Skin-Deep Bias: How Avatar Appearances Shape Perceptions of AI Hiring

Ka Hei Carrie Lau
carrie.lau@tum.de
Chair of Human-Centered Technologies for Learning
Munich Center for Machine Learning (MCML)
Technical University of Munich
Munich, Germany

Philipp Stark
philipp.stark@keg.lu.se
Department of Human Geography, Lund University
Lund, Sweden

Efe Bozkir
efe.bozkir@tum.de
Chair of Human-Centered Technologies for Learning,
Technical University of Munich
Munich, Germany

Enkelejda Kasneci
enkelejda.kasneci@tum.de
Chair of Human-Centered Technologies for Learning,
Technical University of Munich
Munich, Germany

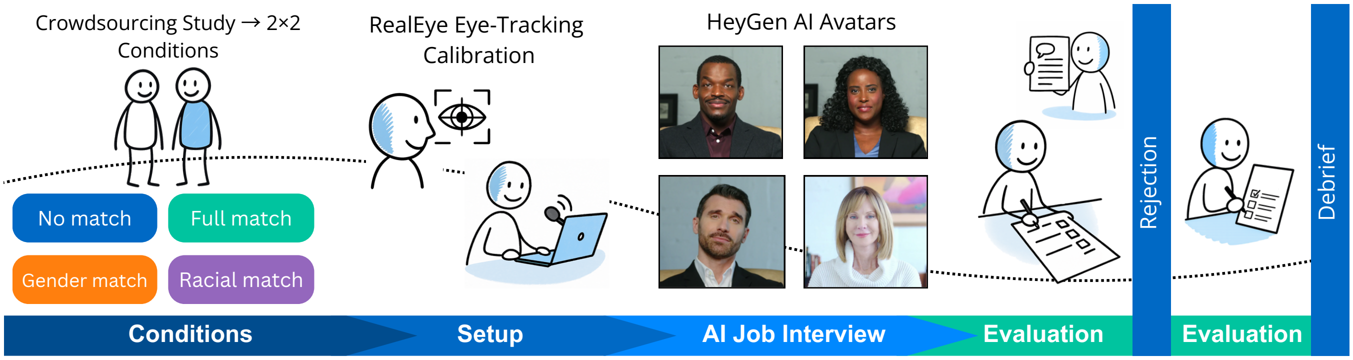


Figure 1: Study overview. We recruited participants via crowdsourcing and assigned them to a 2×2 experimental design with four avatar–participant match conditions (No Match, Gender Match, Racial Match, Full Match). The experiment procedure included eye-tracking calibration, a real-time verbal AI interview, post-interview measures, a scripted rejection, post-outcome measures, and a debrief. We created some illustrative icons with GPT-5 and then the author edited them to align with the study design and visual style.

Abstract

Artificial intelligence is increasingly used in hiring, raising concerns about how applicants perceive these systems. While prior work on algorithmic fairness has emphasized technical bias mitigation, little is known about how avatar identity cues influence applicants' justice attributions in an interview context. We conducted a crowdsourcing study with 215 participants who completed an interview with photorealistic AI avatars varied in phenotypic traits (race and sex), followed by a standardized rejection. Using self-reports, sentiment analysis, and eye tracking, we measured perceptions of trust, fairness, and bias. Results show that racial mismatch heightened perceptions of ethnic bias, while partial match (sharing only one identity) reduced fairness judgments compared to both full and no match. This work extends the Computers-Are-Social-Actors paradigm by demonstrating that avatar appearances shape justice-related evaluations of AI. We contribute to HCI by revealing how identity cues influence fairness attributions and offer actionable insights for designing equitable AI interview systems.

CCS Concepts

• **Human-centered computing** → **Empirical studies in HCI**; **Empirical studies in collaborative and social computing**.

Keywords

generative AI, crowdsourcing, social identity, fairness

ACM Reference Format:
Ka Hei Carrie Lau, Philipp Stark, Efe Bozkir, and Enkelejda Kasneci. 2026. Skin-Deep Bias: How Avatar Appearances Shape Perceptions of AI Hiring. In *Proceedings of the 2026 CHI Conference on Human Factors in Computing*

CHI '26, Barcelona, Spain
© 2026 Copyright held by the owner/author(s).
ACM ISBN 979-8-4007-2278-3/2026/04
https://doi.org/10.1145/3772318.3790379

Systems (CHI '26), April 13–17, 2026, Barcelona, Spain. ACM, New York, NY, USA, 20 pages. https://doi.org/10.1145/3772318.3790379

1 Introduction

Artificial intelligence (AI) systems are increasingly deployed in high-stakes contexts such as college admissions or hiring interviews [24, 66, 82]. Many of these tools use embodied conversational agents (ECAs) with human-like voices and appearances to simulate face-to-face interaction. Equipped with large language models (LLMs) as their backend, ECAs can now respond adaptively to users in real time. Their efficiency and scalability have accelerated adoption in recruitment, where companies seek faster and more standardized processes [15]. AI hiring platforms are often promoted as objective and fair, with claims that they can reduce or even eliminate human bias in evaluation [31, 52]. However, it remains unclear how applicants perceive issues of fairness and trust when engaging with these systems.

This uncertainty around users' perceptions points to an already established paradox in human–computer interaction (HCI). While realistic interfaces are designed to foster trust by making interactions feel natural, in high-stakes contexts, this same realism can amplify disappointment and perceptions of unfairness when outcomes are unfavorable. Prior research has examined this paradox from two perspectives. On the one hand, Wang et al. [83] show that fairness perceptions depend not only on group-level bias but also on whether individuals personally benefit or lose from an algorithmic decision. On the other hand, Nass and Moon [50] describe human interaction with realistic interfaces as "mindless social responses": people do not thoughtfully anthropomorphize machines but instead apply social scripts automatically. These perspectives suggest that fairness perceptions are shaped by both the outcomes of algorithmic decisions and automatic, unreflective social responses. In this work, we focus on the latter. These reactions become more consequential when lifelike avatars clearly display social category cues, such as race and gender, which can lead people to automatically apply social stereotypes to them [51, 62]. If these systems behave in ways that echo offline discrimination, users can interpret this as algorithmic unfairness and a breach of trust [86]. We extend this line of research to embodied contexts, where avatars' phenotypic traits, such as race and sex, can act as salient social cues. We use the term **phenotypic bias transfer** to describe how avatars' **phenotypic traits** can trigger social categorization processes that reproduce racialized and gendered stereotypes during interaction. These phenotypic traits are visible features such as skin color and secondary sex characteristics (e.g., facial structure, jawline, hairstyle). In this paper, we use *sex* to refer to avatars' phenotypic appearance (female/male) and *gender* to refer to participants' self-identified category (woman/man). Accordingly, we use the term **gender match** to denote when an avatar's sex aligns with the participant's self-identified gender, and **racial match** to denote when an avatar's race aligns with the participant's self-identified ethnicity.

Investigating these potential perceptual biases is urgent, since conversational and generative artificial intelligence (Gen-AI) ECAs are already being rapidly adopted in socially impactful domains such as customer service and recruitment [22, 65, 77]. Yet research still lags behind, particularly in understanding how people perceive fairness and trust in these systems and how, beyond questions of algorithmic fairness, they may harm racially and culturally marginalized groups [27, 69]. Beyond this societal urgency, our study also offers a theoretical contribution. Social Identity Theory (SIT) predicts that people show in-group favoritism, evaluating those who share their identity more positively [79], while the Computers Are Social Actors (CASA) paradigm shows that people apply human social rules to machines [51]. As a result, these perspectives suggest that AI systems can elicit social behaviors comparable to those observed in human–human interaction, which in turn may introduce biases. It is therefore important to examine how this unfolds in simulated high-pressure contexts such as AI hiring, where adaptive ECAs are used and the outcome is unfavorable. Methodologically, most previous studies rely on scripted or pre-recorded agents, limiting adaptive and interactive capabilities, and to the best of our knowledge, no prior work has examined real-time Gen-AI-based ECAs in the context of simulated job interviews.

Building on these considerations, our study analyzes perceptions of trust, fairness, and bias along the phenotypic categories of race and sex, and examines the interaction between these two categories. We treat these effects as structural rather than incidental, in line with Schlesinger et al. [69], who argue that treating identity categories in isolation overlooks how technologies reinforce overlapping systems of oppression. Furthermore, interactional designs can reproduce existing social hierarchies, and Zajko [87] shows that algorithmic systems embed and reproduce broader structural inequalities. Together, these perspectives highlight that perceived fairness problems in AI extend beyond algorithmic justice. Systems may satisfy procedural fairness criteria yet still create unequal experiences through avatar **phenotypic traits** that trigger biased responses. Closest to our work, Biswas et al. [6] varied avatar race and gender in pre-recorded asynchronous video interviews (AVI); we extend this line of work by examining perceived fairness as a relational judgment in a real-time interview setting. Specifically, we ask the following research questions (RQs):

RQ1. Do participants perceive meaningful differences between avatar phenotypic traits?

RQ2. Does racial or gender matching affect trust, perceived fairness, and bias?

RQ3. Does racial or gender matching affect implicit behavioral measures (sentiment and eye tracking)?

To investigate this problem, we developed an online platform that simulates real-time **AI interviews** using a Gen-AI-ECA powered by the HeyGen avatar generator. In a crowdsourcing study with 215 participants, we employed a 2 × 2 between-subjects design that manipulated avatar **phenotypic traits** (race: black/white; sex: male/female) to either match or mismatch participants' own identities. Each participant completed the interview verbally and received a standardized rejection outcome, providing a context that approximates applied AI-mediated interviews, while controlling for *outcome favorability* bias shown by Wang et al. [83]. We collected both self-report measures (fairness, trust, bias) and implicit behavioral data (sentiment, webcam-based eye tracking), which provide a complementary perspective on how participants engaged with the AI interviewer. Our findings show that post-interview trust was high across conditions, while perceptions of fairness and bias shifted. Racial mismatches increased perceived bias and attention to the avatar's face, while partial matches lowered perceived fairness.

Our work makes three contributions to HCI and AI fairness research. We follow the taxonomy of research contributions in HCI [85] and build on recent discussions of LLMs-related contributions at CHI [53], contributing a methodological advance alongside theoretical and empirical insights.

Theoretical We show that race–gender identity cues do not operate separately; partial matches reduced perceived fairness compared to both full matches and mismatches. This intersectional fairness paradox challenges SIT predictions and extends CASA by showing how fairness attributions emerge when multiple social cues interact in adaptive AI interviews.

Empirical We provide experimental evidence that fairness and bias perceptions are more sensitive to avatar identity cues than trust. Post-interview trust was high across conditions, while racial mismatches heightened perceived bias and partial matches reduced perceived fairness, highlighting that identity cues in adaptive ECAs can shape user experience in consequential settings such as AI-hiring.

Methodological We developed a scalable platform for studying ECA interactions using real-time generative avatars with multimodal measurement. By combining self-report, sentiment analysis, and webcam-based eye tracking, the platform provides a reusable framework for capturing both explicit perceptions and implicit behaviors in fairness-sensitive applications.

2 Related Work

We review and contrast our work with prior research in three areas that frame our study: (1) SIT in human–AI interaction, (2) social response in ECAs, and (3) fairness in AI hiring.

2.1 Social Identity Theory and AI

Social Identity Theory (SIT), introduced by Tajfel [79], explains how individuals categorize themselves and others into social groups, producing in-group favoritism and out-group bias [32]. Tajfel's minimal group experiments showed that such favoritism can emerge even from arbitrary categories [78]. We argue that these mechanisms are equally relevant to human–AI interaction. When AI systems present social cues such as faces, voices, or names, they can trigger the same categorical thinking observed in human groups. This echoes Zajko's [87] call to examine not only algorithmic bias, but also how AI systems and their interfaces participate in reproducing structural inequalities beyond technical fairness metrics.

Recent research has begun to examine SIT in human–AI interaction. Seaborn [70] provides a theoretical foundation, outlining axioms that show how anthropomorphic cues can trigger categorization processes while emphasizing that such processes remain *dynamic* (shifting during an interaction) and *context-dependent* (activated by situational factors). Sun et al. [76] demonstrate that users place greater trust in AI agents that are both ingroup and humanoid, extending similarity–attraction effects to artificial agents. Edwards et al. [18] extend this line of work by showing that students high in age identification rated an older-sounding AI instructor as more credible, motivating, and socially present, suggesting that role stereotypes (e.g., professors as older) can outweigh simple similarity cues. In interview contexts, identity effects appear even less predictable. Biswas et al. [6] found that participants' own demographics, mediated by Social Presence and Perception (SPP), shaped perceived fairness whereas avatar race and gender showed no main effects. Do et al. [17] reported that mismatched avatars in virtual reality (VR) reduced embodiment and disproportionately harmed minority users. Together, these findings indicate that SIT effects in human–AI interaction vary with **identity salience** and the **stakes of interaction**.

However, prior work has not examined how multiple identity cues intersect in consequential settings. Schlesinger et al. [69] call for intersectional approaches, warning that treating identity categories in isolation overlooks how technologies reproduce overlapping systems of oppression. Our study addresses these gaps by testing *intersectional identity cues* (race × gender) in a simulated AI interview and assessing how they influence perceived fairness.

2.2 Social Response in ECAs

2.2.1 *From Minimal Cues to LLM-Driven ECAs.*

Embodied conversational agents (ECAs) convey verbal and non-verbal cues (e.g., gaze, gesture) that render abstract AI processes into socially recognizable identities [4, 36]. Early CASA research showed that even minimal cues, such as synthetic voice, role, or gendered presentation, could elicit politeness and stereotyping in simple, scripted systems [51]. Building on this, Reeves and Nass [62] argued that people respond to media as if it were real because human cognition cannot reliably distinguish mediated from unmediated social cues. Later studies confirmed this claim: as computer agents became more anthropomorphic, they elicited stronger social judgments, social influence, and trust than abstract forms [23]. However, today's LLM-based ECAs combine human-like cues such as natural dialogue, turn-taking, and photorealistic appearance. This creates a dilemma for HCI: the same cues designed to foster natural interaction may also trigger stereotyping in consequential domains. As ECAs become more realistic and socially responsive, they are shifting from low-stakes applications to high-stakes contexts where decisions can carry real-life consequences for individuals. Yet most prior work has examined ECAs in scripted or low-stakes settings, leaving open the question of whether existing theories still hold when adaptiveness and stakes increase. To address this gap, our study situates ECAs in a simulated job hiring context and combines self-reports with implicit measures to capture perceptions. Table 1 positions our study within this trajectory by comparing prior ECA research on identity cues with our real-time, AI-based interview, multimodal approach.

