] Z. Buçinca, M. B. Malaya, K. Z. Gajos, 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) (2021) 1–21.

[29] H. Vasconcelos, M. Jörke, M. Grunde-McLaughlin, T. Gerstenberg, M. S. Bernstein, R. Krishna, Explanations can reduce overreliance on ai systems during decision-making, Proceedings of the ACM on Human-Computer Interaction 7 (CSCW1) (2023) 1–38.

[30] S. Mohseni, N. Zarei, E. D. Ragan, A multidisciplinary survey and framework for design and evaluation of explainable ai systems, ACM Transactions on Interactive Intelligent Systems (TiiS) 11 (3-4) (2021) 1–45.

[31] M. Raees, I. Meijerink, I. Lykourentzou, V.-J. Khan, K. Papangelis, From explainable to interactive ai: A literature review on current trends in human-ai interaction, International Journal of Human-Computer Studies 189 (2024) 103301.

[32] M. Radensky, D. Downey, K. Lo, Z. Popovic, D. S. Weld, Exploring the role of local and global explanations in recommender systems, in: CHI Conference on Human Factors in Computing Systems Extended Abstracts, 2022, pp. 1–7.

[33] A. Angerschmid, J. Zhou, K. Theuermann, F. Chen, A. Holzinger, Fairness and explanation in AI-informed decision making, Machine Learning and Knowledge Extraction 4 (2) (2022) 556–579.

[34] V. Arya, R. K. Bellamy, P.-Y. Chen, A. Dhurandhar, M. Hind, S. C. Hoffman, S. Houde, Q. V. Liao, R. Luss, A. Mojsilović, et al., One explanation does not fit all: A toolkit and taxonomy of AI explainability techniques, arXiv preprint arXiv:1909.03012 (2019).

[35] M. Riveiro, S. Thill, "that's (not) the output i expected!" on the role of end user expectations in creating explanations of AI systems, Artificial Intelligence 298 (2021) 103507.

[36] X. Wang, M. Yin, Are explanations helpful? a comparative study of the effects of explanations in ai-assisted decision-making, in: 26th international conference on intelligent user interfaces, 2021, pp. 318–328.

[37] E. M. Kenny, C. Ford, M. Quinn, M. T. Keane, Explaining black-box classifiers using post-hoc explanations-by-example: The effect of explanations and error-rates in XAI user studies, Artificial Intelligence 294 (2021) 103459.

[38] C. Chen, A. D. Tian, R. Jiang, When post hoc explanation knocks: Consumer responses to explainable AI recommendations, Journal of Interactive Marketing (2023) 10949968231200221.

[39] B. Y. Lim, A. K. Dey, D. Avrahami, Why and why not explanations improve the intelligibility of context-aware intelligent systems, in: Proceedings of the SIGCHI conference on human factors in computing systems, 2009, pp. 2119–2128.

[40] A. B. Arrieta, N. Díaz-Rodríguez, J. Del Ser, A. Bennetot, S. Tabik, A. Barbado, S. García, S. Gil-López, D. Molina, R. Benjamins, et al., Explainable artificial intelligence (xai): Concepts, taxonomies, opportunities and challenges toward responsible ai, Information fusion 58 (2020) 82–115.

[41] R. Dwivedi, D. Dave, H. Naik, S. Singhal, R. Omer, P. Patel, B. Qian, Z. Wen, T. Shah, G. Morgan, et al., Explainable ai (xai): Core ideas, techniques, and solutions, ACM computing surveys 55 (9) (2023) 1–33.

[42] A. Adadi, M. Berrada, Peeking inside the black-box: a survey on explainable artificial intelligence (xai), IEEE access 6 (2018) 52138–52160.

[43] A. Suh, I. Hurley, N. Smith, H. C. Siu, Fewer than 1% of explainable ai papers validate explainability with humans, in: Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems, 2025, pp. 1–7.

[44] T. Miller, Explanation in artificial intelligence: Insights from the social sciences, Artificial Intelligence 267 (2019) 1–38.

[45] B. Dietvorst, People reject (superior) algorithms because they compare them to counter-normative reference points, Available at SSRN 2881503 (2016).

[46] M. Nourani, S. Kabir, S. Mohseni, E. D. Ragan, The effects of meaningful and meaningless explanations on trust and perceived system accuracy in intelligent systems, in: Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, Vol. 7, 2019, pp. 97–105.

[47] R. R. Hoffman, M. Johnson, J. M. Bradshaw, A. Underbrink, Trust in automation, IEEE Intelligent Systems 28 (1) (2013) 84–88.