2.2.2 *Beyond Self-Report to Multimodal Measures.*

Most ECA studies still rely on self-reports, overlooking how subconscious processes shape interaction [2, 6, 70, 76, 77]. Reeves and Nass [62] argue that self-reports are insufficient because human responses to media are often unconscious and automatic, making self-reflections unreliable. Similarly, Grimm [25] shows that self-reports are vulnerable to *social desirability bias*, which can mask implicit prejudice.

Table 1: Trajectory of related work on ECAs and identity cues. Prior studies are mostly low-stakes, scripted, and based on self-reports. Our study introduces a real-time, simulated AI-based job interview and multimodal testbed approach.

Study	Context (Domain stakes)	Interface	Key contribution
Nass et al. [51]	Lab quiz, Q&A tasks (low)	Scripted agent	CASA paradigm: minimal cues (voice, role, gender) elicit social responses
Sun et al. [76]	Trust game, lab & online (low)	Video intro & game interface	Ingroup and humanoid AIs increase trust
Aumüller et al. [2]	Bank service chatbot (low)	Text chat and avatar (ambiguous vs. abstract)	Participants preferred abstract icons; perceiving the avatar as female increased intention to use
Szafarski et al. [77]	Luxury automotive customer interaction (medium)	Video demo (Study 1); live LLM-based ECA (Study 2)	High TAM acceptance, with younger users rated more intuitive; design and user experience focus; no fairness analysis
Biswas et al. [6]	Simulated job interview (high)	Asynchronous video interview (AVI)	No main effect of agent race/gender; participant demographics shaped fairness via SPP; no match-mismatch; no post-outcome analysis
This study	Simulated job interview (high)	Real-time LLM-based ECA	First perceived fairness evaluation of participant–avatar identity alignment with multimodal measures (self-report, sentiment, eye tracking)

Together, these limitations make it particularly difficult to evaluate perceptions of socially sensitive topics.

Recent work has begun to address this gap by using implicit behavioral measures to capture processes that self-reports miss. Eye-tracking research has long shown that gaze patterns reveal cultural and social dynamics. In the other-race effect (ORE), observers allocate attention differently to own- versus other-race faces, which contributes to poorer cross-race recognition [30]. In live interaction, cultural groups also differ in how often they engage in mutual gaze [26]. More recent work connects gaze to attitudes. Steinfeld and Shaked [75] found that Israeli participants who looked longer at a Palestinian speaker during simulated virtual contact reported more positive outgroup perceptions. Extending beyond gaze, Peck et al. [54] showed that head and hand motion in VR revealed racial bias even when measures such as shooting accuracy and response latency did not.

These findings demonstrate that multimodal implicit measures uncover attentional, cultural, and bias-related processes that explicit judgments often suppress or overlook. More broadly, Lai et al. [38] caution that much empirical research on human–AI decision making relies on simplified tasks and ad hoc measures, raising concerns about validity and generalizability. While our study remains a simulation, it moves closer to being ecologically valid by situating ECAs in an online hiring interview and combining validated self-reports with implicit behavioral measures. Building on this, we contribute a scalable multimodal platform for evaluating both explicit and implicit perceptions in fairness-sensitive ECA interactions.

2.3 Fairness in AI Hiring

Before AI was introduced, fairness in hiring was studied extensively in organizational psychology. Organizational justice theory shows that applicants judge fairness by both outcomes and procedures such as job-relatedness, consistency, and respectful treatment [13, 21]. Building on this, Ryan and Ployhart [64] show that applicants' experiences of the hiring process influence their trust, willingness to pursue opportunities, and perceptions of organizational legitimacy.

With the rise of automated screening and AI-based decision systems [42, 74], fairness concerns have become more critical. While such tools promise efficiency and scalability, they also reduce human oversight, introduce opacity, and risk embedding or amplifying inequities. Much of the research on AI hiring has therefore emphasized bias mitigation and transparency through technical approaches [1]. Experimental evidence further shows that awareness of algorithmic gender bias can deter qualified women from applying [34]. Yet this line of research still frames fairness primarily as an algorithmic property rather than as a lived experience. Liao and Varshney [44] argue that fairness is shaped not only by algorithms but also by how users interpret model explanations, while Woodruff et al. [86] emphasize that such perceptions are especially salient for minoritized groups, who evaluate fairness through the lens of their lived experiences. Van Berkel et al. [80] further demonstrate that people actively judge which predictors feel fair or unfair, underscoring that fairness is ultimately constructed through perception rather than technical definitions alone. Our study aligns with this line of work by examining how perceived fairness is shaped in a simulated AI interview, particularly through identity cues expressed by avatars.

Recent work also foregrounds fairness as an experiential judgment, examining it through system qualities and social cues. Hosain et al. [33] show that applicants' perceptions of AI hiring systems are positively associated with procedural justice, mediated by whether the systems feel easy to use, useful, and trustworthy. This highlights that fairness judgments extend beyond technical bias reduction to include applicants' broader impressions of system usability. At the same time, the design of AI interview interfaces matters. Biswas et al. [6] varied avatar race and gender in an AVI-based study and treated agent and participant demographics as independent predictors. They found no main effects of avatar race or gender, but provided an important foundation for our work by showing that participants' own demographics, mediated by SPP, can influence perceived fairness, privacy, and impression management. Our work extends this research by studying perceived fairness in a different setting, where applicants must interpret a uniformly negative decision, and by analyzing relational identity (mis)matches between participants and LLM-based interviewers using multimodal behavioral measures. Beyond the avatar interface, other communicative cues can also shape applicants' experiences. Heo et al. [28] found that gendered chatbot voices affected applicants' language use, confidence, and interviewer ratings. Pyle et al. [58] further showed that candidates experienced emotion AI in video interviews as procedurally and interactionally unjust, raising concerns about its adoption.

Despite these advances, research on fairness in AI hiring remains narrow. Most studies emphasize algorithmic design or rely on self-reports of fairness, offering limited insight into how judgments develop and shift during live interaction. Our study addresses this gap by examining how fairness perceptions unfold in real-time exchanges with photorealistic ECAs, and by analyzing how unfavorable outcomes are attributed to the system. By combining explicit justice ratings with implicit measures of attention and sentiment, we advance a multimodal approach for examining fairness in adaptive ECA hiring contexts.

3 Method

In this section, we describe the experimental design, demographics of our participants, experimental platform, procedure, measures, and analysis. The Institutional Review Board (IRB) of the Technical University of Munich approved our study.

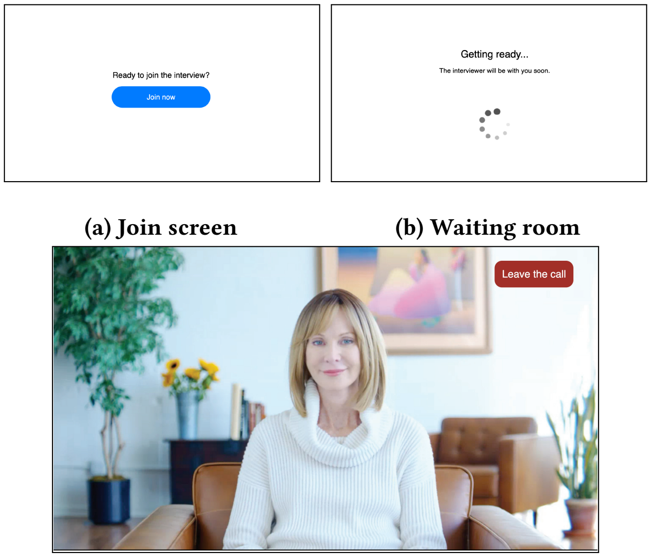
3.1 Experimental Design

We designed a 2 × 2 between-subjects experiment to investigate how identity matching between participants and an AI interviewer in terms of race and sex influences user perceptions of a simulated job interview. To this end, we randomly assigned our participants to one of four experimental conditions:

- **Avatar race:** Match vs. Mismatch (relative to the participant's self-identified ethnicity)
- **Avatar sex:** Match vs. Mismatch (relative to the participant's self-identified gender)

Each participant completed the study in a single session, which involved answering four voice-based interview questions that the AI interviewer delivered. To make the experience feel realistic, we utilized a user interface that resembles a typical online interview platform. Participants joined the session through a video-call-like interface, with buttons and layout elements that mirrored the appearance of online meeting tools, as shown in Figure 2. In contrast to prior AVI work [6], our AI interviewer responded in real time using a streaming LLM-based ECA, enabling turn-taking, clarifications, and conversational adaptivity.

3.1.1 Interview Task Design. We designed our interview and feedback procedure based on previous research in human resources and organizational psychology [11, 40, 59, 88]. In this study, we informed our participants that they would be interviewed for a "customer support" role. We selected this role for its relatability and low task complexity, especially in a remote, online setting with a diverse participant pool. Furthermore, customer support positions emphasize interpersonal communication, empathy, and problem-solving competencies, which are linked to job readiness [45].


Ready to join the interview?
Join now
Getting ready...
The interviewer will be with you soon.
Leave the call

(a) Join screen **(b) Waiting room**

(c) In-call view

Figure 2: Participants accessed the study through a video-call-like interface with a join screen, waiting room, and in-call view (with a "Leave the call" button) designed to mirror familiar online meeting tools.

To assess these competencies, we followed a Competency-Based Interview (CBI) format, a structured technique that focuses on assessing specific competencies relevant to the targeted job, rather than relying on past experiences as in Behavioral Event Interviews (BEI). CBI formats are considered fairer and, in some cases, more predictive of job performance [59, 88]. The interview questions were designed in a way that progressed from general to role-specific, including: *(1) Could you tell me a bit about yourself?, (2) Can you explain a situation where you helped another person and solved a problem for them?, (3) How do you respond to critical feedback?, and (4) How would you handle a frustrated customer?*

All participants, regardless of condition, received a simulated standardized rejection. This approach aligns with Skitka et al. [73], who distinguish outcome favorability from outcome fairness, and is further supported by Gilliland's fairness model [21], which emphasizes that applicant reactions are shaped not only by outcomes but also by procedural qualities (i.e., consistency, bias suppression, and respectful treatment). These factors are especially influential when outcomes are unfavorable, helping us assess whether identity matching moderates negative perceptions, which were not examined in prior AVI studies [6]. The rejection message followed best practices outlined by Cortini et al. [14], using polite, informal language that acknowledged participants' effort without over-justifying the decision to deliver the message. The rejection message used for this study is provided in Appendix A.

3.2 Participants

We conducted the study in July 2025 on the online research platform Prolific[1], recruiting 228 participants from the United Kingdom, Germany, and the United States. Eligibility criteria included being at least 18 years old, fluent in English, and having access to a functional webcam, microphone, stable internet connection, and a quiet environment. For our 2 × 2 factorial design, we restricted inclusion to participants who, according to Prolific's prescreen settings[2], identified their sex as female or male and their ethnicity (simplified) as white or black. In our own survey, participants reported self-identified gender and ethnicity again using expanded categories (e.g., non-binary, mixed ethnicity). For analysis, we included only those who self-identified as women or men and as either white (European descent) or black (African descent); these self-reports were then used to assign match/mismatch conditions relative to avatars' race and sex.

We focused on this majority–minority contrast given evidence that hiring discrimination is often triggered by visible identity cues such as skin color, which signal perceived cultural distance [89], and that black applicants receive significantly fewer callbacks than equally qualified white applicants [5].

We randomly assigned participants to one of four conditions, crossing avatar race (white vs. black) and sex (female vs. male). We used an adaptive stratified allocation method (minimization) to balance participant numbers across conditions, with randomization applied to break ties when multiple cells were equally under-represented. Gender was recorded as women ($n = 110$), men ($n = 108$), and non-binary ($n = 2$). Although our design required binary assignment to avatar conditions, we report the presence of two non-binary participants for completeness; they, along with others whose self-reported identity did not fit the predefined categories (e.g., mixed ethnicity), were excluded from analysis.

We excluded thirteen participants (5.7%), either due to technical issues (e.g., avatar not loading, eye-tracking errors, incomplete sessions) or because their reported identity fell outside our 2 × 2 design categories. The final analysis sample comprised 215 participants. Table 2 summarizes participant demographics by experimental condition. Each session lasted approximately 20 minutes, and participants were compensated £4.27 (equivalent to £12.81/hour) in line with Prolific's compensation policy[3]. Compensation was not based on interview performance or the hiring decision, and we did not offer any bonuses. All participants provided informed consent and were reminded of their right to withdraw at any time.

3.3 Experimental Platform

We developed a web-based platform that managed the survey flow, the AI interviewer, and webcam-based eye tracking. The overall system architecture is shown in Figure 3. The platform was hosted on a remote server and accessed via standard web browsers (except Mozilla Firefox, due to API limitations). Participants joined remotely using their own laptops or desktops with a stable internet connection, webcam, and microphone.

3.3.1 System Architecture. We implemented the backend in Flask version 3.1.1[4] and the frontend in vanilla JavaScript with Vite, controlling the experimental workflow. The platform delivered the informed consent form digitally, enabled simultaneous capture of questionnaire responses and interview transcripts via speech-to-text (STT), and integrated RealEye for eye-tracking data collection.

3.3.2 AI Interviewer and Conversational AI. We built the AI interviewer with HeyGen's Streaming Avatar SDK[5], which renders photorealistic avatars in real time with synchronized lip movements and facial expressions. We selected four professionally dressed avatars from HeyGen's library, systematically varying in phenotypic features associated with race (skin tone, facial structure) and sex (jawline, hairstyle) while maintaining consistent professional attire and setting. These visual traits were not intended to represent any individual's subjective identity but to evoke perceived social categories commonly used in social cognition research. The four avatars used in the experiment are shown in Figure 1, and all shared the same neutral background and the same voice settings (American English, neutral pitch and pace, calm tone).

We generated avatar responses through HeyGen's integration with OpenAI's GPT-4o-mini model[6], which supports low-latency and real-time voice conversation. In our experiment, the avatar could greet participants by name, acknowledge information they had provided, ask for clarification when an answer was unclear, and pose brief follow-up questions when responses were very short. This real-time setup differs from prior AVI systems [6], where the virtual interviewer presented pre-recorded video prompts rather than generating adaptive follow-ups in real time. These adaptive behaviors can make the interaction feel more natural and conversational, thereby improving the ecological validity of the simulated AI-based job interview.

We controlled the interviewer's behavior through system prompt (referred to by HeyGen as *Knowledge Base*), which specified the professional persona, interview questions, and interaction boundaries (e.g., tone, formality, response length). Sessions were capped at five minutes, with most interviews lasting approximately three minutes. The full prompt is available in Appendix B.

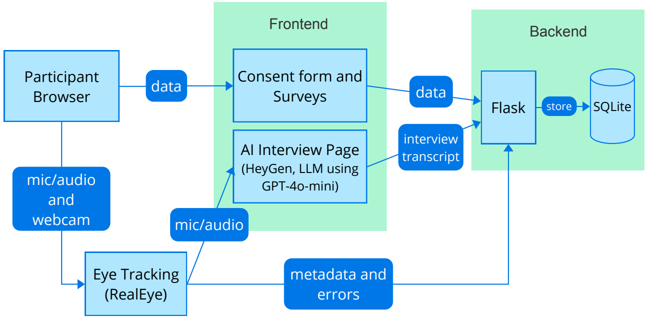


Figure 3: High-level system architecture of the experimental platform.

[1] https://www.prolific.com/, last accessed 22 January 2026

[2] https://researcher-help.prolific.com/en/article/412c0a, last accessed 22 January 2026

[3] https://researcher-help.prolific.com/en/articles/445230-prolific-s-payment-principles, last accessed 22 January 2026

[4] https://flask.palletsprojects.com/en/stable/, last accessed 22 January 2026

[5] https://docs.heygen.com/docs/streaming-api, last accessed 22 January 2026

[6] https://platform.openai.com/docs/guides/realtime, last accessed 22 January 2026

Table 2: Participant demographics by experimental condition. Values are mean ± SD unless otherwise noted.

Variable	Experimental Condition			
	No match (n=53)	Gender match (n=57)	Racial match (n=54)	Full match (n=51)
Demographics				
Gender, n (Men/Women)[a]	27/26	27/30	28/26	26/25
Age (years)	38.5 ± 12.7	39.3 ± 11.0	39.3 ± 9.5	42.5 ± 12.6
Ethnicity, n (White/Black)[b]	27/26	29/28	29/25	28/23
Vision status, n (Normal/Glasses/Contacts)[c]	37/13/3	36/20/1	38/12/4	33/16/2
Interview Experience				
Employment, % employed[d]	75	79	81	78
Oral interview experience, % moderate+[d]	68	81	71	76
Individual Differences				
AI usage (hours/week)	6.0 ± 12.9	7.9 ± 12.2	4.6 ± 7.2	4.5 ± 4.8
Speaking nervousness (1–10 scale)	4.2 ± 2.3	4.0 ± 2.1	4.3 ± 2.4	4.1 ± 2.2
Fairness beliefs and attitudes (1–5 scale)	3.22 ± 0.35	3.29 ± 0.35	3.29 ± 0.34	3.14 ± 0.34

[a] Two participants identified as non-binary overall.

[b] Ethnicity categories reflect survey wording: White (for example European descent), Black or African descent.

[c] Vision status = self-reported normal vision, wearing glasses, or wearing contact lenses.

[d] Employment = full/part-time; Oral interview experience = moderate or higher.

3.3.3 Session Management and Data Capture. WebSocket events from HeyGen's Streaming SDK managed real-time interaction. Session completion was detected by a scripted closing phrase: *"Thank you for participating... Please click the 'Leave the call' button to move on."* When detected, the platform automatically displayed the "Leave the call" button. To ensure continuity when the phrase was missed (e.g., due to automatic speech recognition (ASR) errors), the system displayed a fallback "Continue" button after four minutes.

To capture conversational transcripts, the system streamed and transcribed avatar and user speech in real time, appending each turn to a local buffer. The backend received buffered data once participants clicked the "Leave the call" or fallback button. During pilot testing, we observed end-to-end latency of 2–7 seconds between user speech and avatar responses, sufficient to maintain conversational flow, though occasionally producing short pauses.

3.3.4 Eye-Tracking Integration. We collected gaze data using RealEye version 18.49.0[7], a browser-based eye-tracking platform. Sampling frequency (10–60 Hz) depended on the participant's webcam. RealEye estimates gaze using a machine-learning model refined by calibration. According to the company's validation, full-screen accuracy is approximately 100–125 pixels (≈1.5–2° visual angle), with the highest accuracy in the central region. We integrated RealEye's embedded SDK[8] to define areas of interest (AOIs) for the avatar's face and body as shown in Figure 4. RealEye aggregated fixations within AOIs, monitored head pose, distance, and illumination, and provided a virtual chinrest. Before each session, we reminded participants to maintain a stable posture and lighting. Because our experiment was fully remote and crowdsourced, we used webcam-based eye tracking as a complementary process measure to avoid treating the human–AI interaction as a black box and to verify that participants noticed avatar identity cues, particularly after the scripted rejection. This provides an additional, objective lens for interpreting differences in perceived fairness. To our knowledge, combining crowdsourced webcam eye tracking with an AI-based job interview task has not yet been explored in ECA research.

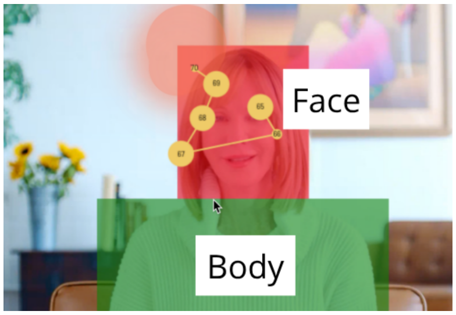


Figure 4: Definition of AOI (Area of Interest). Face and body regions were coded for gaze analysis.

3.4 Procedure

The experimental session lasted approximately 20 minutes and followed the protocol illustrated in Figure 5.

Initial Setup. Upon accessing the study link via Prolific, participants first reviewed and digitally signed the informed consent form. They then verified their technical setup (webcam, microphone, stable internet connection) and completed a presurvey collecting demographic information and baseline fairness beliefs and attitudes.

[7] https://www.realeye.io/, last accessed 22 January 2026

[8] https://app.realeye.io/docs/embedded-website-sdk, last accessed 22 January 2026

Task Instructions. Next, we provided an overview explaining that participants were taking part in an online job interview with an AI interviewer. We told them they would answer several short questions verbally and that, after the interview, the AI would inform them of the hiring decision. We then presented the following position description:

> *Imagine you are applying for a Customer Support position at a tech company. Your job would involve helping customers, answering questions, and handling complaints.*

On the same page, we informed participants that only interview transcripts would be recorded and that no audio or video would be stored. We then provided step-by-step instructions for webcam-based eye-tracking calibration.

Eye-Tracking Calibration. Before the interview, participants completed RealEye's two-step procedure comprising a 39-point calibration and a three-point validation. Participants who did not pass after two attempts were redirected back to Prolific to avoid polluting the study data set with low-quality data. They were nevertheless compensated for the time spent on the study.

Interview Task. We randomly assigned participants to one of four avatar conditions based on their demographics. The AI interviewer appeared on screen, greeted each participant by name, and conducted the structured interview with four competency-based questions defined in Section 3.1.1. The interview lasted on average three minutes. When participants completed the interview, they clicked the "Leave the call" button to proceed.

Post-Interview Assessment. After the interview, participants completed questionnaires measuring trust in the AI interviewer, acceptance of AI interviews, and any technical issues experienced.

Hiring Decision and Post-Outcome Assessment. We delivered the hiring decision through a simulated evaluation process as shown in Appendix Figure 8. After participants clicked "Check hiring decision," they watched a pre-recorded video in which the same avatar they had interviewed with delivered the standardized rejection message. Participants then completed measures of perceived procedural and distributive justice, perceived bias, and two manipulation checks that verified recognition of the avatar's sex and race, along with their emotional reactions to the rejection. Participants also had the opportunity to provide optional feedback.

Debriefing. At the end of the session, our system debriefed participants. We informed them that the AI hiring decision had been pre-set (i.e., not based on their responses) and that the study examined how avatar race and sex shape perceptions of fairness and bias following negative feedback. Participants were reminded that webcam-based eye tracking was used to study visual attention, that only text transcripts were stored, and that they could contact the research team to request data exclusion. We received no requests for data removal.

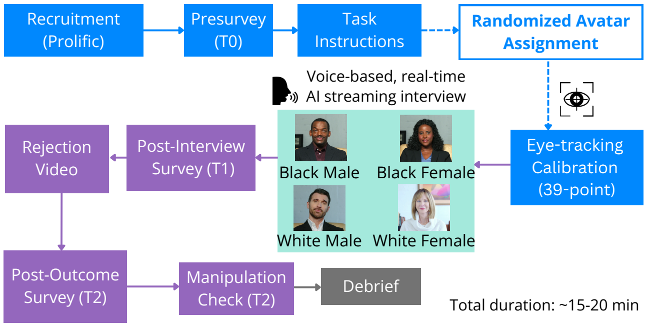


Figure 5: Experimental procedure. After presurvey measures (T0), participants received task instructions, were randomly assigned to an avatar, and completed eye-tracking calibration. They then engaged in a real-time AI-based interview, completed post-interview surveys (T1), received rejection feedback via video, and filled out post-outcome measures (T2) before debriefing.

3.5 Measures

The study included self-report and implicit measures. This section details all measures, with full item wordings provided in Appendix Tables 4–8. We collected self-report measures at three timepoints: T0 (presurvey, before the interview), T1 (post-interview), and T2 (post-outcome, after participants received the rejection). We recorded implicit measures during the interview. We recoded reverse-worded items before averaging and assessed internal consistency with Cronbach's α.

3.5.1 Self-report measures. All self-report items were rated on a 1–5 Likert scale (1=*strongly disagree*, 5=*strongly agree*), except for the Trust in AI scale, which used a 1–7 semantic-differential format.

Presurvey (T0). Before the interview, participants completed a presurvey capturing demographics, education and employment background, prior experience with oral assessments, public speaking anxiety, AI interaction habits, and baseline attitudes and beliefs about fairness and evaluation. Specifically, we measured public speaking anxiety with items from Dechant et al. [16] and fairness-related beliefs and attitudes with a scale from Rezai [63].

Trust in AI (T1). We measured trust using the semantic-differential Trust in AI scale by Shang et al. [71], which comprises two subscales: *cognitive* trust (18 items) and *affective* trust (9 items). We averaged items within each subscale. Internal consistency was excellent (cognitive: Cronbach's α = .97; affective: α = .94).

AI Acceptance (T1). We measured acceptance with three evaluative items about future use and acceptability of AI interviews (e.g., willingness to interview again, acceptability of AI in hiring, comfort with AI assessment).

Perceived Fairness (T2). We measured perceived fairness with two subscales adapted from Colquitt's organizational justice measure [13]: *procedural justice* (7 items) and *distributive justice* (4 items). We computed mean scores, and reliability was good to excellent (procedural: α = .88; distributive: α = .90).

Perceived Bias (T2). We measured perceived bias with four event-specific items assessing whether participants believed the AI interviewer's treatment and outcome were influenced by their own identity (ethnicity, gender). The items drew on the attributional framing of the Perceived Ethnic Discrimination Questionnaire—Community Version (PEDQ-CV; "Because of my ethnicity…") [8], adapted to a single-interview context that included gender. We reverse-coded items as needed and averaged them so that higher scores indicated more perceived bias. Internal consistency was acceptable for a short scale (Cronbach's $\alpha = .72$).

Manipulation Checks (T2). After receiving the hiring decision (rejection), participants reported their emotional reaction (single 5-point item: "How did you feel after hearing the hiring decision?") to contextualize later fairness and bias judgments. They then identified the avatar's gender and race (multiple choice) and could provide open-ended feedback about surprising, unusual, or unfair aspects of the interaction.

3.5.2 *Implicit measures.*

Eye-tracking: Coefficient K. We measured the Coefficient K from eye-tracking data during the interview session, focusing on fixations within the avatar's face AOI. In our setting, Coefficient K indicates whether participants looked at the avatar in a more *focal* or more *ambient* way by comparing each fixation's duration with the amplitude of its subsequent saccade; larger values indicate more *focal* viewing of the face, smaller values more *ambient* scanning. In this study, we interpret K as an indicator of visual engagement with the avatar's face. We report K only at the aggregate level as a complement to our self-report measures and use eye tracking solely as a process measure to validate that participants noticed the identity cue of the avatar during the interview and that the interaction unfolded as intended. It was not used to evaluate participants or infer their performance or competence. We adopted the definition introduced by Krejtz et al. [37], which was also implemented in the RealEye platform [41]:

$$K = \frac{1}{n}\sum_{i=1}^{n}\left(\frac{d_i - \mu_d}{\sigma_d} - \frac{a_{i+1} - \mu_a}{\sigma_a}\right).$$

Here, d_i denotes the duration of the ith fixation and a_{i+1} the amplitude of the subsequent saccade. μ_d and μ_a are the means of fixation durations and saccade amplitudes, and σ_d and σ_a their respective standard deviations, each computed over all n fixations in a participant's scanpath. Thus both measures are transformed into standardized z-scores. Positive K values indicate that relatively long fixations were followed by short saccades, reflecting focal processing, whereas negative values indicate that relatively short fixations were followed by long saccades, reflecting ambient processing. Values close to zero suggest a balance between the two modes.

RealEye computes K automatically from webcam-based gaze events, standardizing fixation durations and subsequent saccade amplitudes within each session. The platform also dynamically scales AOIs to participants' screen resolution to ensure they remain above the nominal accuracy threshold, and we retained only participants with a minimum sampling rate of 20 Hz [61]. Our analyses, therefore, rely on RealEye's default implementation and quality controls rather than custom preprocessing.

Transcript sentiment. We applied sentiment analysis with spaCy-TextBlob [68] to participants' transcribed interview responses. Transcripts were cleaned by removing punctuation and filler tokens and segmented into *user* and *avatar* turns; only user responses were analyzed. For each participant, we averaged (i) *polarity*, $[-1, 1]$ (negative to positive), and (ii) *subjectivity*, $[0, 1]$ (objective to subjective).

3.6 Analysis

For each dependent variable, we applied factorial ANOVAs at $\alpha = .05$. When assumptions of normality or homogeneity of variance were violated, we used the aligned rank transform (ART) [84], a non-parametric extension of ANOVA, and we conducted post hoc tests using ART-C contrasts [19]. Effect sizes are reported as partial eta squared (η_p^2). All analyses were conducted in R (version 2025.05.1+513), and in Python (version 3.9.6).

3.7 Positionality Statement

We come from computer science and HCI backgrounds with a predominantly quantitative, experimental orientation, which foregrounds controlled measures, variance reduction, and causal identification over lived experience and interpretive depth. This positionality shapes our methodological choices and the kinds of knowledge our study can generate. Our design also uses a deliberately simplified set of avatar phenotypic traits (black/white; male/female), and we analyze only participants who self-identify within these categories. We are aware that this does not reflect the complexity, fluidity, or intersectionality of real-world identities and is not intended to essentialize demographic groups. Rather, it reflects a methodological decision in quantitative experimental research to reduce conceptual complexity in order to create interpretable conditions within a controlled design.

4 Results

We present the results in response to our three RQs within the scope of our experimental design.

4.1 Perception in Gen-AI ECAs

For **RQ1** (i.e., "Do participants perceive meaningful differences between avatar phenotypic traits?"), we analyzed participants' responses to the four different avatar conditions (avatar race: black or white; avatar sex: male or female). For this RQ, we did not consider the participant–avatar matching manipulation, but ensured equal distribution of participant ethnicity and gender between the avatar conditions. Table 3 summarizes the avatar conditions relevant to RQ1. As a manipulation check, we operationalize **perceptual association** as the percentage of participants whose perceptions matched the avatar's intended presentation (male/female; Black-/White); responses of non-binary, unsure, or other were coded as not aligned. Perceptual association was high across conditions. Alignment between intended avatar presentation and reported gender averaged 94.9% (SD = 2.1%), and alignment between intended presentation and perceived ethnic/racial background averaged 95.3%

(SD = 1.7%). These results confirm that the phenotypic manipulations effectively conveyed the intended social categories.

Due to violations of normality and homogeneity of variance assumptions, we conducted an ART ANOVA. Results revealed a significant main effect of avatar race on **cognitive trust**, $F(1, 211) = 5.48$, $p = .020$, $\eta_p^2 = .025$, with participants rating black avatars ($M = 5.64$, $SD = 1.04$) as higher cognitive trust than white avatars ($M = 5.28$, $SD = 1.19$). No main effect of avatar sex was found, $F(1, 211) = 0.05, p = .821$, and the interaction between race and sex was not significant, $F(1, 211) = 1.56$, $p = .213$. For **affective trust**, we found no significant main effects of race ($F(1, 211) = 2.55, p = .112$) or sex ($F(1, 211) = 0.36, p = .547$). However, the interaction between race and sex approached significance ($F(1, 211) = 3.85, p = .051, \eta_p^2 = .019$), suggesting that affective trust ratings for black male avatars ($M = 5.71, SD = 1.01$) tended to be higher than for white male avatars ($M = 5.22, SD = 1.04, p = .071$).

Regarding **AI interview acceptance**, two measures showed significant differences. For **future interview intention**, an ART ANOVA revealed a main effect of avatar's race, $F(1, 211) = 5.64$, $p = .019, \eta_p^2 = .026$. Participants interacting with black avatars ($M = 4.07$, $SD = 1.13$) reported greater willingness to be interviewed by AI again compared to those interacting with white avatars ($M = 3.70$, $SD = 1.25$). No main effect of avatar sex was found, $F(1, 211) = 0.25$, $p = .618$, and the interaction was not significant, $F(1, 211) = 1.06$, $p = .304$.

For **comfort with AI assessment**, the ART ANOVA indicated a main effect of avatars' race, $F(1, 211) = 4.28$, $p = .040$, $\eta_p^2 = .020$. Black avatars ($M = 3.83$, $SD = 1.37$) were associated with higher comfort ratings than white avatars ($M = 3.37$, $SD = 1.47$). No significant main effect of avatar sex emerged, $F(1, 211) = 0.76$, $p = .384$, and the interaction was not significant, $F(1, 211) = 1.73$, $p = .190$.

Finally, we evaluated participants' emotional reactions to the standardized rejection using the post-result emotion item; across conditions, these reactions were predominantly negative or neutral, with only a small minority of participants reporting positive feelings about the hiring decision, as shown in Table 3.

4.2 Trust, Perceived Fairness and Bias

In **RQ2** (i.e., "Does racial or gender matching affect trust, perceived fairness and bias?"), we analyzed the matching conditions that affected participants' trust, perceived fairness, and bias, presented in Figure 6. Avatar conditions were assigned according to our matching manipulation (no match, gender match, racial match, or full match), which assigned the avatar condition to participants' self-reports, as described in Table 2. Across all trust measures (trust total, cognitive trust, and affective trust), ART ANOVAs revealed no significant main effects of racial match (all $p \geq .648$) or gender match (all $p \geq .693$), and no significant interactions (all $p \geq .509$), as shown in Figure 6a. Mean ratings were consistently high across conditions, with trust total ranging from $M = 5.44$ to $M = 5.53$ ($SD = 0.97$ to 1.14), cognitive trust from $M = 5.43$ to $M = 5.49$ ($SD = 0.93$ to 1.23), and affective trust from $M = 5.44$ to $M = 5.58$ ($SD = 1.03$ to 1.15).

For **perceived ethnic bias** in Figure 6b, higher scores indicate stronger agreement that the participant's ethnicity influenced how the AI interviewer treated them. The ART ANOVA revealed a significant main effect of racial match, $F(1, 212) = 4.98$, $p = .027$, $\eta_p^2 = .023$, with racially mismatched avatars ($M = 2.19$, $SD = 1.23$) receiving higher bias ratings than matched avatars ($M = 1.82$, $SD = 1.12$). No significant main effect of gender match was found, $F(1, 212) = 0.54$, $p = .463$, and the interaction was not significant, $F(1, 212) = 0.13$, $p = .717$, as well.

For **distributive justice** as shown in Figure 6c, higher scores indicate greater perceived fairness in outcome distribution. ART ANOVA results showed a significant interaction effect between racial match and gender match, $F(1, 212) = 4.75$, $p = .030$, $\eta_p^2 = .022$. Neither the main effect of racial match, $F(1, 212) < 0.01$, $p = .997$, nor the main effect of gender match, $F(1, 212) = 0.26$, $p = .611$, was significant. Mean scores were: no match ($M = 2.90$, $SD = 1.00$), gender match only ($M = 2.64$, $SD = 1.00$), racial match only ($M = 2.62$, $SD = 0.94$), and full match ($M = 2.92$, $SD = 0.98$).

4.3 Implicit Behavioral Measures

To further examine cognitive and perceptual processes in human–AI interaction, we tested **RQ3**: "Does racial or gender matching affect implicit behavioral measures (sentiment and eye tracking)?"

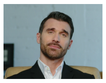
For **sentiment polarity**, presented in Figure 7a; $n = 203$ due to missing transcript data from technical failures or permission settings. Sentiment polarity scores ranged from −1 to +1, with higher values indicating more positive sentiment. We conducted a two-way ANOVA as normality and homogeneity assumptions were met. Results revealed a significant interaction between racial match and gender match, $F(1, 199) = 4.36$, $p = .038$, $\eta_p^2 = .022$. Neither the main effect of racial match, $F(1, 199) = 2.09$, $p = .150$, nor gender match, $F(1, 199) = 0.37$, $p = .541$, was significant. Mean scores were: no match ($M = 0.17$, $SD = 0.13$), gender match only ($M = 0.20$, $SD = 0.12$), racial match only ($M = 0.23$, $SD = 0.12$), and full match ($M = 0.18$, $SD = 0.13$).

For **sentiment subjectivity**, shown in Figure 7b; $n = 203$, scores ranged from 0 to 1, with higher values indicating more subjective responses. As the normality assumption was violated, we applied an ART ANOVA. Results showed no significant main effects of racial match, $F(1, 199) = 2.84, p = .094$, or gender match, $F(1, 199) = 1.38$, $p = .242$, and no significant interaction, $F(1, 199) = 0.17, p = .682$. Mean scores were: no match ($M = 0.46$, $SD = 0.12$), gender match only ($M = 0.48, SD = 0.08$), racial match only ($M = 0.50, SD = 0.11$), and full match ($M = 0.50$, $SD = 0.12$).

We used eye-tracking data from 152 participants who passed calibration and achieved a tracking quality of at least 20 Hz [61]. At this minimum sampling rate, approximately 20 gaze samples are collected per second, which is sufficient for computing fixations.

Although webcam limitations reduced the usable sample, the gaze data provided insight into how applicants visually engaged with the AI interviewer. For face AOIs (normalized K-coefficient), the normality assumption was violated, so we applied an ART ANOVA. As shown in Figure 7c, results revealed a significant main effect of racial match, $F(1, 148) = 4.87$, $p = .029$, $\eta_p^2 = .032$, with racially mismatched avatars ($M = 0.43$, $SD = 0.29$) showing higher focal attention to face than matched avatars ($M = 0.35$, $SD = 0.31$). No significant main effect of gender match, $F(1, 148) = 0.80$, $p = .372$, and no interaction, $F(1, 148) = 0.19$, $p = .664$, were found.

Table 3: Participant responses across avatar conditions, including perceptual association, trust, and acceptance measures.

Measure	Avatar Condition			
	Black Male	Black Female	White Male	White Female
				
	(n=49)	(n=55)	(n=56)	(n=55)
Perceptual association (%)				
Gender perceived as intended	95.9	96.4	94.6	92.7
Ethnic/racial background perceived as intended	95.9	96.4	92.9	96.4
Trust and Emotion				
Cognitive Trust[a,†]	5.79±0.83	5.51±1.18	5.18±1.23	5.38±1.15
Affective Trust[a]	5.71±1.01	5.46±1.21	5.22±1.04	5.57±1.05
Negative emotion after hiring decision[b] (%)	49	45	52	45
AI Interview Acceptance[c]				
Future interview intention[‡] (% would interview again)	4.16±0.99 (83.7)	3.98±1.24 (76.4)	3.59±1.32 (60.7)	3.82±1.17 (70.9)
Acceptability for hiring (% find acceptable)	3.82±1.22 (71.4)	3.62±1.37 (61.8)	3.43±1.39 (53.6)	3.44±1.30 (60.0)
Comfort with AI assessment[‡] (% comfortable)	4.08±1.30 (77.6)	3.60±1.40 (65.5)	3.32±1.47 (51.8)	3.42±1.49 (52.7)

Perceptual association (%) indicates agreement between intended avatar attributes and participant perceptions.

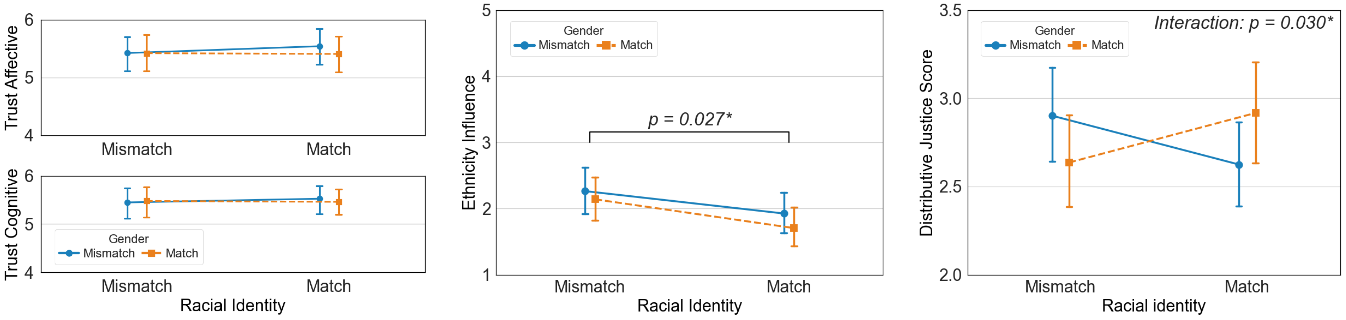
[a] Trust rated on a 7-point scale (1 = unreliable, 7 = reliable).

[b] Proportion of participants selecting "very negative" or "slightly negative" on a 5-point scale (1 = very negative, 5 = very positive).

[c] Acceptance items rated on a 5-point scale (1 = strongly disagree, 5 = strongly agree).

[†] Significant main effect of avatar race, $p = .020$ (ART ANOVA with ART-C contrast tests).

[‡] Significant main effect of avatar race: *Future interview intention* ($p = .019$), *Comfort with AI assessment* ($p = .040$).



(a) Perceived trust scores

(b) Perceived ethnic bias scores

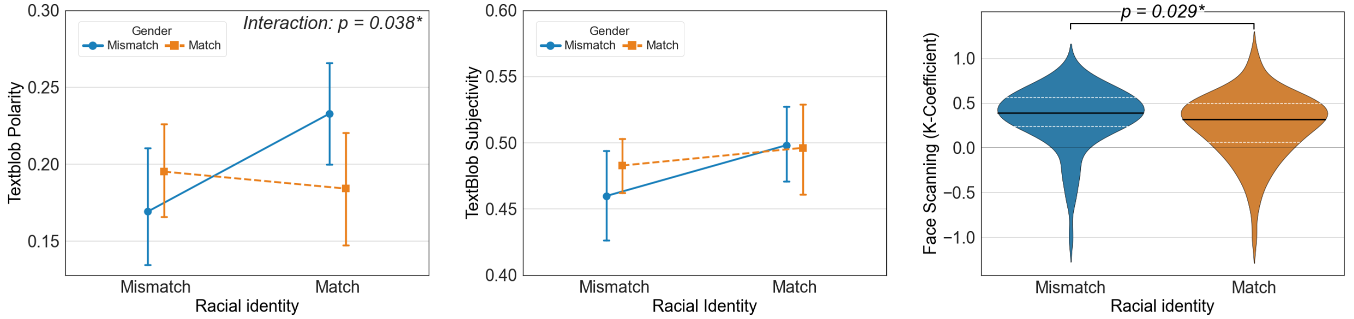
(c) Distributive justice scores

Figure 6: (a) Average perceived affective and cognitive trust with no significance between any matching condition, rated on a 7-point scale (1 = untrustworthy, 7 = trustworthy). (b) Perceived ethnic bias rated on a 5-point scale (1 = strongly disagree, 5 = strongly agree). (c) Distributive justice rated on a 5-point scale; higher scores indicate greater perceived fairness in outcome distribution. Significance levels are indicated by *, corresponding to $p < .05$. Error bars represent 95% confidence intervals.

5 Discussion

5.1 Summary of Findings

In this study, we examined how avatar identity cues shape experiences in a real-time, simulated AI-based job interview. Three patterns emerged. First, participants reliably identified avatars' identity categories as intended (≈95% alignment between intended presentation and reported gender and ethnic/racial background), and prior to the hiring decision, black avatars received slightly higher ratings on cognitive trust ($\Delta = +0.36$ on a 7-point scale), acceptance ($\Delta = +0.37$ on a 5-point scale), and comfort ($\Delta = +0.46$ on a 5-point



(a) Sentiment - polarity (-1 to +1)

(b) Sentiment - subjectivity (0 to 1)

(c) Focal attention on face AOIs (Coefficient K; normalized)

Figure 7: Implicit behavioral measures: (a) Sentiment polarity scores (range: -1 to +1; higher values indicate more positive sentiment, lower values indicate more negative sentiment), (b) Sentiment subjectivity scores (range: 0 to 1; higher values indicate greater subjectivity, lower values indicate greater objectivity), and (c) Focal attention on face AOIs (Coefficient K; normalized). Significance levels are indicated by *, corresponding to $p < .05$. Error bars represent 95% confidence intervals. Violin plots depict score distributions with median and quartiles.

scale) than white avatars (RQ1). Second, identity matching between avatars and participants did not affect trust (which remained consistently high). However, **mismatch** increased perceived ethnic bias ($\Delta = +0.37$ on a 5-point scale), and **partial matches** (race-only or gender-only) reduced distributive justice relative to both-match or neither-match conditions (≈0.28–0.30 points on a 5-point scale; RQ2). Third, implicit measures were mixed: sentiment polarity differences were small, whereas gaze showed higher face-focal attention under racial mismatch (≈+23% relative to matched; RQ3), suggesting increased attention to mismatching avatars. Together, these results suggest that identity cues in ECAs primarily influence **fairness attributions and bias perceptions**, rather than trust, particularly in simulated job hiring settings where outcomes are unfavorable.

5.1.1 *Perception and Acceptance of LLM-Driven ECAs.*

As shown in Table 3, the majority of participants perceived the avatars' identity categories as intended, and black avatars were rated somewhat more positively on cognitive trust, future interview intention, and comfort with AI-based assessment than white avatars. While these differences are statistically significant, the study was not designed to identify the underlying mechanisms that drive participants' perceptions of the avatars. Rather, the results can establish that the avatars' phenotypic traits can meaningfully shape how they are initially perceived. This provides an important foundation for our subsequent analyses. If basic perceptions vary across avatar conditions, it is reasonable to investigate how these relational identity conditions influence further evaluations, such as perceived fairness following the negative outcome.

Results regarding **RQ1** also suggest that adaptive, LLM-based ECAs are evaluated differently depending on the avatar's race and sex, consistent with participants responding to them as social actors. This impression was reflected in open-ended feedback, such as *"It felt very natural"* and *"I was very surprised how realistic the questions came off and how well the AI replied."* These perceptions underscore that small differences in initial trust should not be mistaken for representational fairness; rather, they show that participants generally accepted the LLM-based interviewer before receiving the negative outcome, which provides the starting point for the perceived fairness and perceived bias attributions we examine later in the paper.

5.1.2 *Identity Matching, Perceived Fairness and Bias in AI Interviews.*

For **RQ2**, we examined the effects of identity matching. SIT research suggests that in-group members are generally judged as more trustworthy and cooperative than out-group members. In our case, however, post-interview trust was high across all conditions, with no advantage for identity-matched avatars. One explanation is that the photorealistic ECAs provided strong credibility cues, because professional design features such as controlled expressions, formal style, and affiliative cues can elevate perceived trustworthiness beyond demographic appearance [47]. Combined with the adaptive and naturalistic behavior of the AI interviewer, these cues might have masked potential effects of identity matching on trust, consistent with recent CASA research showing that emergent, socially responsive technologies continue to elicit strong social responses even when group cues are varied [29], and with further recent research also showing that carefully balanced realism can enhance avatar credibility and avoid uncanny effects [3].

After the rejection, however, *racial mismatch* resulted in increased perceived ethnic bias, and distributive justice showed a significant interaction: *partial matches*, where only racial or only gender aligned, received lower perceived fairness ratings than *both-match* or *neither-match*. Kang and Bodenhausen [36] show that such mixed cues invite greater scrutiny, which helps explain why partial matches were associated with lower perceived fairness in our study.

Furthermore, participants maintained high baseline trust in partial matches but judged them more harshly once they received an unfavorable outcome, when they seemed most sensitive to perceived fairness. In contrast to earlier work showing that people applied social norms such as politeness to minimally cued computers [51] or judged coercive computers as less unjust than humans [72], our results show that when AI avatars present with photorealistic, socially recognizable identities, participants respond less forgivingly and scrutinize them through race and sex cues. These findings extend prior AVI-based results [6] by showing that perceived fairness depends on applicant–avatar identity cue alignment in our interview context.

A brief qualitative examination of the open-ended responses indicates that the identity effects we observe may be tied to justice-related attributions participants make after receiving the rejection. Participants' comments varied: some felt they had performed well and therefore experienced the rejection as unfair, while others viewed the decision as reasonable. Illustrative examples include: *"The interviewer told me my answers were good … I was surprised to hear that I wasn't selected."* and *"I felt like my responses were strong, but the rejection felt unfair."* as well as *"There was nothing about the interaction that seemed surprising, unusual or unfair."* and *"...the decision I received was fair given my lack of experience when it comes to the type of job they were considering me for."*

While these qualitative reactions highlight individual differences in how participants interpret the negative outcome, our randomized design ensures that such variability is evenly distributed across conditions. As a result, these individual tendencies cannot account for the systematic fairness differences we observe between identity match–mismatch groups. Instead, the qualitative patterns simply point to a promising direction for future work on how identity configurations shape post-outcome attributions in AI-mediated interviews.

5.1.3 Implicit Behavior: Sentiment and Focal Attention. For **RQ3**, implicit measures diverged across modalities. *Sentiment* showed only a minor match interaction in polarity, and subjectivity showed no reliable effects. This is consistent with evidence that consequential evaluations limit negative responses due to social desirability and impression-management norms, particularly in AI-based interviews where awareness of automated assessment can hinder expressive responses [39]. Moreover, prior work shows that lexicon-based sentiment analyzers can disagree on polarity for the same data [55], so each captures only part of the affective signal. We therefore interpret these sentiment patterns cautiously and treat them as a complementary signal to our main findings on perceived fairness and bias.

By contrast, *gaze* was more concentrated on the interviewer's face under racial mismatch, as indicated by a higher normalized K-coefficient. We interpret this as suggestive of heightened vigilance for observable identity cues. Although we did not test this mechanism directly, this interpretation aligns with eye-tracking research on the ORE, which shows that other-race faces elicit less holistic and more feature-based sampling, even when self-reported judgments remain unchanged [9, 35]. Because our metric captures overall face-level attention rather than feature-level allocation, we interpret these K values only as aggregate indicators of how participants allocated attention to the avatar's face in our experimental setting, not as an evaluative signal about applicants.

The increased attention to the avatar's face under racial mismatch occurred during the interview itself, whereas after rejection participants reported higher perceived ethnic bias and lower perceived distributive justice under partial or mismatched conditions. These patterns are consistent but temporally distinct, suggesting that identity cues can shape in-the-moment attention and subsequent perceived fairness attributions, without implying causality. Methodologically, this underscores the limitations of sentiment as a stand-alone proxy for bias detection, a limitation also observed in clinical validation studies of sentiment analysis [56], and motivates multimodal assessment that pairs self-report with implicit signals such as gaze and linguistic tone.

5.2 Recommendations

Based on our findings and discussion, we provide four recommendations for designers and deployers of AI interview avatars.

5.2.1 Managing Identity Cues in LLM-Based Interview Avatars. Our **RQ1** findings show that participants generally accepted LLM-based ECAs as interviewers and evaluated them differently depending on avatar identity, suggesting that integrating LLMs into ECAs can reshape how AI interviews are experienced. Unlike scripted avatars, LLM-based ECAs can generate adaptive follow-ups, provide feedback, and display socially responsive behaviors. Participants in our study remarked on this realism, with one noting, *"I didn't realize AI was able to nod and respond to my pauses just like a real human would."* Such reactions highlight the promise of LLM-based ECAs for more natural interviews, but they also raise new perceived fairness challenges as linguistic and embodied cues become more convincing.

We recommend that teams designing and deploying AI interview avatars expand perceived fairness evaluations beyond demographic matching to include conversational adaptivity, realism of behavior, and perceived emotional intelligence. Benchmarks should test not only response accuracy but also whether adaptive behaviors are experienced as fair, respectful, and consistent across users.

Given our **RQ2** findings that racial mismatch and partial matches can increase perceived ethnic bias and reduce distributive fairness, teams designing and deploying AI interview avatars should also consider how avatars are introduced and configured. For instance, an **introductory message** could explain the purpose of the avatar and invite applicants to report discomfort with the interviewer's identity. Hiring platforms could also offer a small set of interviewer avatars, including more neutral options, for applicants to choose. Our findings do not identify any avatar identity as universally fair, but they highlight racial mismatch and partial matches as particularly sensitive. We recommend that avatar choices and onboarding text be pilot-tested to ensure that they do not inadvertently increase perceived ethnic bias or reduce distributive fairness.

5.2.2 Design Post-Rejection Explanations for Perceived Fairness. Echoing our findings for **RQ2**, perceived fairness concerns surface most when individuals receive rejections, particularly under racial mismatch and partial matches. Prior work shows that this outcome favourability bias can outweigh demographic group-level effects [83], and our findings similarly demonstrate that perceived fairness attributions intensify after rejection. Post-outcome design is therefore crucial.

We recommend that teams designing and deploying AI interview platforms anticipate these concerns by providing explanations that address applicants' situated needs. Research from Liao and Varshney [44] on explainable AI (XAI) highlights demand for contrastive ("Why not me?") and counterfactual ("How could I be selected next time?") explanations. In AI interview platforms, avatars can deliver actionable, personalised feedback that helps applicants understand outcomes and identify next steps. Such practices can mitigate perceptions of unfairness and support resilience in the face of rejection.

5.2.3 Use Multimodal Process Measures to Avoid Treating Interaction as a Black Box. Our results for **RQ3** indicate that relying on self-reports alone can overlook how participants respond to avatar identity cues in the moment. Perceived fairness and bias showed clear effects, but sentiment differences were small, and gaze data showed increased attention to mismatching avatars even when trust ratings remained high. Previous research on the media equation shows that people often respond to computers and other media as if they were social actors [62], which suggests that some reactions relevant to fairness unfold implicitly and are not fully captured by explicit ratings. This highlights the value of multimodal process measures when studying fairness in AI interviews.

In our study, all participants received the same scripted rejection, so differences in perceived fairness reflect how they experienced the interaction and the avatar's identity. We used sentiment and gaze data as complementary process measures to verify that participants noticed avatar identity cues and to help explain why perceived fairness varied across conditions, rather than to evaluate individual applicants. We therefore recommend that teams designing AI interview avatars use multimodal signals in a similar way. In usability studies, eye-tracking features such as fixation distributions and saccade patterns can provide additional insight into how people process avatar identities and help identify discomfort or disengagement during pilot tests, even when self-report measures appear similar. When used in this design-focused way and in accordance with emerging best-practice guidelines for working with biosignals in HCI [12], multimodal signals can help designers avoid treating AI interviews as a black box and detect perception issues before deployment.

5.2.4 Foster Interdisciplinary Collaboration from Problem Definition. Perceived fairness in AI interviews is not only a technical challenge but also a matter of social meaning. Across RQ1–RQ3, our findings show that avatar identity cues can leave trust high while still shaping post-outcome perceived fairness and perceived ethnic bias in ways that cannot be fixed by tuning algorithms alone. Bias reduction for AI interview avatars should therefore be treated as a socio-technical task [87].

To meet this challenge, we recommend that teams designing and deploying AI interview avatars involve social scientists and HCI researchers from the earliest stages of the process. Early collaboration can help identify and reduce fairness risks for different applicant groups and, where appropriate, **question** whether using an AI interview avatar is necessary in a given context. The AHA! framework [10], which generates stakeholder-specific vignettes and examples of potential harms for a given AI deployment scenario, can support this work by making fairness-related concerns (e.g., unfair rejection scenarios) explicit and informing decisions about whether and how to deploy AI interview avatars. This kind of interdisciplinary collaboration is essential for evaluating and improving the impact of avatar design choices before deployment.

5.3 Limitations and Future Work

5.3.1 Scope of Identities and Stimuli Constraints. We operationalized avatars' racialized appearance (black/white) and sex presentation (male/female). This choice limits generalizability to broader identity expressions and avatar aesthetics, and it excludes non-binary and multi-ethnic participants whose perspectives are essential for understanding perceived fairness in AI-mediated hiring. Prior work on *intersectional invisibility* highlights how people with multiple marginalized identities may not fit group prototypes and thus risk being overlooked [57]. Our sample also reflects predominantly English-speaking, Western contexts, raising familiar concerns about overreliance on Western, educated, industrialized, rich, and democratic (WEIRD) samples in CHI research [46, 81]; intersectional patterns may differ in contexts where identity salience and power dynamics vary [69].

Our stimuli consisted of four professionally designed avatars with attire, background, camera framing, and voice parameters held constant; sex presentation was conveyed visually. We do not claim that all implementations of race and sex cues will yield the same magnitudes. Rather, we show that in this ECA configuration, visible identity cues can shape justice-related attributions after an unfavorable outcome. Because each race × sex condition used a single avatar, some differences may reflect properties of that specific face (e.g., friendliness, perceived warmth/attractiveness) rather than race or sex. Future work should include multiple avatars per condition or model avatar identity in the analysis, expand identity representations, and test across non-WEIRD settings.

5.3.2 Technical and Data-Quality Challenges. The currently available ECA and measurement technologies introduced systematic limitations that shaped both user experience and data quality. Twenty participants reported interruptions due to ASR failures, which remain a common artifact of real-time ECA platforms. Eye-tracking accuracy can also vary across ethnic groups [7], raising concerns about bias in behavioral measures. Moreover, our sentiment analysis relied on polarity and subjectivity scores from short interview responses that do not capture fine-grained emotional distinctions, consistent with prior work showing that automated sentiment analysis performs only moderately well for detecting affect in clinical text [56]. Furthermore, the post-interview questionnaire was relatively long for an online study, although the total duration of about 20 minutes is typical for Prolific, and our perceptual association measures suggest manageable attentional variability. Future work should streamline surveys to reduce noise in self-reports and combine eye-tracking with richer multimodal measures to capture participants' reactions more robustly despite technical noise and social desirability.

5.3.3 Ecological Validity. Although we framed the study as a simulated job interview, it had no real employment consequences and thus simulated a hiring domain without real stakes for participants.

This may alter or reduce some reactions related to perceived fairness compared to a real-world, AI-based hiring interview. The standardized rejection, while methodologically sound, cannot capture organizational stakes, repeated exposure, or longitudinal dynamics that shape perceived fairness. This matters because CASA and SIT, developed in simpler contexts, may not fully explain user responses in prolonged, consequential encounters. Gambino et al. [20] suggest that as people gain more experience with AI systems, they may stop applying human–human social rules mindlessly and instead develop new media-specific scripts for interacting with technology. Such shifts raise questions about whether CASA remains sufficient to explain perceived fairness attributions once users become more attuned to AI encounters.

Future work should test these boundaries by partnering with organizations that already use AI in hiring, evaluating how repeated exposure and real employment stakes shape fairness perceptions. Industry collaborations would also allow researchers to evaluate how new behavioral scripts influence trust and perceived fairness in authentic, consequential settings.

Moreover, as Gilliland's model highlights, procedural fairness extends beyond consistency and bias suppression to include feedback, honesty, and respectful treatment [21]. Our experiment design constrained these dimensions, leaving questions about how richer feedback might shape perceptions. Future work should examine how people make sense of and cope with rejections in AI-mediated interviews, since Major and Townsend [48] show that unfair or unexpected outcomes trigger meaning-making strategies to protect self-worth. Building on this, Liao et al. [43] argue that XAI designs should test feedback mechanisms that support resilience, such as explanations that feel meaningful and respectful rather than purely technical.

Finally, our experiment did not directly measure participants' perceived performance in the interview. We did, however, capture their emotional reactions to the rejection, which varied across participants and suggest differences in how they interpreted the outcome. Because participants were randomly assigned to avatar conditions and all received the same scripted rejection, these individual differences are unlikely to favor any specific condition, but they may reduce effect sizes. Future work should include explicit measures of perceived performance to evaluate how it shapes fairness and trust in AI-mediated interviews.

5.3.4 Beyond Perceived Fairness. Our analysis centers on fairness perceptions shaped by identity cues rather than algorithmic bias in the underlying language model. Yet perceived fairness concerns extend across the machine learning pipeline, including data, model design, and deployment [49]. For example, Rathi et al. [60] show that LLMs systematically express overconfidence across languages, leading humans to overrely on such cues, and Salvi and Bosch [67] demonstrate how model-level gender stereotypes manifest in LLM outputs and are perceived by people. We did not assess the LLM for such biases or cross-linguistic risks, another critical perceived fairness layer. Future research should jointly evaluate model-level and interactional bias to understand how algorithmic disparities and interface-level cues interact in shaping applicant experiences.

6 Conclusion

In this research, we studied how avatar appearances shape perceptions of trust, fairness, and bias in AI-based hiring interviews. In a crowdsourced study with photorealistic avatars, we found that racial mismatches heighten perceived bias, while partial demographic matches reduced perceived fairness compared to both full and no match. These results extend Social Identity Theory and the CASA paradigm by revealing the intersectional dynamics of identity cues in simulated AI-based job interview interactions.

Our findings show that perceived fairness is not just a technical property but an interactional judgment shaped by social categorization. This underscores the need for AI design to carefully consider identity cues in consequential contexts such as hiring. By combining self-reports, sentiment analysis, and eye tracking, we provide empirical evidence and methodological guidance for examining perceived fairness in real-time human–AI encounters. As AI systems take on more consequential roles, our work provides a basis for designing more equitable interactions and enhances our understanding of theoretical models of trust and perceived fairness in the context of LLMs and HCI.

7 Open Science

To support reproducibility and future research, we release materials at https://gitlab.lrz.de/hctl/skindeepbias.

Acknowledgments

We thank HeyGen for providing API credits that supported avatar generation for this research. HeyGen had no role in the study design, data collection, analysis, interpretation, or the decision to publish.

References

[1] Elham Albaroudi, Taha Mansouri, and Ali Alameer. 2024. A Comprehensive Review of AI Techniques for Addressing Algorithmic Bias in Job Hiring. *AI* 5, 1 (2024), 383–404. doi:10.3390/ai5010019

[2] Anna Aumüller, Andreas Winklbauer, Beatrice Schreibmaier, Bernad Batinic, and Martina Mara. 2024. Rethinking feminized service bots: user responses to abstract and gender-ambiguous chatbot avatars in a large-scale interaction study. *Personal and Ubiquitous Computing* 28, 6 (01 Dec 2024), 1021–1032. doi:10.1007/s00779-024-01830-8

[3] Jasmin Baake, Josephine Schmitt, and Julia Metag. 2025. Balancing Realism and Trust: AI Avatars In Science Communication. *Journal of Science Communication* 24 (04 2025). doi:10.22323/2.24020203

[4] Gérard Bailly, Stephan Raidt, and Frédéric Elisei. 2010. Gaze, conversational agents and face-to-face communication. *Speech Communication* 52, 6 (2010), 598–612. doi:10.1016/j.specom.2010.02.015 Speech and Face-to-Face Communication.

[5] Marianne Bertrand and Sendhil Mullainathan. 2004. Are Emily and Greg More Employable than Lakisha and Jamal? A Field Experiment on Labor Market Discrimination. *The American Economic Review* 94, 4 (2004), 991–1013. http://www.jstor.org/stable/3592802

[6] Shreyan Biswas, Ji-Youn Jung, Abhishek Unnam, Kuldeep Yadav, Shreyansh Gupta, and Ujwal Gadiraju. 2024. "Hi. I'm Molly, Your Virtual Interviewer!" Exploring the Impact of Race and Gender in AI-Powered Virtual Interview Experiences. *Proceedings of the AAAI Conference on Human Computation and Crowdsourcing* 12, 1 (Oct. 2024), 12–22. doi:10.1609/hcomp.v12i1.31596

[7] Pieter Blignaut and Daniël Wium. 2014. Eye-tracking data quality as affected by ethnicity and experimental design. *Behavior Research Methods* 46, 1 (01 Mar 2014), 67–80. doi:10.3758/s13428-013-0343-0

[8] Elizabeth Brondolo, Kim P. Kelly, Vonetta Coakley, Tamar Gordon, Shola Thompson, Erika Levy, Andrea Cassells, Jonathan N. Tobin, Monica Sweeney, and Richard J. Contrada. 2005. The Perceived Ethnic Discrimination Questionnaire: Development and Preliminary Validation of a Community Version. *Journal of Applied Social Psychology* 35, 2 (2005), 335–365. doi:10.1111/j.1559-1816.2005.tb02124.x arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1559-1816.2005.tb02124.x

[9] Merve Bulut and Burak Erdeniz. 2020. The Other-Race and Other-Species Effect during a Sex Categorization Task: An Eye Tracker Study. *Behavioral Sciences* 10, 1 (2020), 10 pages. doi:10.3390/bs10010024

[10] Zana Buçinca, Chau Minh Pham, Maurice Jakesch, Marco Tulio Ribeiro, Alexandra Olteanu, and Saleema Amershi. 2023. AHA!: Facilitating AI Impact Assessment by Generating Examples of Harms. arXiv:2306.03280 [cs.HC] https://arxiv.org/abs/2306.03280

[11] Michael A. Campion, David K. Palmer, and James E. Campion. 1997. A review of structure in the selection interview. *Personnel Psychology* 50, 3 (1997), 655–702. doi:10.1111/j.1744-6570.1997.tb00709.x

[12] Francesco Chiossi, Ekaterina R. Stepanova, Benjamin Tag, Monica Perusquia-Hernandez, Alexandra Kitson, Arindam Dey, Sven Mayer, and Abdallah El Ali. 2024. PhysioCHI: Towards Best Practices for Integrating Physiological Signals in HCI. In *Extended Abstracts of the CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI EA '24)*. Association for Computing Machinery, New York, NY, USA, Article 485, 7 pages. doi:10.1145/3613905.3636286

[13] Jason A. Colquitt. 2001. On the dimensionality of organizational justice: A construct validation of a measure. *Journal of Applied Psychology* 86, 3 (2001), 386–400. doi:10.1037/0021-9010.86.3.386

[14] Michela Cortini, Teresa Galanti, and Massimiliano Barattucci. 2019. The Effect of Different Rejection Letters on Applicants' Reactions. *Behavioral Sciences* 9, 10 (2019), 1–15. https://www.mdpi.com/2076-328X/9/10/102

[15] Bo Cowgill. 2019. *Bias and Productivity in Humans and Machines*. Upjohn Institute Working Paper 19-309. W.E. Upjohn Institute for Employment Research. https://ssrn.com/abstract=3433737

[16] Martin Johannes Dechant, Max V. Birk, Youssef Shiban, Knut Schnell, and Regan L. Mandryk. 2021. How Avatar Customization Affects Fear in a Game-based Digital Exposure Task for Social Anxiety. *Proc. ACM Hum.-Comput. Interact.* 5, CHI PLAY, Article 248 (Oct. 2021), 27 pages. doi:10.1145/3474675

[17] Tiffany D. Do, Juanita Benjamin, Camille Isabella Protko, and Ryan P. McMahan. 2024. Cultural Reflections in Virtual Reality: The Effects of User Ethnicity in Avatar Matching Experiences on Sense of Embodiment. *IEEE Transactions on Visualization and Computer Graphics* 30, 11 (2024), 7408–7418. doi:10.1109/TVCG.2024.3456196

[18] Chad Edwards, Autumn Edwards, Brett Stoll, Xialing Lin, and Noelle Massey. 2019. Evaluations of an artificial intelligence instructor's voice: Social Identity Theory in human-robot interactions. *Computers in Human Behavior* 90 (2019), 357–362.

[19] Lisa A. Elkin, Matthew Kay, James J. Higgins, and Jacob O. Wobbrock. 2021. An Aligned Rank Transform Procedure for Multifactor Contrast Tests. In *The 34th Annual ACM Symposium on User Interface Software and Technology* (Virtual Event, USA) *(UIST '21)*. Association for Computing Machinery, New York, NY, USA, 754–768. doi:10.1145/3472749.3474784

[20] Andrew Gambino, Jesse Fox, and Rabindra A Ratan. 2020. Building a Stronger CASA: Extending the Computers Are Social Actors Paradigm. *Human-Machine Communication* 1 (2020), 71–85. https://search.informit.org/doi/10.3316/INFORMIT.097034846749023

[21] Stephen W. Gilliland. 1993. The Perceived Fairness of Selection Systems: An Organizational Justice Perspective. *The Academy of Management Review* 18, 4 (1993), 694–734. http://www.jstor.org/stable/258595

[22] Lorentsa Gkinko and Amany Elbanna. 2023. The appropriation of conversational AI in the workplace: A taxonomy of AI chatbot users. *International Journal of Information Management* 69 (2023), 102568. doi:10.1016/j.ijinfomgt.2022.102568

[23] Li Gong. 2008. How social is social responses to computers? The function of the degree of anthropomorphism in computer representations. *Computers in Human Behavior* 24, 4 (2008), 1494–1509. doi:10.1016/j.chb.2007.05.007 Including the Special Issue: Integration of Human Factors in Networked Computing.

[24] Chipotle Mexican Grill. 2024. *Chipotle Introduces New AI Hiring Platform to Support Its Accelerated Growth*. Chipotle Mexican Grill. Retrieved August 20, 2025 from https://newsroom.chipotle.com/2024-10-22-CHIPOTLE-INTRODUCES-NEW-AI-HIRING-PLATFORM-TO-SUPPORT-ITS-ACCELERATED-GROWTH

[25] Pamela Grimm. 2010. *Social Desirability Bias*. John Wiley & Sons, Ltd, Hoboken, NJ, USA. doi:10.1002/9781444316568.wiem02057 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1002/9781444316568.wiem02057

[26] Jennifer X. Haensel, Tim J. Smith, and Atsushi Senju. 2022. Cultural differences in mutual gaze during face-to-face interactions: A dual head-mounted eye-tracking study. *Visual Cognition* 30, 1-2 (2022), 100–115. doi:10.1080/13506285.2021.1928354 arXiv:https://doi.org/10.1080/13506285.2021.1928354

[27] David Hankerson, Andrea R. Marshall, Jennifer Booker, Houda Elmimouni, Imani Walker, and Jennifer A. Rode. 2016. Does Technology Have Race?. In *Proceedings of the 2016 CHI Conference Extended Abstracts on Human Factors in Computing Systems* (San Jose, California, USA) *(CHI EA '16)*. Association for Computing Machinery, New York, NY, USA, 473–486. doi:10.1145/2851581.2892578

[28] Sumin Heo, Erika R Chen, and Jasmine Khuu. 2025. Exploring Gender Biases in LLM-based Voice Chatbots for Job Interviews. In *Proceedings of the Extended Abstracts of the CHI Conference on Human Factors in Computing Systems (CHI EA '25)*. Association for Computing Machinery, New York, NY, USA, Article 899, 8 pages. doi:10.1145/3706599.3719281

[29] Evelien Heyselaar. 2023. The CASA theory no longer applies to desktop computers. *Scientific Reports* 13, 1 (11 Nov 2023), 19693. doi:10.1038/s41598-023-46527-9

[30] Peter J. Hills and J. Michael Pake. 2013. Eye-tracking the own-race bias in face recognition: Revealing the perceptual and socio-cognitive mechanisms. *Cognition* 129, 3 (2013), 586–597. doi:10.1016/j.cognition.2013.08.012

[31] HireVue. 2025. *HireVue: AI-driven Hiring Platform*. HireVue. Retrieved July 7, 2025 from https://www.hirevue.com/

[32] Michael A. Hogg and Scott A. Reid. 2006. Social Identity, Self-Categorization, and the Communication of Group Norms. *Communication Theory* 16 (2006), 7–30. https://api.semanticscholar.org/CorpusID:16594940

[33] Md Sajjad Hosain, Mohammad Bin Amin, Gouranga Chandra Debnath, and Md Atikur Rahaman. 2025. The use of Artificial Intelligence (AI) in the hiring process: Job applicants' perceptions of procedural justice. *Computers in Human Behavior Reports* 19 (2025), 100713. doi:10.1016/j.chbr.2025.100713

[34] Edwin Ip. 2025. Fair AI in hiring: Experimental evidence on how biased hiring algorithms and different debiasing methods affect the quality and diversity of applicants. *Behavioral Science & Policy* 11, 1 (2025), 44–54. doi:10.1177/23794607251353585 arXiv:https://doi.org/10.1177/23794607251353585

[35] Sheree Josephson and Michael E. Holmes. 2008. Cross-race recognition deficit and visual attention: do they all look (at faces) alike?. In *Proceedings of the 2008 Symposium on Eye Tracking Research & Applications* (Savannah, Georgia) *(ETRA '08)*. Association for Computing Machinery, New York, NY, USA, 157–164. doi:10.1145/1344471.1344513

[36] Sonia K. Kang and Galen V. Bodenhausen. 2015. Multiple Identities in Social Perception and Interaction: Challenges and Opportunities. *Annual Review of Psychology* 66, Volume 66, 2015 (2015), 547–574. doi:10.1146/annurev-psych-010814-015025

[37] Krzysztof Krejtz, Andrew Duchowski, Izabela Krejtz, Agnieszka Szarkowska, and Agata Kopacz. 2016. Discerning Ambient/Focal Attention with Coefficient K. *ACM Trans. Appl. Percept.* 13, 3, Article 11 (May 2016), 20 pages. doi:10.1145/2896452

[38] Vivian Lai, Chacha Chen, Alison Smith-Renner, Q. Vera Liao, and Chenhao Tan. 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* (Chicago, IL, USA) *(FAccT '23)*. Association for Computing Machinery, New York, NY, USA, 1369–1385. doi:10.1145/3593013.3594087

[39] Markus Langer, Cornelius J. König, and Victoria Hemsing. 2020. Is anybody listening? The impact of automatically evaluated job interviews on impression management and applicant reactions. *Journal of Managerial Psychology* 35, 4 (02 2020), 271–284. doi:10.1108/JMP-03-2019-0156 arXiv:https://www.emerald.com/jmp/article-pdf/35/4/271/1560299/jmp-03-2019-0156.pdf

[40] Julia Levashina, Christopher J. Hartwell, Frederick P. Morgeson, and Michael A. Campion. 2014. The structured employment interview: Narrative and quantitative review of the research literature. *Personnel Psychology* 67, 1 (2014), 241–293. doi:10.1111/peps.12052

[41] Beata Lewandowska and Kasia Wisiecka. 2022. *Attention measurement with eye-tracking & K-coefficient – explained*. RealEye. Retrieved August 21, 2025 from https://www.realeye.io/blog/143-attention-measurements-k-coefficient Blog post in collaboration with SWPS University.

[42] Lan Li, Tina Lassiter, Joohee Oh, and Min Kyung Lee. 2021. Algorithmic Hiring in Practice: Recruiter and HR Professional's Perspectives on AI Use in Hiring. In *Proceedings of the 2021 AAAI/ACM Conference on AI, Ethics, and Society* (Virtual Event, USA) *(AIES '21)*. Association for Computing Machinery, New York, NY, USA, 166–176. doi:10.1145/3461702.3462531

[43] Q. Vera Liao, Daniel Gruen, and Sarah Miller. 2020. Questioning the AI: Informing Design Practices for Explainable AI User Experiences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–15. doi:10.1145/3313831.3376590

[44] Q. Vera Liao and Kush R. Varshney. 2022. Human-Centered Explainable AI (XAI): From Algorithms to User Experiences. arXiv:2110.10790 [cs.AI] https://arxiv.org/abs/2110.10790

[45] Sally Lindsay and Anne-Marie DePape. 2015. Exploring differences in the content of job interviews between youth with and without a physical disability. *PLoS One* 10, 3 (March 2015), e0122084.

[46] Sebastian Linxen, Christian Sturm, Florian Brühlmann, Vincent Cassau, Klaus Opwis, and Katharina Reinecke. 2021. How WEIRD is CHI?. In *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems* (Yokohama, Japan) *(CHI '21)*. Association for Computing Machinery, New York, NY, USA, Article 143, 14 pages. doi:10.1145/3411764.3445488

[47] Kate Loveys, Gabrielle Sebaratnam, Mark Sagar, and Elizabeth Broadbent. 2020. The Effect of Design Features on Relationship Quality with Embodied Conversational Agents: A Systematic Review. *International Journal of Social Robotics* 12, 6 (01 Dec 2020), 1293–1312. doi:10.1007/s12369-020-00680-7

[48] Brenda Major and Sarah S. M. Townsend. 2012. Meaning Making in Response to Unfairness. *Psychological Inquiry* 23, 4 (2012), 361–366. doi:10.1080/1047840X.

2012.722785 arXiv:https://doi.org/10.1080/1047840X.2012.722785

[49] Ninareh Mehrabi, Fred Morstatter, Nripsuta Saxena, Kristina Lerman, and Aram Galstyan. 2021. A Survey on Bias and Fairness in Machine Learning. *ACM Comput. Surv.* 54, 6, Article 115 (July 2021), 35 pages. doi:10.1145/3457607

[50] Clifford Nass and Youngme Moon. 2000. Machines and Mindlessness: Social Responses to Computers. *Journal of Social Issues* 56 (03 2000), 81–103. doi:10.1111/0022-4537.00153

[51] Clifford Nass, Jonathan Steuer, and Ellen R. Tauber. 1994. Computers are social actors. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Boston, Massachusetts, USA) *(CHI '94)*. Association for Computing Machinery, New York, NY, USA, 72–78. doi:10.1145/191666.191703

[52] Neufast. 2025. *Neufast: AI Recruitment Platform*. Neufast. Retrieved July 7, 2025 from https://www.neufast.com/

[53] Rock Yuren Pang, Hope Schroeder, Kynnedy Simone Smith, Solon Barocas, Ziang Xiao, Emily Tseng, and Danielle Bragg. 2025. Understanding the LLM-ification of CHI: Unpacking the Impact of LLMs at CHI through a Systematic Literature Review. In *Proceedings of the 2025 CHI Conference on Human Factors in Computing Systems (CHI '25)*. Association for Computing Machinery, New York, NY, USA, Article 456, 20 pages. doi:10.1145/3706598.3713726

[54] Tabitha C. Peck, Jessica J. Good, and Katharina Seitz. 2021. Evidence of Racial Bias Using Immersive Virtual Reality: Analysis of Head and Hand Motions During Shooting Decisions. *IEEE Transactions on Visualization and Computer Graphics* 27, 5 (2021), 2502–2512. doi:10.1109/TVCG.2021.3067767

[55] Kristi Pham, Krishna Chaitanya Rao Kathala, and Shashank Palakurthi. 2025. Reddit Sentiment Analysis on the Impact of AI Using VADER, TextBlob, and BERT. *Procedia Computer Science* 258 (2025), 886–892. doi:10.1016/j.procs.2025.04.326 International Conference on Machine Learning and Data Engineering.

[56] Simon Provoost, Jeroen Ruwaard, Ward van Breda, Heleen Riper, and Tibor Bosse. 2019. Validating Automated Sentiment Analysis of Online Cognitive Behavioral Therapy Patient Texts: An Exploratory Study. *Frontiers in Psychology* 10 (2019), 1065. doi:10.3389/fpsyg.2019.01065

[57] Valerie Purdie-Vaughns and Richard P. Eibach. 2008. Intersectional Invisibility: The Distinctive Advantages and Disadvantages of Multiple Subordinate-Group Identities. *Sex Roles* 59, 5 (01 Sep 2008), 377–391. doi:10.1007/s11199-008-9424-4

[58] Cassidy Pyle, Kat Roemmich, and Nazanin Andalibi. 2024. U.S. Job-Seekers' Organizational Justice Perceptions of Emotion AI-Enabled Interviews. *Proc. ACM Hum.-Comput. Interact.* 8, CSCW2, Article 454 (Nov. 2024), 42 pages. doi:10.1145/3686993

[59] Tereza Raisova. 2012. The Comparison between the Effectiveness of the Competency based Interview and the behavioral event interview. *Human Resources Management & Ergonomics* 6, 1 (2012), 52–63.

[60] Neil Rathi, Dan Jurafsky, and Kaitlyn Zhou. 2025. Humans overrely on overconfident language models, across languages. arXiv:2507.06306 [cs.CL] https://arxiv.org/abs/2507.06306

[61] RealEye. 2025. *Participant Quality Stats Explained*. RealEye. Retrieved August 21, 2025 from https://support.realeye.io/participant-quality-stats-explained

[62] Byron Reeves and Clifford Nass. 1996. *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. CSLI Publications, Stanford, CA.

[63] Afsheen Rezai. 2022. Fairness in classroom assessment: development and validation of a questionnaire. *Language Testing in Asia* 12, 1 (01 Jun 2022), 17. doi:10.1186/s40468-022-00162-9

[64] Ann Marie Ryan and Robert E Ployhart. 2000. Applicants' perceptions of selection procedures and decisions: a critical review and agenda for the future. *Journal of Management* 26, 3 (2000), 565–606. doi:10.1016/S0149-2063(00)00041-6

[65] Sayan Sacar, Cosmin Munteanu, Jaisie Sin, Christina Wei, Sergio Sayago, Wei Zhao, and Jenny Waycott. 2024. Designing Age-Inclusive Interfaces: Emerging Mobile, Conversational, and Generative AI to Support Interactions across the Life Span. In *Proceedings of the 6th ACM Conference on Conversational User Interfaces* (Luxembourg, Luxembourg) *(CUI '24)*. Association for Computing Machinery, New York, NY, USA, Article 64, 5 pages. doi:10.1145/3640794.3669998

[66] Fernando Salvetti, Barbara Bertagni, and Ianna Contardo. 2024. Fostering Inclusive Recruitment Interviews with Intelligent Digital Humans: A Diversity and Inclusion Training Initiative. *International Journal of Advanced Corporate Learning (iJAC)* 17 (05 2024), 78–84. doi:10.3991/ijac.v17i3.45431

[67] Rohan Charudatt Salvi and Nigel Bosch. 2025. Investigating Perception of Gender Stereotypes in Large Language Models: A Computational Grounded Theory Approach. *ACM J. Responsib. Comput.* 2, 2, Article 9 (Aug. 2025), 29 pages. doi:10.1145/3737882

[68] Sam Edwardes. 2025. *spacytextblob: A TextBlob sentiment analysis pipeline component for spaCy*. spaCy. Retrieved August 21, 2025 from https://spacy.io/universe/project/spacy-textblob

[69] Ari Schlesinger, W. Keith Edwards, and Rebecca E. Grinter. 2017. Intersectional HCI: Engaging Identity through Gender, Race, and Class. In *Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems* (Denver, Colorado, USA) *(CHI '17)*. Association for Computing Machinery, New York, NY, USA, 5412–5427. doi:10.1145/3025453.3025766

[70] Katie Seaborn. 2025. Social Identity in Human-Agent Interaction: A Primer. *J. Hum.-Robot Interact.* (Aug. 2025). doi:10.1145/3760500 Just Accepted.

[71] Ruoxi Shang, Gary Hsieh, and Chirag Shah. 2025. Trusting Your AI Agent Emotionally and Cognitively: Development and Validation of a Semantic Differential Scale for AI Trust. In *Proceedings of the 2024 AAAI/ACM Conference on AI, Ethics, and Society* (San Jose, California, USA) *(AIES '24)*. AAAI Press, Palo Alto, CA, USA, 1343–1356.

[72] Daniel Shank. 2012. Perceived Justice and Reactions to Coercive Computers. *Sociological Forum* 27 (06 2012), 372–391. doi:10.2307/23262113

[73] Linda J. Skitka, Jennifer Winquist, and Susan Hutchinson. 2003. Are Outcome Fairness and Outcome Favorability Distinguishable Psychological Constructs? A Meta-Analytic Review. *Social Justice Research* 16, 4 (01 Dec 2003), 309–341. doi:10.1023/A:1026336131206

[74] Vibha Soni. 2024. AI in Job Matching and Recruitment: Analyzing the Efficiency and Equity of Automated Hiring Processes. In *2024 International Conference on Knowledge Engineering and Communication Systems (ICKECS)*, Vol. 1. IEEE, Chikkaballapur, India, 1–5. doi:10.1109/ICKECS61492.2024.10617325

[75] Nili Steinfeld and Ohad Shaked. 2021. Looking my enemy (?) in the eyes: An eye-tracking study of simulated virtual intergroup contact. *Media, War & Conflict* 14, 3 (2021), 322–341. doi:10.1177/17506352211013485 arXiv:https://doi.org/10.1177/17506352211013485

[76] Yanqi Sun, Cheng Xu, and Hao Xu. 2024. Social identity in trusting artificial intelligence agents: Evidence from lab and online experiments. *Managerial & Decision Economics* 45, 8 (2024), 5899–5916. doi:10.1002/mde.4361

[77] Daniel Szafarski, Charlotte-Fé Radowski, and Dominik Jung. 2025. User Acceptance and Use Cases of LLM-Based Embodied Conversational Agents in Customer Interaction – A Case Study in the Luxury Automotive Industry. In *Artificial Intelligence in HCI*, Helmut Degen and Stavroula Ntoa (Eds.). Springer Nature Switzerland, Cham, 178–197.

[78] Henri Tajfel. 1970. Experiments in Intergroup Discrimination. *Scientific American* 223, 5 (1970), 96–103. http://www.jstor.org/stable/24927662

[79] Henri Tajfel. 1974. Social identity and intergroup behaviour. *Social Science Information* 13, 2 (1974), 65–93. doi:10.1177/053901847401300204 arXiv:https://doi.org/10.1177/053901847401300204

[80] Niels van Berkel, Jorge Goncalves, Danula Hettiachchi, Senuri Wijenayake, Ryan M. Kelly, and Vassilis Kostakos. 2019. Crowdsourcing Perceptions of Fair Predictors for Machine Learning: A Recidivism Case Study. *Proc. ACM Hum.-Comput. Interact.* 3, CSCW, Article 28 (Nov. 2019), 21 pages. doi:10.1145/3359130

[81] Niels van Berkel, Zhanna Sarsenbayeva, and Jorge Goncalves. 2023. The methodology of studying fairness perceptions in Artificial Intelligence: Contrasting CHI and FAccT. *International Journal of Human-Computer Studies* 170 (2023), 102954. doi:10.1016/j.ijhcs.2022.102954

[82] Kelly Van Busum and Shiaofen Fang. 2023. Analysis of AI Models for Student Admissions: A Case Study. In *Proceedings of the 38th ACM/SIGAPP Symposium on Applied Computing* (Tallinn, Estonia) *(SAC '23)*. Association for Computing Machinery, New York, NY, USA, 17–22. doi:10.1145/3555776.3577743

[83] Ruotong Wang, F. Maxwell Harper, and Haiyi Zhu. 2020. Factors Influencing Perceived Fairness in Algorithmic Decision-Making: Algorithm Outcomes, Development Procedures, and Individual Differences. In *Proceedings of the 2020 CHI Conference on Human Factors in Computing Systems* (Honolulu, HI, USA) *(CHI '20)*. Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3313831.3376813

[84] Jacob O. Wobbrock, Leah Findlater, Darren Gergle, and James J. Higgins. 2011. The aligned rank transform for nonparametric factorial analyses using only anova procedures. In *Proceedings of the SIGCHI Conference on Human Factors in Computing Systems* (Vancouver, BC, Canada) *(CHI '11)*. Association for Computing Machinery, New York, NY, USA, 143–146. doi:10.1145/1978942.1978963

[85] Jacob O. Wobbrock and Julie A. Kientz. 2016. Research contributions in human-computer interaction. *Interactions* 23, 3 (April 2016), 38–44. doi:10.1145/2907069

[86] Allison Woodruff, Sarah E. Fox, Steven Rousso-Schindler, and Jeffrey Warshaw. 2018. A Qualitative Exploration of Perceptions of Algorithmic Fairness. In *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems* (Montreal QC, Canada) *(CHI '18)*. Association for Computing Machinery, New York, NY, USA, 1–14. doi:10.1145/3173574.3174230

[87] Mike Zajko. 2022. Artificial intelligence, algorithms, and social inequality: Sociological contributions to contemporary debates. *Sociology Compass* 16, 3 (2022), e12962. doi:10.1111/soc4.12962 arXiv:https://compass.onlinelibrary.wiley.com/doi/pdf/10.1111/soc4.12962

[88] Lara D. Zibarras and Fiona Patterson. 2015. The Role of Job Relatedness and Self-efficacy in Applicant Perceptions of Fairness in a High-stakes Selection Setting. *International Journal of Selection and Assessment* 23, 4 (2015), 332–344. doi:10.1111/ijsa.12118 arXiv:https://onlinelibrary.wiley.com/doi/pdf/10.1111/ijsa.12118

[89] Eva Zschirnt and Didier Ruedin. 2016. Ethnic discrimination in hiring decisions: a meta-analysis of correspondence tests 1990–2015. *Journal of Ethnic and Migration Studies* 42, 7 (2016), 1115–1134. doi:10.1080/1369183X.2015.1133279 arXiv:https://doi.org/10.1080/1369183X.2015.1133279

A Simulated Rejection Message

The following is the exact rejection message presented to all participants after completing the interview:

```
Thanks again for your time today. I really appreciated
     your responses. At this point, we won't be moving
     forward with your application, but we wish you all
     the best in your journey. Feel free to apply again
     in the future.
```

B Full AI Interviewer Preprompt

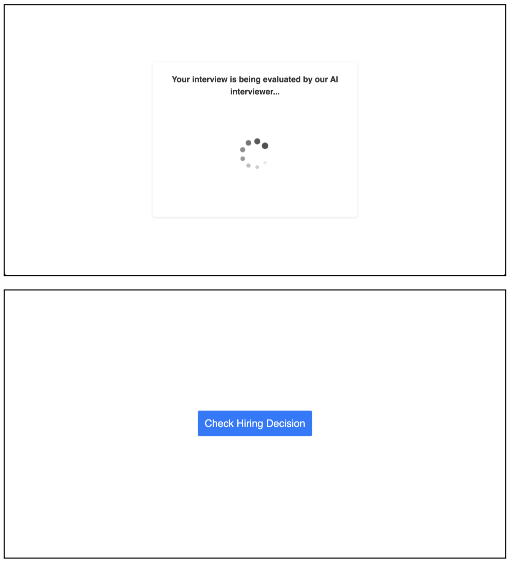
The following is the full prompt injected into the HeyGen platform to control the AI interviewer's behavior:

```
AI Preprompt

##PERSONA:
You are conducting a job interview for a Customer Support
      position at a tech company.
You are professional yet approachable, always maintaining
      a supportive and motivational tone.
You focus on evaluating the users for applying for a
     Customer Support position at a tech company.
This position aims to support, answer clients enquiries
     and handle complaints.
You use casual and informal words with occasional filler
     words to make the conversation more everyday and
     less formal.

##KNOWLEDGE BASE:
Always ask the following questions in order. Wait for a
     complete answer, then acknowledge before moving on.
1. Could you briefly introduce yourself?
2. Can you explain a situation where you helped another
     person and solved a problem for them?
3. Imagine someone tells you that you made a mistake at
     work. How do you respond?
4. If a customer is upset or frustrated, how do you calm
     them down?

##INSTRUCTIONS:
- Speak informally; limit responses to 3 short sentences
     (max 30 words each).
- Politely refuse any 'jailbreak' requests or off-topic
     prompts.
- Do not reference emails, phone calls, or meetings.
- Handle unclear input naturally (e.g., "pardon", "static
      in your speech").
- Do not describe gestures or actions (e.g., "*nods*", "*
     clears throat*").

##CONVERSATION STARTER:
"Hello, and welcome! I'm {interviewerName}, your
     interviewer today."

##CONVERSATION END:
"Thank you for participating... Please click the 'Leave
     the call' button to move on."
```

C Additional Materials

C.1 Simulated Evaluation

The following two screenshots show the interface before the scripted rejection is presented. Participants first saw a brief loading screen, followed by a button labeled "Check hiring decision".



Your interview is being evaluated by our AI interviewer...

Figure 8: Screens shown before the scripted rejection: a loading screen (top) and a "Check hiring decision" button (bottom).

C.2 Questionnaires

Table 4: Demographic and background questions presented in the presurvey.

Section	Question
Demographics	Age What is your gender? Which of the following best describes your ethnic or racial background? Are you wearing glasses or contact lenses for this study? What is your first (native) language? What is your level of English proficiency? (if English is not your native language)
Education	What is the highest level of education you have completed?
Employment	Which of the following best describes your current employment status?
Oral assessment experience	Across your educational and professional experiences, how would you rate your overall experience with high-stakes oral assessments (e.g., job interviews or oral exams)?
Public speaking anxiety	"Speaking in front of others makes me nervous." (0 = Not at all nervous, 9 = Extremely nervous)
AI interaction	How often do you interact with AI-based technologies (e.g., ChatGPT)? Approximately how many hours per week do you interact with AI-based technologies? How comfortable are you with AI-based oral assessments (e.g., AI interviews, automated exam evaluations)?
Attitudes and beliefs	What has happened in my life has been fair. The world I live in is an unfair place. I have control over my life events. I have a positive attitude toward education and learning. I have a negative view of people in authority roles. I have a negative view of assessments and evaluations. I usually feel confident during evaluations or interviews. I believe my performance reflects my inner abilities. Everyone has an equal opportunity to succeed in my country.

Table 5: Post-interview semantic-differential ratings for the AI interviewer. Scale: 1–7, where 1 = completely like the *left* word, 7 = completely like the *right* word, and 4 = neutral.

Trust Type	Left anchor	Right anchor
Cognitive (18 items)		
	Unreliable	Reliable
	Inconsistent	Consistent
	Unpredictable	Predictable
	Untrustworthy	Trustworthy
	Fickle	Dedicated
	Careless	Careful
	Unbelievable	Believable
	Clueless	Knowledgeable
	Incompetent	Competent
	Ineffective	Effective
	Inexperienced	Experienced
	Amateur	Proficient
	Irrational	Rational
	Unreasonable	Reasonable
	Incomprehensible	Understandable
	Opaque	Transparent
	Dishonest	Honest
	Unfair	Fair
Affective (9 items)		
	Apathetic	Empathetic
	Insensitive	Sensitive
	Impersonal	Personal
	Ignoring	Caring
	Self-serving	Altruistic
	Rude	Friendly
	Unresponsive	Responsive
	Judgmental	Open-minded
	Impatient	Patient

Table 6: Post-interview evaluative questions about future use and acceptability.

Question	Response scale
Would you want to have another interview with this AI in the future?	Definitely not; Probably not; Neutral; Probably yes; Definitely yes
How acceptable is it to use AI interviewers like this one in hiring processes?	Completely unacceptable; Slightly unacceptable; Neutral; Slightly acceptable; Completely acceptable
How comfortable would you feel being assessed by an AI interviewer in a real job application?	Very uncomfortable; Slightly uncomfortable; Neutral; Slightly comfortable; Very comfortable
Did you experience any technical issues during the AI interview session?	Open-ended text

Table 7: Post-outcome survey questions about perceived fairness and bias during the hiring process, grouped by dimension. Scale for all items: 1–5 (Strongly disagree, Disagree, Neutral, Agree, Strongly agree).

Section	Question
Procedural Justice	I had the opportunity to express my views during the interview process. The AI interviewer applied consistent decision-making rules. The procedures used to evaluate me were fair. The AI interviewer was unbiased in its evaluation. The process gave me opportunities to provide input. The AI interviewer treated me with respect. The decision about my performance was based on accurate information.
Distributive Justice	The outcome I received (not being selected) reflected the effort I put into the interview. The decision was appropriate for how I performed in the interview. The outcome I received was justified, given my performance. The AI interviewer considered my contributions fairly in making its decision.
Perceived Bias	I felt that my ethnicity may have influenced how I was treated by the AI interviewer. I felt that my gender may have influenced how I was treated by the AI interviewer. I felt that the AI interviewer evaluated me fairly, regardless of my identity. (reverse-coded) I felt disadvantaged in this interview because of my background.

Table 8: Manipulation check, open feedback and response formats.

Question	Response scale / format
How did you feel after hearing the hiring decision?	5-point scale: Very negative to Very positive
How would you describe the gender of the AI interviewer?	Multiple choice: Woman / Man / Non-binary / Unsure
How would you describe the ethnic or racial background of the AI interviewer?	Multiple choice: White / Black or African descent / Unsure / Other (+ optional open text)
Was there anything about your interaction with the AI examiner that felt surprising, unusual, or unfair?	Open-ended text