[48] S. Mohseni, F. Yang, S. Pentyala, M. Du, Y. Liu, N. Lupfer, X. Hu, S. Ji, E. Ragan, Trust evolution over time in explainable AI for fake news detection, in: Workshop on Human-Centered Approaches to Fair and Responsible AI, 2020, pp. 1–4.

[49] E. Horvitz, Principles of mixed-initiative user interfaces, in: Proceedings of the SIGCHI conference on Human Factors in Computing Systems, 1999, pp. 159–166.

[50] D. Honeycutt, M. Nourani, E. Ragan, Soliciting human-in-the-loop user feedback for interactive machine learning reduces user trust and impressions of model accuracy, in: Proceedings of the AAAI Conference on Human Computation and Crowdsourcing, Vol. 8, 2020, pp. 63–72.

[51] A. Vultureanu-Albişi, C. Bădică, A survey on effects of adding explanations to recommender systems, Concurrency and Computation: Practice and Experience 34 (20) (2022) e6834.

[52] R. Kocielnik, S. Amershi, P. N. Bennett, Will you accept an imperfect ai? exploring designs for adjusting end-user expectations of AI systems, in: Proceedings of the 2019 CHI Conference on Human Factors in Computing Systems, 2019, pp. 1–14.

[53] B. Lavender, S. Abuhaimed, S. Sen, Effects of explanation types on user satisfaction and performance in human-agent teams, International Journal on Artificial Intelligence Tools 33 (03) (2024) 2460004.

[54] M. Vered, T. Livni, P. D. L. Howe, T. Miller, L. Sonenberg, The effects of explanations on automation bias, Artificial Intelligence 322 (2023) 103952.

[55] T. Kulesza, S. Stumpf, M. Burnett, S. Yang, I. Kwan, W.-K. Wong, Too much, too little, or just right? ways explanations impact end users' mental models, in: 2013 IEEE Symposium on visual languages and human centric computing, IEEE, 2013, pp. 3–10.

[56] M. Szymanski, M. Millecamp, K. Verbert, Visual, textual or hybrid: the effect of user expertise on different explanations, in: 26th international conference on intelligent user interfaces, 2021, pp. 109–119.

[57] O. Kostopoulou, M. Sirota, T. Round, S. Samaranayaka, B. C. Delaney, The Role of Physicians' First Impressions in the Diagnosis of Possible Cancers without Alarm Symptoms, Medical Decision Making 37 (1) (2017) 9–16. doi:10.1177/0272989X16644563.

[58] W. E. Remus, J. E. Kottemann, Toward intelligent decision support systems: An artificially intelligent statistician, Mis Quarterly (1986) 403–418Publisher: JSTOR.
URL https://www.jstor.org/stable/249197

[59] B. W. Swider, T. B. Harris, Q. Gong, First impression effects in organizational psychology., Journal of Applied Psychology 107 (3) (2022) 346.

[60] P. E. Tetlock, Accountability and the perseverance of first impressions, Social psychology quarterly (1983) 285–292.

[61] S. E. Asch, Forming impressions of personality, The Journal of Abnormal and Social Psychology 41 (3) (1946) 258–290, place: US Publisher: American Psychological Association. doi:10.1037/h0055756.

[62] S. Xu, X. Jiang, I. J. Walsh, The influence of openness to experience on perceived employee creativity: The moderating roles of individual trust, The Journal of Creative Behavior 52 (2) (2018) 142–155.

[63] S. Tolmeijer, U. Gadiraju, R. Ghantasala, A. Gupta, A. Bernstein, Second chance for a first impression? trust development in intelligent system interaction, in: Proceedings of the 29th ACM Conference on user modeling, adaptation and personalization, 2021, pp. 77–87.

[64] A. Kim, M. Yang, J. Zhang, When algorithms err: Differential impact of early vs. late errors on users' reliance on algorithms, ACM Transactions on Computer-Human Interaction 30 (1) (2023) 1–36.

[65] K. Inoue, K. Hara, D. Lala, S. Nakamura, K. Takanashi, T. Kawahara, A job interview dialogue system with autonomous android erica, in: Increasing Naturalness and Flexibility in Spoken Dialogue Interaction: 10th International Workshop on Spoken Dialogue Systems, Springer, 2021, pp. 291–297.

[66] K. Anderson, E. André, T. Baur, S. Bernardini, M. Chollet, E. Chryssafidou, I. Damian, C. Ennis, A. Egges, P. Gebhard, et al., The tardis framework: intelligent virtual agents for social coaching in job interviews, in: International conference on advances in computer entertainment technology, Springer, 2013, pp. 476–491.

[67] I. Stanica, M.-I. Dascalu, C. N. Bodea, A. D. B. Moldoveanu, VR job interview simulator: where virtual reality meets artificial intelligence for education, in: 2018 Zooming innovation in consumer technologies conference (ZINC), IEEE, 2018, pp. 9–12.

[68] B. Lee, B. Kim, Development of an AI-based interview system for remote hiring, International Journal of Advanced Research in Engineering and Technology (IJARET) 12 (3) (2021) 654–663.

[69] Y.-C. Chou, F. R. Wongso, C.-Y. Chao, H.-Y. Yu, An AI mock-interview platform for interview performance analysis, in: 2022 10th International Conference on Information and Education Technology (ICIET), IEEE, 2022, pp. 37–41.

[70] A. Heimerl, S. Mertes, T. Schneeberger, T. Baur, A. Liu, L. Becker, N. Rohleder, P. Gebhard, E. André, Generating personalized behavioral feedback for a virtual job interview training system through adversarial learning, in: International Conference on Artificial Intelligence in Education, Springer, 2022, pp. 679–684.

[71] I.-C. Stănică, F. Moldoveanu, Considerations for a virtual training system to improve job interview skills for software engineers, Romanian Conference on Human-Computer Interaction (2017).

[72] N. Takeuchi, T. Koda, Initial assessment of job interview training system using multimodal behavior analysis, in: Proceedings of the 9th International Conference on Human-Agent Interaction, 2021, pp. 407–411.

[73] K. Yadav, A. Seemendra, A. Singhania, S. Bora, P. Dubey, V. Aggarwal, Interviewing the interviewer: Ai-generated insights to help conduct candidate-centric interviews, in: Proceedings of the 28th International Conference on Intelligent User Interfaces, 2023, pp. 723–736.

[74] D. R. Pertiwi, M. A. D. Kusumaningrum, Developing english job interview skill using artificial intelligence technology, KACANEGARA Jurnal Pengabdian pada Masyarakat 4 (2) (2021) 221–228.

[75] B. Rammstedt, O. P. John, Measuring personality in one minute or less: A 10-item short version of the big five inventory in english and german, Journal of research in Personality 41 (1) (2007) 203–212.

[76] R. Haydon, Trust in artificial intelligence: How personality and risk experience affect human-AI relationships, Ph.D. thesis, San Francisco State University (2020).

[77] A. Schepman, P. Rodway, Initial validation of the general attitudes towards artificial intelligence scale, Computers in human behavior reports 1 (2020) 100014.

[78] C. K. Sigelman, S. F. Elias, P. Danker-Brown, Interview behaviors of mentally retarded adults as predictors of employability., Journal of Applied Psychology 65 (1) (1980) 67.

[79] A. S. Dibner, Cue-counting: A measure of anxiety in interviews., Journal of Consulting Psychology 20 (6) (1956) 475.

[80] A. R. Feiler, D. M. Powell, Behavioral expression of job interview anxiety, Journal of Business and Psychology 31 (2016) 155–171.

[81] R. Miller, B. Gayfer, D. Powell, Influence of vocal and verbal cues on ratings of interview anxiety and interview performance. personnel assessment and decisions, 4 (2), 26-41 (2018).

[82] M. Körber, Theoretical considerations and development of a questionnaire to measure trust in automation, in: Proceedings of the 20th Congress of the International Ergonomics Association (IEA 2018) Volume VI: Transport Ergonomics and Human Factors (TEHF), Aerospace Human Factors and Ergonomics 20, Springer, 2019, pp. 13–30.

[83] M. Madsen, S. Gregor, Measuring human-computer trust, in: 11th australasian conference on information systems, Vol. 53, Citeseer, 2000, pp. 6–8.

[84] G. E. Noether, Sample size determination for some common nonparametric tests, Journal of the American Statistical Association 82 (398) (1987) 645–647.

[85] M. Nourani, C. Roy, J. E. Block, D. R. Honeycutt, T. Rahman, E. Ragan, V. Gogate, Anchoring bias affects mental model formation and user reliance in explainable AI systems, in: 26th International Conference on Intelligent User Interfaces, 2021, pp. 340–350.

[86] N. Scharowski, S. A. Perrig, M. Svab, K. Opwis, F. Brühlmann, Exploring the effects of human-centered ai explanations on trust and reliance, Frontiers in Computer Science 5 (2023) 1151150.

[87] R. R. Hoffman, S. T. Mueller, G. Klein, J. Litman, Metrics for explainable ai: Challenges and prospects, arXiv preprint arXiv:1812.04608 (2018).

Declaration of interests

☒ The authors declare that they have no known competing financial interests or personal relationships that could have appeared to influence the work reported in this paper.

☐ The authors declare the following financial interests/personal relationships which may be considered as potential competing interests: