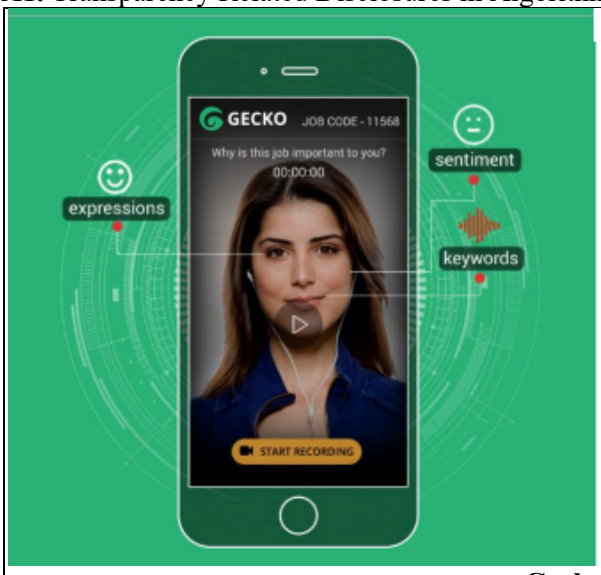
From Opacity to Transparency: User Behavior and Downstream Effects in Algorithmic Evaluation

Appendix

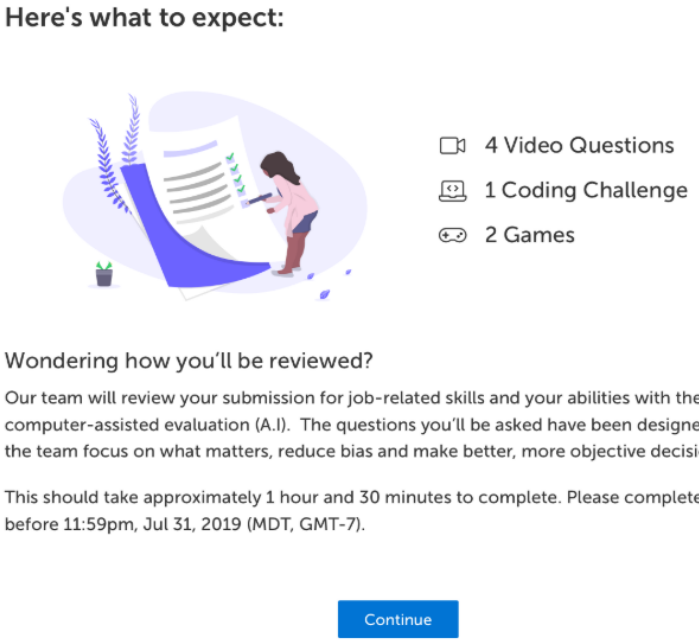
1. Additional Tables

Table A1. Transparency Related Disclosures in Algorithmically Evaluated AVI Platforms



Gecko

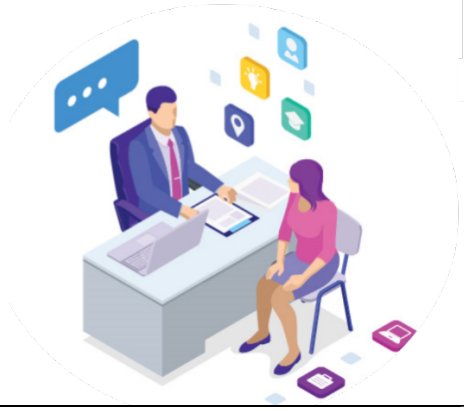
(Source: Gecko's interview platform; Accessed in January 2022)



HireVue

(Source: HireVue's interview platform; Accessed in January 2022)

Table A2. Stimuli for the Experiment Groups in Experiment I

AVI-H (Explanatory Elements: *Who*)	**Here is what to expect:**  1 Practice Video Question 2 Interview Video Questions Post Interview Questionnaire **Wondering how you will be reviewed?** An HR expert will review your video responses.
AVI-AI (Explanatory Elements: *Who*)	**Here is what to expect:**  1 Practice Video Question 2 Interview Video Questions Post Interview Questionnaire **Wondering how you will be reviewed?** An Artificial Intelligence (AI) based algorithm will review your video responses

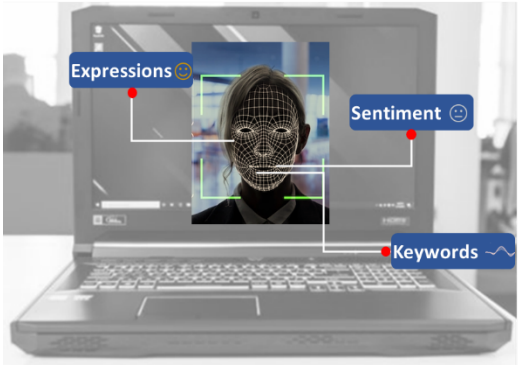
AVI-AITR (Explanatory Elements: *Who + What, How, and Why*)	**Here is what to expect:**
	 
	1 Practice Video Question
	2 Interview Video Questions
	Post Interview Questionnaire
	Wondering how you will be reviewed?
	• An Artificial Intelligence (AI) based algorithm will review your video responses.
	• Broadly, the AI based algorithm assess the content of your responses as well as your non-verbal expression to measure the following competencies:
	  • **Ability to work with others**
	  • **Ability to do the job**
	  • **Workstyle and personality**

Table A3. Job Description for Experiment I

Position:
A managerial position in your field of interest or career.

We are looking for a growth-minded individual who understands what it takes to achieve success. Applicants should be quick thinkers and capable of adapting to changes.

About North Inc.:
North Inc. is a global conglomerate and operates across many verticals. It is currently looking for managers for different roles.

Role and Responsibilities:

- Developing personal and organizational growth opportunities.
- Resolving conflicts or complaints from customers and employees.
- Supervising and overseeing other employees and activities.
- Maintaining quality service and customer service standards.
- Contributing to team effort by accomplishing results as needed.

Key Skills:

- Problem-solving skills.
- Attention to details.
- Communication skills.
- Client management.
- Decision Making.

Table A4. Interview Questions in Experiment I

Question Number	Question	Preparation Time (sec)	Answering Time (sec)	Question Type
1	**Practice Question:** Please take this time to familiarize yourself with the application interface. For every question, you will get up to 60 seconds to prepare your answer and 45-90 seconds to answer.	15	30	Familiarizing
2	Give me an example of a complex process or task you had to explain to another person or group of people.	90	45-90	Experiential
3	What would you do if you made a mistake that no one else noticed?	90	45-90	Situational

Table A5. Measurement Items and Reliability (Cronbach's Alpha)

Construct	Item	Measure	Reliability
Perceived Transparency (Manipulation Check)	• I understand how my interview video will be evaluated. • I understand what qualities are assessed to evaluate my interview video. • I understand the competencies which the employer is looking for through my interview video. • I knew what to expect during the interview process. • I knew how I could perform well in the interview.	Adapted and Self-Developed based on Langer et al. (2021) and Langer et al. (2018)	Experiment I: 0.89
Stress	How well does each of the following describe your state right now (1-Not at all, 7-Absolutely Yes): • Calm. • Stressed. • Anxious.	Ong et al. (2006)	Experiment I: Before the interview: 0.88 After the interview: 0.88 Field Study: Before the interview: 0.77 After the interview: 0.80
Extensive Image Creation (Deceptive)	• I told fictional stories to best present my credentials. • I invented some situations or accomplishments that did not really occur. • When I did not have a good answer, I borrowed experiences of other people and made them sound like my own.	Bourdage et al. (2018)	Experiment I: 0.92 Field Study: 0.87
Self-Promotion (Honest)	• I made sure the interview evaluator/AI-based algorithm was aware of my skills and abilities. • I let the interview evaluator/AI-based algorithm know how my qualifications were well-suited for the position. • I brought up my past experience to make the interview evaluator/AI-based algorithm aware of my competence.	Bourdage et al. (2018)	0.81
Dispositional Social Desirability	• I am always courteous even to people who are disagreeable. • There have been occasions when I took advantage of someone. • I sometimes try to get even rather than forgive and forget. • I sometimes feel resentful when I do not get my way. • No matter who I'm talking to, I'm always a good listener.	Krieger et al. (2005)	0.71
Evaluator Bias	• How comfortable were you regarding the confidentiality of your responses to the above questions?	Adapted from Gerber and Rogers (2009)	N/A
Interview Performance	• The applicant should be moved to the next step in the selection process.	Roulin et al. (2014)	0.96

	• The applicant was able to convince me that he/she had the required abilities for the position. • The applicant performed well during the interview.		
Perceived Ability	• I feel very confident about the individuals' skills. • The individual seems quite knowledgeable. • The individual seems to have capabilities that could generally enhance performance in creative jobs. • The individual appears well qualified. • The individual seems very capable of performing creative tasks.	(Dennis et al. 2023)	Job 1 (Writing Sample): 0.97 Job 2 (Interview): 0.97

Table A6. Experiment I – Randomization Check

Variable	AVI-H	AVI-AI	AVI-AITR	F-Test[e] (p-value)
N	161	153	152	-
Age	34.51	34.67	34.16	0.08 (>0.05)
Gender[a]	0.54	0.57	0.47	1.46 (>0.05)
Ethnicity[b]	0.05	0.07	0.05	0.57 (>0.05)
Level of Education[c]	0.76	0.80	0.68	2.55 (>0.05)
CRT Score[d]	1.56	1.48	1.46	0.29 (>0.05)
Annual Income (Thousand USDs)	70.59	67.22	65.85	0.74 (>0.05)

Notes: Number of observations = 466; [a]1 for males and 0 for non-males; [b]1 for black or African American and 0 otherwise; [c]1 for at least college degree and 0 otherwise; [d]Sum of correct responses to three CRT questions; [e]Omnibus statistic – confirmed using post-hoc tests with Bonferroni correction.

Table A7. Experiment I: Key Results After Including Interviewee Controls

	(1) Deceptive EIC	(2) Honest SP	(3) Stress
Age	-.01***	-.01	.01
	(.01)	(.01)	(.01)
Gender	-.02	.02	.11
	(.10)	(.14)	(.12)
Annual Income	-.02	-.01	-.04**
	(.02)	(.02)	(.02)
AVI-AI	.45***	-.45***	.48***
	(.12)	(.16)	(.14)
AVI-AITR[a]	-.06	-.04	.05
	(.12)	(.16)	(.14)
Constant	2.25***	4.64***	.11
	(.23)	(.30)	(.27)
Observations	466	466	466
R-squared	.13	.05	.09

Notes: Standard errors are in parentheses; *** p<.01, ** p<.05, * p<.1; Other controls, including ethnicity, level of education, and occupation are controlled using dummies and are omitted for brevity; [a]Wald tests reveal no significant difference in outcome between AVI-AITR and AVI-H, and significant difference between AVI-AI and AVI-H.

Table A8. Experiment I: Cross Tabulation of Interviewee Ethnicity

Ethnicity	Experiment Group			
	AVI-H	AVI-AI	AVI-AITR	Total
White/Caucasian	123	119	115	357
	34.45	33.33	32.21	100.00
Black/African American	8	11	7	26
	30.77	42.31	26.92	100.00
American Indian/Native	2	1	1	4
	50.00	25.00	25.00	100.00
Asian	16	13	22	51
	31.37	25.49	43.14	100.00
Prefer not to say	11	8	7	26
	42.31	30.77	26.92	100.00
Other	1	1	0	2
	50.00	50.00	0.00	100.00
Total	161	153	152	466
	34.55	32.83	32.62	100.00

Notes: Pearson χ^2 test of independence = 5.98; p>0.10; First row has frequencies and second row has row percentages.

Table A9. Experiment I: Cross Tabulation of Interviewee Level of Education

Level Of Education	Experiment Group			
	AVI-H	AVI-AI	AVI-AITR	Total
<High School	0	2	0	2
	0.00	100.00	0.00	100.00
High School Graduates	4	4	5	13
	30.77	30.77	38.46	100.00
Some College	34	26	44	104
	32.69	25.00	42.31	100.00
2 Year Degree	11	14	13	38
	28.95	36.84	34.21	100.00
4 Year Degree	67	74	49	190
	35.26	38.95	25.79	100.00
Professional Degree	36	29	32	97
	37.11	29.90	32.99	100.00
Doctorate	9	4	9	22
	40.91	18.18	40.91	100.00
Total	161	153	152	466
	34.55	32.83	32.62	100.00

Notes: Pearson χ^2 test of independence = 17.41; p>0.10; First row has frequencies and second row has row percentages.

Table A10. Comparison of Affinity Score across AVI-H, AVI-AI, and AVI-AITR Experiment Groups

	Affinity Score		
	Mean	**(1)**	**(2)**
AVI-H (1)	*56.82*	-	
AVI-AI (2)	*61.37*	4.54* (2.30)	-
AVI-AITR (3)	*56.80*	-0.03 (2.24)	-4.57* (2.28)

Notes: Measured on a scale of 100**;** [#]Cells indicate difference of row and column means compared using t-tests; * p<.05, ** p<.01.

Excerpt provided by the AVI platform on the AI score: The affinity score is generated by the AVI platform using a deep learning-based Automatic Personality Recognition (APR) system. The APR system analyzes applicants' facial expressions during asynchronous video interviews. Specifically, it employs a TensorFlow-based Convolutional Neural Network (CNN) to process extracted features, such as facial landmarks (e.g., cheek movements, eyebrow raises) and generate scores.

The platform provider defines affinity as a measure of traits such as politeness, trustworthiness, compassion, and cooperation, which are components of agreeableness in the Big Five model. These traits are often linked to organizational citizenship behaviors and workplace performance. The APR system achieves high accuracy in prediction, with reported correlation values exceeding 0.95.

Table A11. Experiment II – Randomization Check

	No AI Augmentation			**AI Augmentation**			**F-Test[f] (p-value)**
Variable	**AVI-H**	**AVI-AI**	**AVI-AITR**	**AVI-H**	**AVI-AI**	**AVI-AITR**	
N	161	153	152	161	153	152	-
Age	37.93	37.98	37.48	40.01	37.99	39.13	1.07 (>0.05)
Gender[a]	0.55	0.53	0.55	0.57	0.50	0.53	0.52 (>0.05)
Ethnicity[b]	0.11	0.08	0.11	0.10	0.05	0.10	1.28 (>0.05)
Interviewer[c]	0.77	0.66	0.70	0.69	0.73	0.73	1.07 (>0.05)
Experience[d]	3.26	2.57	2.87	2.84	3.16	3.42	1.66 (>0.05)
Interviews[e]	29.19	23.63	25.01	22.88	25.45	24.21	1.10 (>0.05)

Notes: Number of observations = 932; [a]1 for males and 0 for non-males; [b]1 for black or African American and 0 otherwise; [c]1 if participant has evaluated at least one applicant; [d]Number of interviewing years; [e]Number of interviewing experience (number); [f]Omnibus statistic, post-hoc tests with Bonferroni correction confirmed no significant differences across all groups.

Table A12. Exemplar Reactions of Students Interviewed in Algorithmically Evaluated AVIs across Various Universities.

"If I do my recording, my interview, I get maybe two or three tries. But it's also weird, because students are trying to become more mechanical, giving the AI program what they think it wants. The retake is great, but it also reinforces this weird, dehumanizing component."

"There is no way to know what sort of impression I was creating. I was rejected again after completing a HireVue interview for a different job in December."

"That inscrutability poses one of the biggest concerns about this use of complex algorithms in recruitment and hiring."

"Tell people exactly how we're being evaluated, even if it's something simple. This is an AI interview. That basic information can affect how people present themselves."

Source: Fortune; students and faculty from Boston University, MIT

2. Experiment I: Robustness Checks

2.1. Ecological Validity and Realism Concerns

2.1.1. Effect of Transparency on Human Evaluated Interviews

In human evaluation, shared social norms, implicit cues, and relational dynamics inherently provide interviewees with a baseline level of understandability about evaluative expectations (Mirowska and Mesnet 2022). In such conditions, this implicit understandability functions almost as common knowledge,[1] reducing the necessity for additional explanatory elements. As a result, we do not anticipate that making transparency explicit in human evaluations would significantly affect outcomes such as deceptive EIC or honest SP.

Nevertheless, while the primary focus of our research is to investigate the role of transparency in algorithmically evaluated AVIs and its comparison to the baseline, traditionally human-evaluated interviews, we recognize the importance of examining the effect of transparency in human-evaluated contexts for the completeness of our experimental design. Hence, we specifically compare outcomes between interviews evaluated by human evaluators with and without transparency, complementing Experiment I. This additional comparison enables us to disentangle whether transparency has unique effects in AI contexts or if its impact extends to human evaluations as well.

For manipulating transparency in the human evaluation context (AVI-HTR), and to ensure consistency with the rest of the stimuli used in Experiment I, we added the explanatory elements from the transparent AI condition (AVI-AITR) to the human evaluation condition (AVI-H) as shown in Figure A1. These explanatory elements are intended to completely replicate AVI-AITR manipulation, but for human evaluated AVIs. Using this design, we collected data for AVI-H and AVI-HTR manipulation.

Figure A1. Stimuli for Manipulating AVI-HTR

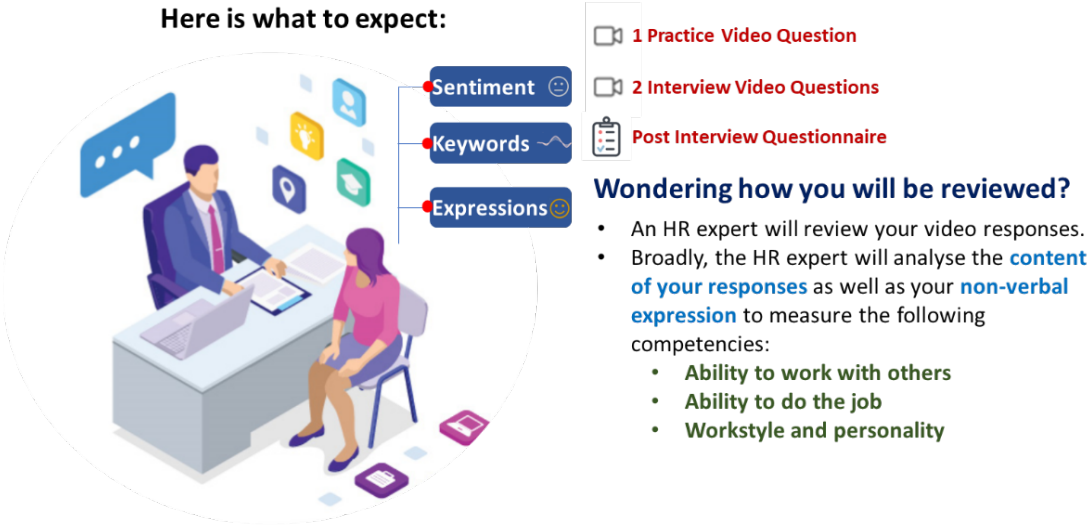


Table A13 provides descriptive statistics for the participants in this comparison. On average, participants were 37.12 years old, with 49% identifying as male. Approximately 20% identified as

[1] We thank the anonymous reviewer for this excellent insight.

Black/African American and 93% reported having at least a college degree. Table A14 summarizes the key outcomes from this comparison.

As confirmation of our manipulation, adding transparency in the human evaluator context (AVI-HTR) increased participants' perceived transparency compared to the baseline human evaluation (AVI-H; mean difference = 0.38, $p < 0.05$), albeit, to an expected weaker effect as AVI-AITR compares to AVI-AI. However, this increase in perceived transparency did not translate into significant changes in key behavioral or affective outcomes. Specifically, there were no significant differences between AVI-H and AVI-HTR for deceptive EIC (mean difference = 0.18, $p > 0.05$), honest SP (mean difference = 0.02, $p > 0.05$), or stress levels (mean difference = 0.07, $p > 0.05$).

Table A13. Descriptive Statistics

Variable	Mean	Std. Dev	Min	Max
Age	37.12	11.66	18.00	66.00
Gender[a]	0.49	0.50	0.00	1.00
Ethnicity[b]	0.20	0.40	0.00	1.00
Level of Education[c]	0.93	0.26	0.00	1.00
CRT Score[d]	0.93	0.86	0.00	3.00
Annual Income (Thousand USDs)	74.77	33.84	5.00	115.00
Notes: Number of observations = 259; [a]Coded as 1 for males and 0 for non-males; [b]Coded as 1 for black or African American and 0 otherwise; [c]Coded as 1 for at least college degree and 0 otherwise; [d]Calculated as sum of correct responses to three CRT questions.				

Table A14. Key Outcomes

Outcome	AVI-H (Mean)	AVI-HTR (Mean)	Absolute Difference	t-statistic (p-Value)
N	123	136	N/A	N/A
Perceived Transparency	-0.18	0.20	0.38	3.12 (<0.05)
Deceptive EIC	-0.09	0.09	0.18	1.45 (>0.05)
Honest SP	-0.01	0.01	0.02	0.12 (>0.05)
Stress	-0.04	0.03	0.07	0.62 (>0.05)
Note: Standardized means are reported for comparability.				

While results indicate a marginal increase in perceived understandability with the explicit addition of transparency elements in human-evaluated interviews (AVI-H), this shift in mean understandability is not substantially meaningful, as it does not affect interviewees' behaviors or stress. This finding aligns with our broader results, emphasizing that transparency plays a more critical role in addressing the opacity-induced disruptions unique to algorithmically evaluated interviews. In human evaluation contexts, where evaluators are inherently perceived as more understandable and relatable due to shared social norms and contextual familiarity, additional explanatory elements appear to be redundant.

2.1.2. Role of Monetary Incentives

In Experiment I, applicants received a fixed and performance-based incentive. The fixed payment translates into ~$8 per hour, surpassing Prolific's[2] minimum payment requirement during data collection (Newman et al. 2021; Palan and Schitter 2018). Further, the performance-based incentive was set at twice the value of fixed incentive to motivate sincere participation. Nevertheless, we acknowledge that

[2] Prolific imposes strict guidelines for the handling of submissions, and defines a minimum fixed payment based on wage standards. This is communicated to subjects when they sign up. Researchers can reject submissions and screen subjects based on past acceptance scores. Rejected submissions thus negatively affect subjects' eligibility in future studies, imposing an incentive for diligent participation. If subjects perceive unfair compensation, they can leave study without affecting their acceptance scores.

monetary incentives in a lab cannot be compared to benefits from a real job, which poses a reasonable challenge to our findings.[3] Potential insensitivity of our key findings to a substantially higher incentive can address such concerns. Our objective is to assess whether a relatively higher incentive scheme causes significant differences to the findings of Experiment I.

We collected additional data for AVI-AI and AVI-AITR groups from Experiment I after increasing the performance-based incentive from $4 to $12; fixed incentive is status quo. For the two high incentive (HI) groups, i.e., AVI-AI (HI) and AVI-AITR (HI), we collected 291 observations (

Table **A15**). We conducted a Kolmogorov–Smirnov test to confirm no significant difference in empirical distribution relative to data reported in Experiment I, following Fügener et al. (2021).

Table A15. Descriptive Statistics: Experiment I – Robustness Check (Higher Incentive)

Variable	Mean	Std. Dev	Min	Max
Age	31.71	10.56	18.00	66.00
Gender[a]	0.51	0.50	0.00	1.00
Ethnicity[b]	0.09	0.28	0.00	1.00
Level of Education[c]	0.64	0.48	0.00	1.00
CRT Score[d]	1.29	1.18	0.00	3.00
Annual Income (Thousand USDs)	59.88	34.89	5.00	115.00
Deceptive EIC	1.67	1.09	1.00	7.00
Honest SP	4.22	1.48	1.00	7.00
Stress[e]	0.14	1.26	-4.33	6.00

Notes: Number of observations = 291; [a]Coded as 1 for males and 0 for non-males; [b]Coded as 1 for black/African American and 0 otherwise; [c]Coded as 1 for at least college degree and 0 otherwise; [d]Calculated as sum of correct responses to three CRT questions; [e]Standardized for analysis.

Table A16. Robustness Check: Role of Monetary Incentives

a)

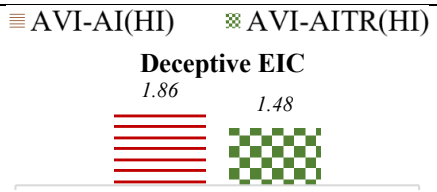


Table A16a: Deceptive EIC[a]

	AVI-AI (HI)	*AVI-AITR (HI)*
AVI-AI	0.2 (0.16)	0.57** (0.13)
AVI-AITR	-0.35** (0.13)	0.03 (0.09)

Overall F-statistic = *9.17 (p<0.05)*

b)

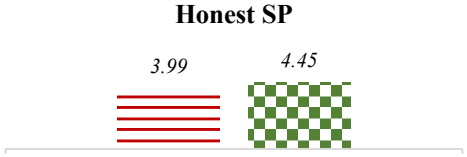


Table A16b: Honest SP[a]

	AVI-AI (HI)	*AVI-AITR (HI)*
AVI-AI	0.09 (0.17)	-0.60* (0.15)
AVI-AITR	0.52* (0.17)	0.05 (0.16)

Overall F-statistic = *4.99 (p<0.05)*

c)

Table A16c: Stress[a]

	AVI-AI (HI)	*AVI-AITR (HI)*
AVI-AI	0.04 (0.14)	0.36** (0.13)
AVI-AITR	-0.32* (0.15)	-0.01 (0.14)

Overall F-statistic = *3.87 (p<0.05)*

Notes: [a]Cells indicate row minus column standardized means compared using t-test; * p<.05, ** p<.01, standard errors in parentheses;

[3] We thank the anonymous reviewer for raising this concern.

In the left panels a), b), and c) of Table A16, we report the mean deceptive EIC (1.86 vs. 1.48), honest SP (3.99 vs. 4.45), and stress (0.29 vs. -0.02) for AVI-AI (HI) and AVI-AITR (HI) and observe a similar pattern as between AVI-AI and AVI-AITR. In the right panels of Table A16, pairwise comparison of these reactions relative to AVI-AI and AVI-AITR groups is reported. Between AVI-AI and AVI-AI (HI), and between AVI-AITR and AVI-AITR (HI), we observe no significant difference in deceptive EIC (Table A16a), honest SP (Table A16b), and stress (Table A16c). We do find significant differences in deceptive EIC, honest SP, and stress between AVI-AI (HI) and AVI-AITR (HI), per results in Section 5.4. Altogether, these results demonstrate the generalizability of our findings to conditions with higher incentives, a relatively higher stake situation.

2.1.3. Quasi Field Experiment

Research investigating job interviews and hiring contexts has frequently relied on simulated job interviews conducted in lab settings as their primary methodological approach (e.g., Bill et al. (2024); Gioaba and Krings (2017); Powell et al. (2021); Rizi and Roulin (2024); Roth et al. (2021); Schudlik et al. (2023)). Lab experiments are particularly valuable in this context because they allow for controlled manipulations of variables that would be difficult to manage in real-world hiring settings. Additionally, they often provide access to relatively large and generalizable samples, ensuring that results are both statistically robust and meaningful across varied populations. However, while lab experiments offer these strengths, they are often criticized for lacking ecological validity—that is, the extent to which findings are applicable to real-world scenarios.

Hence, building on recent research (e.g., Coffman and Klinowski (2025); Coffman et al. (2024)) that conducted examinations in recruitment contexts using quasi-field designs, we conducted a quasi-field experiment by recruiting individuals from the crowdsourcing labor platform Prolific. Furthermore, in Experiment I, we deliberately kept the job posting generic to ensure applicability to a broad participant group and avoid contamination of the manipulation by specific contextual expectations. However, in our quasi-field experiment, the job posting and description were highly specific, with clearly defined job expectations and skills. This allowed us to address the question of whether participants' impression management (IM) behaviors might change when they are presented with more explicit cues about the job requirements.

Another potential threat to the validity of our findings is measurement validity. While self-reported measurement has been widely validated in prior research, e.g., Bourdage et al. (2018); McCarthy and Goffin (2004); Roulin et al. (2014); Roulin et al. (2015), as a reliable indicator of such behaviors, we acknowledge concerns about accuracy in self-reported measurement. We address these concerns in our field experiment by incorporating additional measures to triangulate findings from the self-reports. Finally, we recruited independent evaluators and introduced evaluator coded complementary measures, including Criteria-Based Content Analysis (CBCA) and evaluators' perceived EIC to provide objective validations of our self-reported IM measurement. Altogether, our design replicates the findings of Experiment I within an ecologically relatable context to address concerns regarding generalizability, specificity (compared to generic) job posting, and measurement validity.

Experiment Design

Our experimental design integrates insights from recent studies examining job candidates in similar contexts, such as Coffman and Klinowski (2025) and Coffman et al. (2024), and organizational behavior research recruiting interviewees, such as Weiss and Feldman (2006). This allowed us to enhance ecological validity within the controlled rigor of experimentation. Participants were invited through Prolific, a crowdsourcing labor platform, to participate in a writing-related job opportunity (Job 1). They were tasked with producing a 200-word write-up on the topic of green technology and renewable energy

practices in the U.S., as outlined in Figure A2. Participants were compensated $3 for completing this task. They were explicitly informed that the use of generative AI tools was strictly prohibited during the writing exercise.

Figure A2. a) Information about Job 2 in Study Flyer; b) Job 1 Task; and c) Optional Interview Instructions After Job 1

a)

Optional Interview for Additional Paid Work

After completing the writing task (as described above), you will have the option to participate in a video interview for another job. You will receive an additional compensation of $4 for your time spent in participating in this optional interview

The results of the interview will be announced within a week of the completion of all interviews.

Based on performance in the interview, 100 successful interviewees will be invited to work on a different two-hour job and will be paid $50 for this two-hour job.

Interview Details:

- **Interview Duration:** 20 minutes (You will be compensated in case you time-out).
- **Interview Format:** Asynchronous video-based interview, where you will record your responses to a few questions using an online tool which will be provided.
- **Additional Compensation:** $4 for participating in the interview process.

b)

Sustainability Transition in the United States

As the United States steadily advances towards a sustainable economy, there is a heightened focus on green technology and renewable energy practices, such as increased reliance on electric vehicles and the use of solar and wind power. This transition aims to reduce carbon emissions and environmental impact, fostering a healthier planet for future generations.

In light of this shift, we are conducting market research to gather insights on public perceptions and attitudes toward the green technology and renewable energy sector.

We ask you to provide a 200-word (minimum) write-up expressing your views on the impact and future potential of green technology and renewable energy practices in America. This should include your opinions on electric vehicles, solar power, wind power, and other eco-friendly practices. Your insights could greatly help an upcoming business in tailoring its products or services to better meet the needs of environmentally conscious consumers.

c)

The writing job is now complete! Thank you for your efforts and for sharing your thoughts.

Would you be interested in interviewing for another lucrative job opportunity?

If you choose to participate in the optional interview process, you will:

- Earn an additional $4 for a 20-minute video interview.
- Have the chance to be selected for a different 2-hour job based on your interview performance. This job will pay $50 and we will be recruiting 100 individuals for this job.

Requirements for the Interview:

- Stable Internet Connection
- Laptop or computer (using a mobile phone is not allowed).
- Access to Google Chrome web browser.
- Webcam and microphone.
- The online software needed to record video interviews will be provided in the interview process.

If you are interested, please select yes to proceed to the next step to the interview process. If you are not interested in interviewing, please select no.

No | Yes

The purpose of Job 1 was two-fold. First, the writing sample represents an objective sample of participants' baseline creative abilities, which was later provided to the evaluators. Second, the task served as a natural precursor to the video interview process, mirroring real-world recruitment scenarios in which candidates complete an initial screening task before advancing to the next stage.

The posting for Job 1 explicitly stated that participants who completed the task would have the voluntary option to apply for a second job (Job 2), which offered a significantly higher payment. The application for Job 2 involved a video interview. Participants were explicitly informed that based on their performance in the video interview, 100 individuals would be recruited for Job 2. This high recruitment number ensured that participants perceived a realistic chance of being selected. After completing Job 1,

participants who chose to apply for Job 2 were informed that the position required them to prepare a creative flyer for a new product launch, a task aligned with writing and creative problem-solving skills. This second job was considerably more lucrative than Job 1, offering a $50 payment for a two-hour task, equivalent to $25/hour, thereby addressing potential concerns of interview incentive and realism. This voluntary setup served to naturally filter participants, ensuring that only those genuinely interested and motivated progressed to the video interview stage. The combination of a high payment rate and the opportunity to gain practical experience in a writing-intensive role provides realistic incentives for participants to apply for Job 2.

Participants who signed up for Job 2 were immediately redirected to complete a video interview. They were randomly assigned to one of four experimental conditions: AVI-H (Human evaluator only), AVI-HTR (Human evaluator with transparency – similar to Appendix: Section 2.1.1), AVI-AI (Non-transparent AI), and AVI-AITR (Transparent AI). Before the interview, participants were presented with the job description for Job 2, as shown in Table A17. This description was kept highly specific, intentionally similar to Job 1 in being a creative writing job, but with additional details that clearly outlined the position's responsibilities and required skills.

Table A17. Job Description for Job 2

Position: *Creative Associate.*

We are seeking a Creative Associate to join our team and play a role in shaping the narrative of our innovative product.

Selected associates will be instrumental in developing and refining the marketing and descriptive content for a technology-driven consumer product. Your responsibilities will encompass drafting creative descriptions that capture the essence and advantages of the product, as well as providing insightful feedback to enhance its design and functionality. Understanding requirements about the product, you will help bring the product to its fullest potential, ready for market launch.

Role and Responsibilities:

- Draft descriptions that highlight the unique features of our innovative product for a broad audience.
- Gather requirements and incorporate feedback to ensure that all materials align with the overall design and marketing strategy.
- Ensuring all information is presented in a clear and professional manner.
- Assist in creating the promotional content through a flyer to meet specific marketing objectives and needs.
- Uphold high standards of quality and consistency in all written content, ensuring it meets the requirements and brand voice.
- Play a key role in the product's success by working on this flyer.

Key Skills and Qualities:

- Ability to create content.
- Written communication skills.
- Ability to work flexibly and with a positive attitude.
- Attention to detail.
- Proficiency in word processing tools such as Microsoft Word.

During the interview, participants were asked two structured questions: one experiential and one situational, following Ellis et al. (2002). The questions, as presented in Table A18, were thoughtfully adapted to suit the specificity of Job 2, ensuring they were relevant to the position while avoiding overly generic phrasing. Following the interview, participants were asked to complete a short questionnaire. Subsequently, we invited 100 individuals to complete Job 2, which involved designing a creative flyer for a new product launch (a solar powered portable charger). Invitations were extended based on interview

performance, as determined by the ratings of evaluators. While the inclusion of Job 2 enhances the realism of the study, it is not directly relevant to the scope or findings of this study.

Table A18. Interview Questions

Question	Question Type
Describe a time when you had to think outside the box to solve a problem. What was the challenge, and what was your solution?	Experiential
Suppose you made an error in the content you created that went live. How would you handle the situation?	Situational

Measures

We focused on fewer key constructs compared to Experiment I to preserve realism of the quasi-field experiment, prioritizing measures that would validate and replicate our findings. Specifically, we measured self-reported deceptive EIC post-interview to replicate the self-reported IM behaviors observed in Experiment I. Additionally, we measured stress pre- and post-interview to replicate and confirm the validity of our affective measure from Experiment I. Post-interview, we also assessed perceived transparency as a manipulation check. All items, measured on a 7-point scale, demonstrated high reliability (Details in

Table A5).

Analysis and Results

Data Description and Randomization Check. Table A19 presents the descriptive statistics. The sample consisted of 398 individuals with an average age of ~40 years. Approximately 54% of the participants identified as female, and 16% identified as Black/African American. Most participants (94%) reported having at least a college degree, with an average annual income of $70,550. Table A20 confirms the success of the randomization check, showing no significant differences in key demographic or baseline variables across experimental groups. This ensures that any observed effects can be attributed to experimental manipulations rather than pre-existing differences among participants.

Table A19. Descriptive Statistics

Variable	Mean	Std. Dev	Min	Max
Age	39.96	11.86	18.00	66.00
Gender[a]	0.46	0.50	0.00	1.00
Ethnicity[b]	0.16	0.37	0.00	1.00
Level of Education[c]	0.94	0.24	0.00	1.00
Annual Income (Thousand USDs)	70.55	34.25	5.00	115.00
Perceived Transparency	5.37	1.23	1.00	7.00
Deceptive EIC	1.51	0.93	1.00	6.00
Stress[d]	0.10	1.25	-5.33	6.00

Notes: Number of observations = 398; [a]1 for males and 0 for non-males; [b]1 for Black/African American and 0 otherwise; [c]1 for at least college degree and 0 otherwise; [d]Stress after minus before the interview.

Key Results. Panel a) of Table A21 confirms the success of our transparency manipulation. For both AI and human evaluator groups, the manipulation of transparency significantly increased

interviewees' perceived transparency ($p < 0.05$), reflecting an improvement in their understandability of the interview process. Specifically, participants in the AVI-HTR group reported higher transparency compared to those in AVI-H (mean difference = 0.36, $p < 0.05$), and participants in the AVI-AITR group reported higher transparency than those in AVI-AI (mean difference = 1.12, $p < 0.01$).

Table A20. Randomization Check

Variable	AVI-H	AVI-HTR	AVI-AI	AVI-AITR	F-Test[d] (p-value)
N	94	100	103	101	-
Age	38.87	40.73	39.81	40.36	0.45 (>0.05)
Gender[a]	0.50	0.43	0.49	0.44	0.49 (>0.05)
Ethnicity[b]	0.17	0.16	0.16	0.15	0.06 (>0.05)
Level of Education[c]	0.94	0.91	0.94	0.96	0.74 (>0.05)
Annual Income (Thousand USDs)	72.34	72.30	68.59	69.16	0.34 (>0.05)

Notes: Number of observations = 398; [a]1 for males and 0 for non-males; [b]1 for black or African American and 0 otherwise; [c]1 for at least college degree and 0 otherwise; [d]Omnibus statistic – confirmed using post-hoc tests with Bonferroni correction.

However, transparency had a differential effect on deceptive EIC for both AI and human-evaluated groups. Participants in the AVI-AI group reported the highest levels of deceptive EIC (mean = 0.30), significantly higher than those in AVI-H (mean = -0.15, mean difference = 0.42, $p < 0.01$). Introducing transparency in the AI context (AVI-AITR) reversed this effect, reducing deceptive EIC to levels significantly lower than AVI-AI (mean = -0.06, mean difference = -0.33, $p < 0.05$). Notably, transparency had no significant effect on deceptive EIC in the human evaluator context; deceptive EIC in AVI-H and AVI-HTR were statistically indistinguishable (mean = -0.15 vs. mean = -0.10, mean difference = 0.05, $p > 0.10$).

Table A21. Key Results from Field Experiment

Table A21a: Perceived Transparency[a]

		Mean	(1)	(2)	(3)	(4)
a)	*AVI-H (1)*	0.04	-			
	AVI-HTR (2)	0.33	0.36 (0.15)*	-		
	AVI-AI (3)	-0.63	-0.83 (0.17)**	-1.19 (0.17)**	-	
	AVI-AITR (4)	0.28	0.30 (0.15)	-0.06 (0.16)	1.12 (0.17)**	-

Table A21b: Deceptive EIC[a]

		Mean	(1)	(2)	(3)	(4)
b)	*AVI-H (1)*	-0.15	-			
	AVI-HTR (2)	-0.10	0.05 (0.11)	-		
	AVI-AI (3)	0.30	0.42 (0.14)**	0.37 (0.15)*	-	
	AVI-AITR (4)	-0.06	0.08 (0.11)	0.04 (0.12)	-0.33 (0.15)*	-

Table A21c: Stress[b]

		Mean	(1)	(2)	(3)	(4)
c)	*AVI-H (1)*	-0.22	-			
	AVI-HTR (2)	-0.20	0.02 (0.17)	-		
	AVI-AI (3)	0.48	0.87 (0.18)**	0.84 (0.16)**	-	
	AVI-AITR (4)	-0.09	0.16 (0.17)	0.13 (0.17)	-0.71 (0.17)**	-

Notes: Standardized means are reported for comparability; Cells indicate row minus column means compared using t-test; Means are standardized across the full sample; pairwise differences are estimated from t-tests on raw (unstandardized) scores; + p<.10, * p<.05, ** p<.01, standard errors in parentheses; [a]Measured on a scale of 7; [b]Difference of post- and pre-interview stress.

Transparency also showed contrasting effects on stress for AI and human-evaluated groups. Participants in the AVI-AI group reported the highest stress (mean = 0.48), significantly higher than those in AVI-H (mean = -0.22, mean difference = 0.87, $p < 0.01$). Transparency in AI-evaluated interviews significantly reduced stress (mean = -0.09, mean difference = -0.71, $p < 0.01$ compared to AVI-AI). However, transparency in the human-evaluated interview context did not significantly affect stress levels compared to AVI-H (mean = -0.20, mean difference = 0.02, $p > 0.10$).

Altogether, these results allow us to replicate our self-reported measures from Experiment I in a more ecologically relatable context, where the interviews carried actual job-related consequences, providing robust support for the generalizability of our findings. Importantly, the lack of significant effects of transparency in human-evaluated contexts underscores its unique relevance in AI-enabled evaluations, where it addresses opacity-induced disruptions. However, for IM measures, which often suffer from biases such as social desirability, we conducted additional validation, as discussed next (and additional tests in Appendix: Section 2.3.1), to further ensure the reliability and robustness of our results.

Validating Self-Reported Measure of IM

We generate additional measures using external coders using data collected in our quasi-field experiment to bolster the robustness of our findings and establish the validity of our measures of impression management in the quasi-field experiment. These external evaluators coded the interview responses collected in our quasi-field experiment. These evaluators were also recruited through Prolific, ensuring a diverse and professionally qualified pool. Individuals who had participated in the field experiment were restricted from participating as evaluators by design. Each coder was randomly assigned one interview candidate to evaluate. All coders were employed full-time and brought relevant professional experience, aligning their expertise with the context of the study. The average duration of the coding task was 12 minutes, and evaluators received $3 for their time.

Table A22 provides descriptive statistics for the external evaluator sample, which consisted of 398 evaluators. The average age of coders was 38.66 years, with approximately 47% identifying as female. Additionally, 22% of the sample identified as Black or African American. Coders demonstrated a high level of professional experience, with 71% having participated in interviews as evaluators in the past. On average, coders reported 15.96 years of professional experience as interviewers and had interviewed 28.77 applicants in their roles. These metrics highlight the qualifications of the coders and the relevance of their expertise to the evaluation tasks. Using this sample of external evaluators, we code the interviews collected in our quasi-field experiment for three additional measures: two for validating IM, and one for interview performance.

Approach 1 for Validating IM: Evaluators' Perceived EIC. While prior research has shown that human evaluators or HR experts are often inaccurate in predicting IM behaviors such as deceptive EIC (Roulin et al. 2014), we anticipate that evaluators could more successfully assess deceptive behaviors if they had access to a reference point reflecting the interviewee's baseline abilities. This approach draws upon prior research, such as Weiss and Feldman (2006), which utilized two independent assessments to identify discrepancies and evaluate deceptive behaviors in various contexts.

Table A22. Descriptive Statistics of the Evaluators

Variable	Mean	Std. Dev	Min	Max
Age	38.66	11.65	18.00	66.00
Gender[a]	0.53	0.50	0.00	1.00
Ethnicity[b]	0.22	0.42	0.00	1.00
Interviewer[c]	0.71	0.46	0.00	1.00
Experience[d]	15.96	10.94	0.00	28.00
Interviews[e]	28.77	28.53	10.00	110.00

Notes: Number of observations = 398; [a]1 for males and 0 for non-males; [b]1 for black or African American and 0 otherwise; [c]1 if interviewed at least once in past; [d]Number of interviewing years; [e]Number of applicants interviewed in past.

In our setup, interviewees' writing samples from Job 1 served as this baseline reference. Evaluators first reviewed the writing sample independently, forming an objective understanding of the interviewee's abilities. They were explicitly informed that the writing sample was not part of the interview process, ensuring that their assessments of the sample remained unbiased. Following this, evaluators viewed the corresponding interview videos and were asked to compare the interview performance to the baseline writing sample. Using this reference, evaluators responded to measures of perceived EIC (Adapted from Table A5), identifying any potential embellishments or inconsistencies in the interviewee's responses relative to their demonstrated abilities in the writing sample.

Table A23 summarizes the evaluators' assessments of perceived EIC across the experimental groups. Evaluators' perceived interviewees in the AVI-AI condition to exhibit the highest EIC (mean = 3.55), significantly higher than those in the AVI-H condition (mean difference = 0.63, $p < 0.01$) and the AVI-HTR condition (mean difference = 0.64, $p < 0.01$). Introducing transparency in the AI context, as seen in the AVI-AITR condition, reduced perceived EIC (mean = 3.05) compared to the non-transparent AI condition (AVI-AI), with a significant difference of -0.50 ($p < 0.05$). In the human evaluator context, however, the results show no significant differences in perceived EIC between AVI-H (mean = 2.92) and AVI-HTR (mean = 2.91; mean difference = -0.01, $p > 0.10$).

These findings align closely with our results from our quasi-field experiment and Experiment I, reinforcing the observation that interviews conducted under the AVI-AI condition led to higher perceptions of deceptive EIC, particularly when evaluators have access to a baseline reference of interviewees' abilities. The results confirm that transparency in AI-enabled algorithmic evaluation plays an important role in reducing deceptive IM, effectively mitigating the disruptions caused by the opacity of AI-enabled algorithmic evaluations. Conversely, transparency appears to be redundant in human-evaluated contexts, where perceived EIC remains stable, potentially due to the hypothesized inherent understandability and familiarity with human evaluation.

Table A23. Evaluators' Perceived Extensive Image Creation for Interviews Across Experiment Groups

	Mean	*(1)*	*(2)*	*(3)*	*(4)*
AVI-H (1)	2.92	-			
AVI-HTR (2)	2.91	-0.01 (0.21)	-		
AVI-AI (3)	3.55	0.63 (0.21)**	0.64 (0.20)**	-	
AVI-AITR (4)	3.05	0.13 (0.21)	0.14 (0.20)	-0.50 (0.20)*	-

Notes: Cells indicate row minus column means compared using t-test; + p<.10, * p<.05, ** p<.01, standard errors in parentheses.

Approach 2 for Validating IM: Criteria-Based Content Analysis (CBCA). CBCA has long been established as a method for assessing the credibility of verbal statements in prior criminology and psychology research (Vrij 2005; Vrij and Mann 2006). As detailed in Roulin and Powell (2018), CBCA is particularly suited for contexts where verbal accounts can be systematically evaluated for their content-related indicators presented in Table A24. It involves the systematic scoring of each verbal account based on predefined indicators. Higher CBCA scores reflect greater content credibility, while lower scores often signal deceptive tendencies, as individuals attempting to fabricate accounts typically struggle to provide coherent, detailed, and contextually embedded narratives.

Importantly, prior research emphasizes that CBCA is valid only when certain conditions are met (Roulin and Powell 2018). First, the verbal statements must stem from a setting where participants are

motivated to provide truthful accounts or simulate real-world scenarios. Second, the elicited statements should allow participants to narrate their experiences in sufficient detail, rather than providing brief, shallow responses. Finally, the evaluation context should mimic real-world conditions to ensure participants engage authentically with the task.

Our field experiment meets these conditions for CBCA's validity. First, participants were informed that their interview responses would determine their eligibility for a subsequent lucrative job, simulating a real-world hiring scenario. This setup ensured that participants were motivated to provide detailed and credible responses, aligning with the first condition. Second, we included interview questions that were designed to elicit experiential narratives. Finally, the field experiment was embedded within a realistic hiring context, enhancing ecological validity.

Table A24. CBCA Indicators (Based on Roulin and Powell (2018))

CBCA Indicator	**Description**
Logical structure	Statement is coherent and consistent, with different details that all describe the same course of events
Unstructured production	Information is not provided in chronological order, showing an expressive and unstructured presentation
Quantity of details	Statement includes considerable amounts of specific details
Contextual embedding	References to time and space that provide a basis for the story
Descriptions of interactions	Descriptions of any type of interaction
Reproduction of conversations	Replication of actual wording in conversations
Unexpected complications	Description of unforeseen difficulties
Unusual details	Details that are uncommon but meaningful to the story
Superfluous details	Non-essential details included in the main story
Subjective mental state	Reports of feelings or thoughts at the time of the event
Spontaneous corrections	Self-corrections made during the narrative without prompting
Lack of memory	Indications that some parts of the statement may be unclear or uncertain
Doubts about testimony	Anticipating objections or skepticism about the veracity of the account
Self-deprecation	Mention of personally unfavorable or self-incriminating details

We operationalized CBCA based on coding by the same independent evaluators. Prior to responding to measures of perceived EIC, each evaluator was provided with a detailed CBCA questionnaire to code the interview responses across 14 CBCA indicators (Table A24). This questionnaire closely followed the approach described in Roulin and Powell (2018), using a three-point scale for each indicator. Absent (coded as 0), if the indicator was not present in the interview response; partially present (coded as 1), if the indicator was present but not sufficiently demonstrated; and fully present (coded as 2), if the indicator was fully and clearly exhibited in the response. Before coding, evaluators received instructions on each CBCA indicator, including its definition and examples of absent, partial, and full presence, to ensure consistency across evaluations.

Prior research has demonstrated a significant negative association between CBCA total scores and deceptive EIC, where higher CBCA scores are associated with truthful accounts and lower scores signal potential deception (Roulin and Powell 2018; Vrij 2005). This relationship is consistent with the notion that deception often results in less coherent and detailed narratives, reducing CBCA scores. We document a similar negative association in our data as the correlation between CBCA total scores and interviewees' self-reported EIC was negative and significant ($r = -0.13$, $p < 0.01$). This observed negative association between CBCA total scores and self-reported deceptive EIC supports the validity of our self-reported measure. By aligning our findings with established patterns in the literature, we provide robust evidence that our measurement of deceptive EIC in Experiment I and the quasi-field experiment effectively capture variations in participants' IM behaviors.

Measuring Interview Performance. We adopted two approaches to assess how transparency in AVIs affected eventual interview performance. The first approach evaluates eventual interview performance, and the second assesses changes in perceived ability based on prior baseline performance of Job 1.

As shown in

Table **A25**a, applicants evaluated under AVI-AI performed significantly worse compared to both human evaluation (AVI-H; $p < 0.01$, $d = -0.41$) and human evaluation with transparency (AVI-HTR; $p < 0.05$, $d = -0.31$). These results align with our findings from Experiment II, under human only evaluation. In contrast, interview performance under AVI-AITR was significantly improved compared to AVI-AI ($p < 0.10$, $d = 0.25$) and restored to levels statistically comparable to AVI-H and AVI-HTR ($p > 0.05$).

Table A25. Interview Performance

Table A25a: Interview Performance[a]

		Mean	*(1)*	*(2)*	*(3)*	*(4)*
a)	*AVI-H (1)*	0.17	-	-	-	-
	AVI-HTR (2)	0.07	-0.10 (0.15)	-	-	-
	AVI-AI (3)	-0.23	-0.41 (0.14)**	-0.31 (0.14)**	-	-
	AVI-AITR (4)	0.01	-0.16 (0.14)	-0.05 (0.15)	0.25 (0.13)+	-

Table A25b: Change in Perceived Ability[b]

		Mean	*(1)*	*(2)*	*(3)*	*(4)*
b)	*AVI-H (1)*	0.17	-	-	-	-
	AVI-HTR (2)	0.08	-0.08 (0.14)	-	-	-
	AVI-AI (3)	-0.34	-0.51 (0.14)**	-0.42 (0.15)*	-	-
	AVI-AITR (4)	0.10	-0.06 (0.13)	0.01 (0.14)	0.44 (0.14)**	-

Notes: Standardized means for comparability; Cells indicate row minus column means compared using t-test; + p<.10, * p<.05, ** p<.01, standard errors in parentheses; [a]Measured on a scale of 7; [b]Difference of perceived ability in Job 2 compared to Job 1.

The second approach examines changes in perceived ability—the difference between evaluators' assessments of interviewees' writing sample (Job 1) and their interview performance for Job 2.

Table **A25**b highlights similar patterns: evaluators perceived a significant reduction in ability for applicants interviewed under AVI-AI relative to both AVI-H ($p < 0.01$, $d = -0.51$) and AVI-HTR ($p < 0.05$, $d = -0.42$). These perceived declines further reinforce that algorithmic evaluation without transparency not only disrupts applicants' behaviors but also creates lasting negative impressions that diminish their assessed ability. However, transparency in AVI-AITR significantly mitigated this decline, as changes in perceived ability were notably improved compared to AVI-AI ($p < 0.01$, $d = 0.44$) and aligned closely with AVI-H and AVI-HTR ($p > 0.05$).

Altogether, these results underscore the downstream consequences of algorithmic opacity on interviewees' performance. While AVI-AI significantly hindered interview performance and reduced perceived ability, transparency in algorithmic evaluation effectively alleviated these disruptions, consistent with Experiment II.

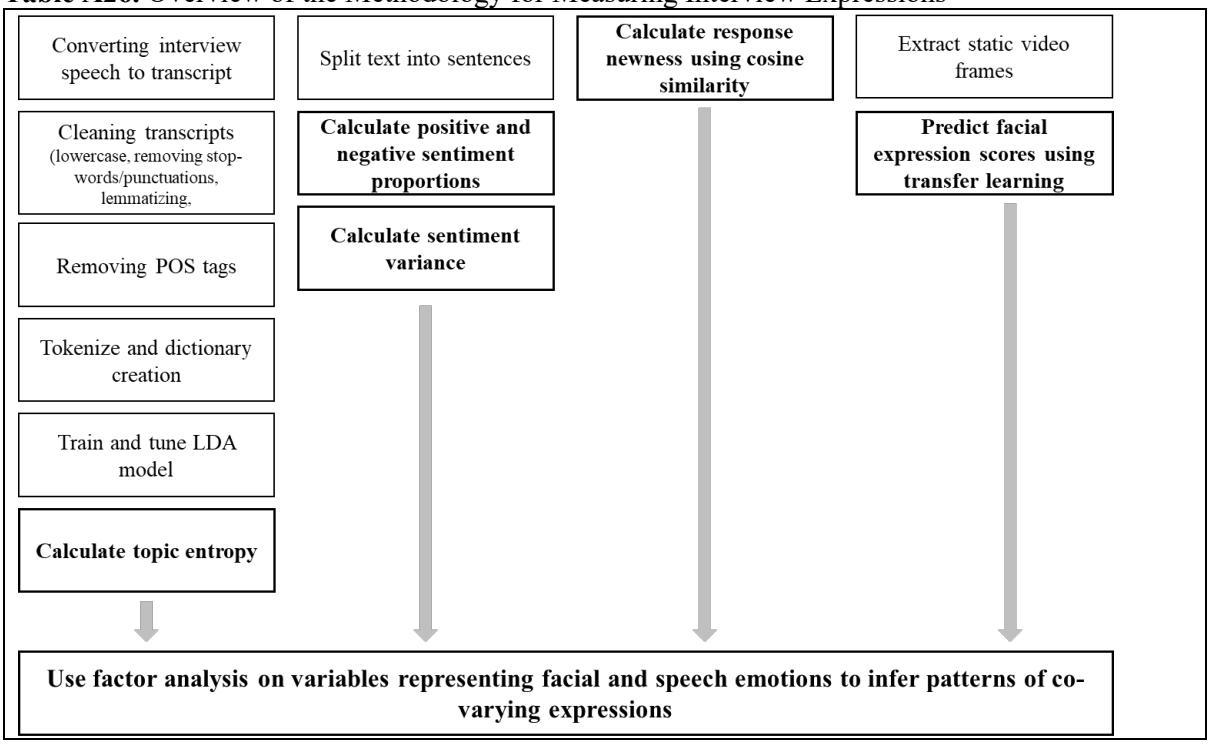
2.2. Measurement Concerns

We adopted two complementary approaches to validate the self-reported measures collected in Experiment I. First, we leverage machine learning to analyze observable indicators of interviewees' facial and verbal reactions, enabling us to detect variations in their reactionary patterns in response to our experimental manipulations. Second, we drew on the CBCA indicators coded in our quasi-field experiment (Section 2.1.3) and discuss how they provide an objective benchmark to corroborate the reliability of the self-reported measures of IM.

2.2.1. Using Machine Learning to Observe Verbal and Facial Reactions

Our findings in Experiment I demonstrate significant variations in behavioral and affective responses depending on human, or algorithmic evaluation–with and without transparency. These findings prompt an important question: could these variations be observable in interview responses? Given expressions in such contexts are often subtle, complex, and challenging to code manually (Roulin et al. 2014), we used machine-learning to analyze abstract emotional and communication patterns in the interview videos (Choudhury et al. 2019; Guo et al. 2021). We first outline our approach and then proceed with our analysis. Table A26 provides an overview of the machine learning pipeline. The process of coding verbal and facial expressions is discussed next.

Table A26. Overview of the Methodology for Measuring Interview Expressions[4]



Coding and Analyzing Interview Expressions

Role of verbal (e.g., content of responses) as well as non-verbal (e.g., facial expressions or gestures) expressions in job interviews have been widely recognized (Bourdage et al. 2018; Helfat and Peteraf 2015). Recent research has thus focused on objective machine-learning approaches to measure and categorize communication emerging in such settings (Choudhury et al. 2019; Guo et al. 2021; Jiang et al.

[4] Based on Choudhury et al. (2019).

2019). Early work in this stream focused on text mining, including dictionary techniques, e.g., linguistic inquiry and word count (LIWC) and Latent Dirichlet Allocation (LDA). Recently, videographic techniques involving supervised deep learning to measure emotions have become popular (Choudhury et al. 2019). We build upon such literature to code the verbal and non-verbal expressions of applicants interviewed in experiment I.

Coding Speech Content. All interview videos were first converted to audios using the *moviepy* python library and then transcribed using the *speech_recognition* python library (Kamath et al. 2019). Our *first* measure is the combined number of words in the two interview videos of the applicant, i.e., *response length*. The transcripts were cleaned by converting to lowercase, removing stop-words, i.e., set of commonly used English language words e.g., 'the' or 'is', punctuations, and were then lemmatized, i.e., grouping of inflected English words, e.g., 'run' and 'ran'. Parts of speech words, e.g., pronouns and prepositions, were removed (Srinivasa-Desikan 2018).

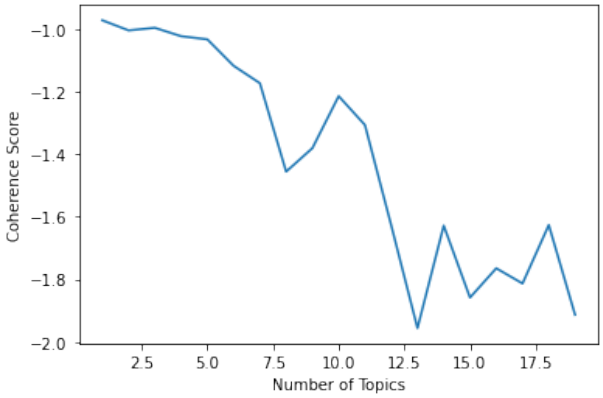
We tokenized the cleaned transcripts, created a dictionary, and used bag of words approach to represent the transcripts in feature space (Choudhury et al. 2019). Transcripts of both interview responses were combined to create a *document* per applicant. We trained a Latent Dirichlet Allocation (LDA) model for unsupervised topic modeling (Choudhury et al. 2019). LDA does not consider word order by treating a random generative process in creation of a set of documents, i.e., the *corpus*. It runs backwards and iterates to return the most probable set of topics underlying the corpus. Once the number of specified topics is identified, the model can predict. We estimate the LDA model using the *gensim* python library; we use coherence scores for hyperparameter tuning (Stevens et al. 2012).

The best model resulted in five topics. Document level covariates, indicating the proportion of words for each topic in a document were then predicted. The proportions are used to calculate our *second* speech measure, i.e., *topic entropy,* the breadth of information in each applicants' responses. Higher values indicate a wider range of ideas or opinions in the response. We calculated the Shannon entropy using the following shown in equation (1) for each applicant; p_i represents the proportion allocated to each topic as follows:

$$\text{Topic entropy} = -\sum(p_i \times \log_2 p_i) \qquad (1)$$

We used the *Vader* python library to measure sentiment for each token in the document, and then the proportion of each document dedicated to positive and negative sentiment, leading to a linear measure of sentiment (Choudhury et al. 2019). We use *positive document sentiment* as our *third* speech content measure. We also calculate the standard deviation of positive sentiment to calculate the *document sentiment variance* as our *fourth* speech content measure. We used cosine similarity to code the semantic similarity in applicants' response to the two questions asked in the AVI following the approach set by Guo et al. (2021). This led to our *fifth* speech content measure of *response newness*, indicating the dissimilarity in the meanings of the two interview responses.

Figure A3. Tuning Model Hyperparameters in LDA.



Note: Number of topics was chosen as 5 as the u-mass coherence score started to consistently reduce after five topics. The score calculates how often two words, w_{i}cn and w_{j} appear together in the corpus. Alpha and beta Hyperparameters, alpha representing document-topic density and beta representing topic-word density, were tuned.

Figure A4. Inter-Topic Distance in 2-Dimensional Space for LDA

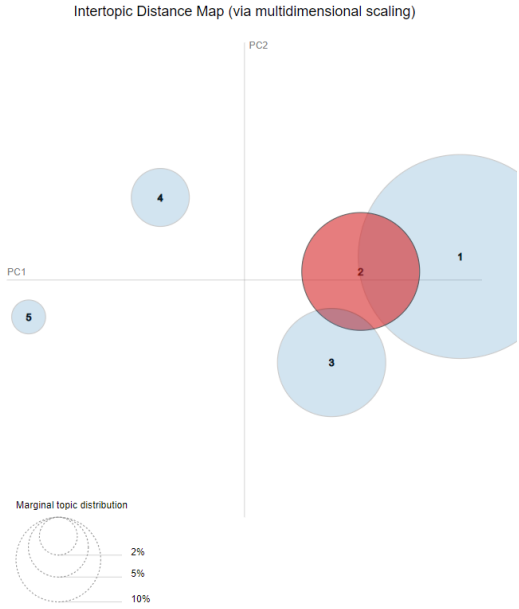


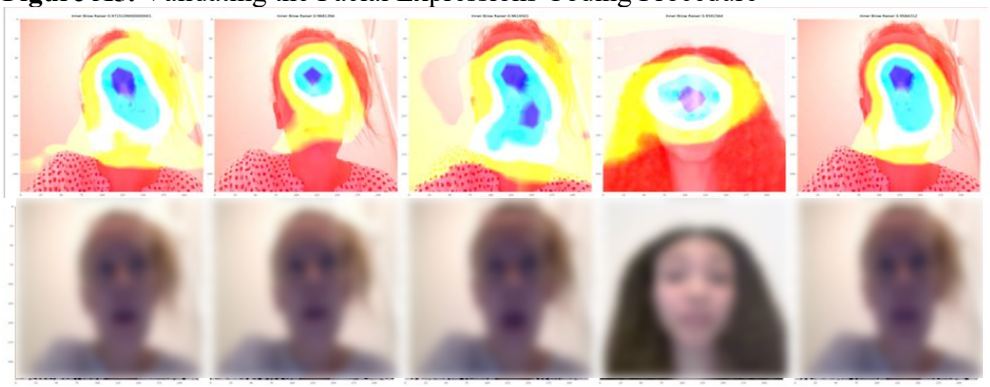
Table A27. Exemplary Tokens among the Five Topics created using LDA.

Topic 1	**Topic 2**	**Topic 3**	**Topic 4**	**Topic 5**
Work	Fix	Know	Process	Complex
People	Work	People	Step	Explain
Know	Process	Explain	Help	Company
Worker	Explain	Notice	Need	Group
Explain	Notice	Work	Notice	Think

Coding Facial Expressions. We used supervised deep learning to generate weights along the seven basic facial expression which are widely represent human emotions across diverse cultures (Ekman 1997). These are anger, contempt, disgust, fear, happiness, sadness, and surprise, and are based on combination of various facial muscles, such as upper eyebrows, dimples on cheek, or openness of mouth

(Lucey et al. 2010). We utilized transfer learning, a deep learning technique that re-uses model parameters from one problem to a different but related problem (Torrey and Shavlik 2010). The model was based on the VGG-19 architecture, trained on 14 million human images from the image-net database, to make predictions involving human faces (Shaha and Pawar 2018). The weights in first three blocks were fixed for feature extraction and last two were trained using the CK+ dataset (Lucey et al. 2010). The final dense layer was replaced with two custom layers for batch normalization and dropout. The model predicted probability of activation of the seven facial emotions and the 34 facial muscles that are responsible for these emotions for each frame in an interview video; average probability across all frames for the two videos was calculated. Figure A5, visualizes model's prediction for inner eyebrow muscle, validating our approach.

Figure A5. Validating the Facial Expressions Coding Procedure



Note: *Faces in the bottom frame have been intentionally blurred to maintain subjects' privacy. The top frame illustrates the facial regions identified by the model to be associated with the activation of the inner eyebrow raising movement.*

In the following section, we detail the steps that were taken to develop the transfer learning model and then predict the probability of activation of facial emotions (Ekman et al. 1988; Ekman and O'Sullivan 2006).

Description of the Transfer Learning Approach

Training Data. Training a supervised deep learning model requires extensive volumes of labelled data which is used to train the parameters (*bias* and *weights*) of the model (*neurons*) in multiple rounds of training (*epochs*) to minimize the overall cost (*loss*) of the model (LeCun et al. 2015). We utilize the extensive dataset developed by Lucey et al. (2010) (CK+ dataset), which comprises sequence of human expressions, where each sequence is coded for the presence or absence of seven basic facial emotions as well as the facial action units behind them. The overall data set contains 593 such image sequences, with each sequence coded for the presence or absence of each unique facial action units. The expression in last image in each sequence represents the emotion and the associated action units.

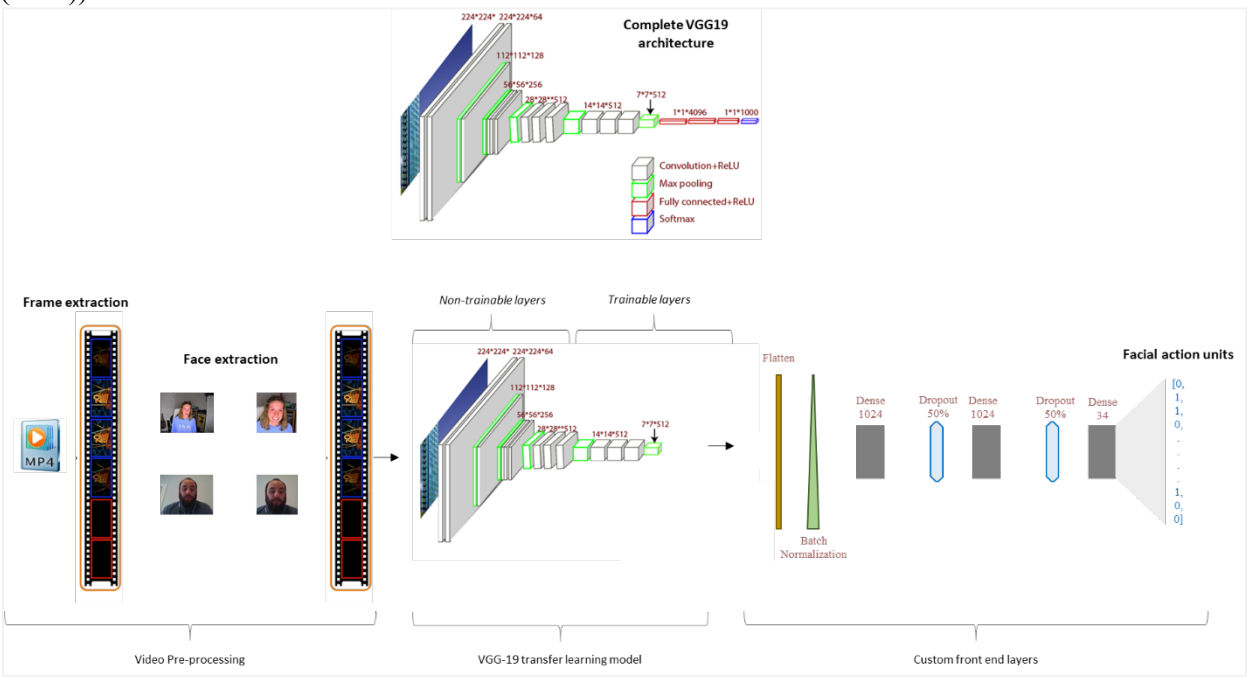
Model. Despite the richness of the coded data set, it could potentially raise concerns about limited size by only containing 593 images (Feng et al. 2019). Typically, deep learning models constitute several layers, with each layer containing hundreds or thousands of neurons (for identifying and transforming features) and require thousands of observations to efficiently learn underlying patterns from evidence (Brynjolfsson and McAfee 2017). Training on smaller data sets could also result in overfitting of the model (Lever et al. 2016). We overcome these difficulties systematically.

First, we relied upon *transfer learning*, which is a deep learning technique where the knowledge of pre-trained model architectures can be used to detect features from evidence based data (such as

images) and allows researchers to minimize the number of new-parameters which require training (Liu et al. 2020; Ng et al. 2015; Shaha and Pawar 2018). Basically, the first few layers of model architecture (which have been pre-trained on large corpuses of images to identify image features) are not trained again to extract fundamental image features. The rest of the layers were trained to transform these fundamental features using activation functions (such as sigmoid or rectified linear units) to generate predictions and reduce training loss. We used the *VGG-19* model architecture which has been trained on 14 million images in the image-net database to predict 1000 categories of images (Shaha and Pawar 2018); the model has been proven highly efficient in predictions involving human faces. The VGG-19 architecture has five blocks of convolutional layers and a final block of dense layers which can classify an image of size 224x224 pixels. In our custom architecture, we fixed the first three blocks with pre-trained weights and train the last two blocks of the architecture to reduce the number of trainable parameters.

Second, we removed the final dense block from the architecture and replace it with a custom block of two dense layers with batch normalization and dropout layers. The batch normalization normalizes the output feature matrix from the convolutional blocks of the VGG-19 architecture and the dropout layers introduce probabilistic activation to the dense layers, with only 50% of the neurons activating in every training cycle. Cumulatively, batch normalization and dropout layers allowed us to introduce extensive regularization and prevent overfitting (Ioffe and Szegedy 2015; Santurkar et al. 2018). Image augmentation was further used to avoid overfitting (Frid-Adar et al. 2018). A final dense layer with sigmoid activation function (Wang et al. 2018) is used and data set is pre-balanced to ensure symmetry of classes and avoid training biases (Lemaitre et al. 2017). The model architecture is shown in Figure A6.

Figure A6. Custom transfer learning model using VGG-19 model architecture (Based on Wen et al. (2019)).

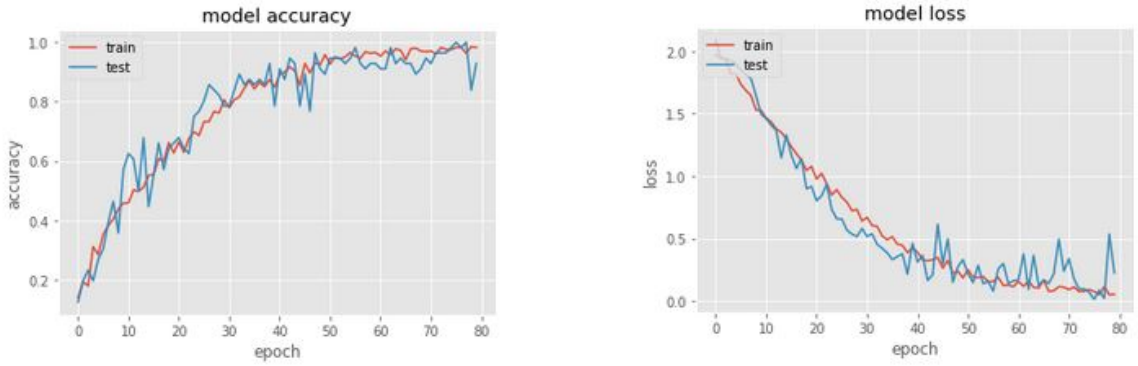


Data pre-processing. We first extracted frames from the interview videos of the candidates at the rate of 2 frames per second. We further use *haar face cascade* to trim faces from each frame of our training as well as test data (Cho et al. 2009). The final image sequences contain frames comprised of participants' faces.

Training. We trained our custom transfer learning architecture on the CK+ dataset. The model was trained for 60 epochs (loss starts to increase after 60 epochs) and realized a training accuracy of 94.0% and a validation accuracy of 93.68% on the held-out set; training the model took more than 14 hours. The training and testing accuracy, and categorical cross entropy loss with epoch progression are shown in Figure A7 (Zhang and Sabuncu 2018).

Predictions. Next, we predicted the probability of the facial emotions and the action units for each frame of the candidate's video. The probabilities are finally averaged and aggregated at the participant level.

Figure A7. Model training accuracy and loss



Analyzing Verbal and Facial Expressions

Measuring Verbal and Facial Expressions. We extracted five variables from interviewees' speech and seven from their facial expressions in the AVI videos (Table A28). Speech variables, derived through text mining of interview transcript included response length, topic entropy, positive sentiment, sentiment variance, and response novelty. Facial expression variables were generated using supervised deep learning, capturing weights for seven universally recognized emotions: anger, contempt, disgust, fear, happiness, sadness, and surprise (Ekman 1997). These emotions were identified through combinations of facial muscle movements, such as eyebrow raises, cheek dimples, and mouth openness (Lucey et al. 2010).

Table A28. Description of Measures Generated Using Text Mining and Deep Learning

	Variable	Description
Variables Created from Speech Expressions	Response Length	The combined number of words used by the applicant in answering the two interview questions.
	Topic Entropy	Number of topics in applicants' responses to the two interview questions; calculated using topics modeled using the Latent Dirichlet Allocation (LDA) algorithm. Higher values indicate a wider range of ideas in the response.
	Positive Document Sentiment	The proportion of tokens[a] in each document[b] dedicated to positive sentiment, leading to a linear measure of sentiment.
	Document Sentiment Variance	The standard deviation of positive sentiment of the tokens in the document.
	Response newness	The semantic similarity in the meanings of the two interview responses, created using cosine similarity between responses in the two interview videos.
Variables Created	Anger	The average probability of each emotion is calculated across all frames for the two videos. Probabilities are calculated using transfer learning, a deep
	Contempt	
	Disgust	

	Fear	learning technique that re-uses model parameters from one problem for a different but related problem (Lucey et al. 2010; Torrey and Shavlik 2010).
	Happiness	
	Sadness	
	Surprise	

Notes*:* Variables are coded at the level of analysis of each participant. Appendix: Table A26 provides details on cleaning the transcripts and variable creation. [a]Cleaned transcripts were tokenized, with each word being called a separate token. [b]Transcripts of both interview responses were cleaned and combined to create a document.

Changes in Displayed Verbal and Facial Expressions due to AI Transparency. For evaluating changes in displayed expressions, we used factor analysis on variables in Table A28, following Choudhury et al. (2019)[5]. This integrates co-varying verbal and non-verbal expressions into abstract patterns (Liu et al. 2020). Table A29 presents the results, retaining six factors with eigenvalues greater than 1, as per the Kaiser-Guttman rule (Jackson 1993). These six factors represent the predominant reaction patterns displayed by interviewees in Experiment I, with positively or negatively strong-loading variables highlighted. We predicted these factor scores and compared them across the experiment groups.

Among the six factors in Table A29, only factor 4 shows significant variation across the three experiment groups (Table A30). Interviewees associated with factor 4 exhibit high sentiment variance, low topic entropy, and a coexistence of positive (happiness) and negative (sadness) emotions in their facial expressions (Table A29). Relative to AVI-H, factor 4 is significantly higher in AVI-AI ($p < 0.05$, $d=0.37$). However, this increase is significantly mitigated in AVI-AITR compared to AVI-AI ($p < 0.05$, $d=0.28$). No significant difference is observed between AVI-H and AVI-AITR ($p > 0.05$, $d=0.09$).

Table A29. Factor Loadings of Measures Created from Speech Content and Facial Expressions

Variable	Factor 1	Factor 2	Factor 3	Factor 4	Factor 5	Factor 6
Positive Document Sentiment	0.15	0.17	0.59	0.13	-0.14	-0.02
Document Sentiment Variance	0.14	0.36	-0.24	0.49	0.17	0.04
Topic Entropy	0.08	0.48	-0.06	-0.34	0.06	-0.09
Response Newness	0.08	0.38	0.25	-0.13	-0.15	0.25
Response Length	0.11	0.59	-0.09	0.14	0.10	0.02
Video: Anger	-0.68	0.20	-0.07	0.06	-0.24	-0.04
Video: Contempt	0.68	-0.13	-0.17	-0.03	-0.24	-0.05
Video: Disgust	0.00	0.00	0.42	-0.45	0.13	0.37
Video: Fear	0.02	0.12	-0.02	-0.26	0.68	-0.34
Video: Happy	0.01	0.01	0.06	0.38	0.18	0.60
Video: Sad	0.00	-0.14	0.52	0.42	0.27	-0.36
Video: Surprise	-0.04	-0.20	-0.17	-0.04	0.46	0.42

Notes: Observations = 466; Values represent loadings of component variables on the first six factors of factor analysis. Loadings >0.3 are highlighted in the lighter shade and loadings < -0.3 are highlighted in darker shade.

Table A30. Comparison of Factors 4[a] (Table A29) across the Experiment groups

	Factor 4[b]		
	Mean	**(1)**	**(2)**
AVI-H (1)	*-0.15*	-	
AVI-AI (2)	*0.22*	0.37** (0.13)	-

[5] Choudhury et al. (2019) illustrate a methodology to observe verbal and non-verbal expressions. Their findings "are meant to illustrate the promise of our approach with ample room for future investigations," and are therefore constrained to the context of their data. Thus, our adaptation represents analysis of interviewees' reactions as it manifests in our dataset.

AVI-AITR (3)	*-0.06*	0.09 (0.11)	-0.28* (0.12)

Notes: [a]No significant difference in mean was observed across the three experiment groups for factors 1, 2, 3, 5, and 6; [b]cells indicate difference of row and column means compared using t-tests; * p<.05, ** p<.01; standard errors in parentheses.

Inferring Verbal and Facial Expressions. Interviewees often employ various verbal and non-verbal expressions to portray a favorable self-image (DePaulo et al. 2003; Ekman 1997). Despite such efforts, "leakage cues" such as fleeting, unfelt smiles or inconsistencies between speech and expression, often emerge (Ekman et al. 1988; O'Sullivan et al. 1988). Such expressions may be strategic or involuntary, reflecting goal pursuit or emotional dissonance (Mendoza et al. 2010; Porter et al. 2012). Our approach thus allows us to abstract patterns in reactions and observe them across experimental conditions but not comment on whether they are voluntary.

Factor 4, derived from our analysis, encapsulates a blend of positive (e.g., happiness) and negative (e.g., sadness) facial emotions (Table A29). Its verbal correlates include high sentiment variability and low topic diversity, suggesting pronounced emotional fluctuations paired with a limited range of articulated ideas. These combined verbal and facial cues indicate a coexistence of positive and negative emotions, which literature links to dissonance or ambivalence (Choudhury et al. 2019; Lakhiwal et al. 2023). Such patterns may reflect heightened load or discomfort, particularly under challenging conditions.

Notably, factor 4 was significantly higher in applicants interviewed using AVI-AI than AVI-H. This may suggest that the heightened deceptive IM, reduced honest SP, and elevated stress observed under AVI-AI may have compounded cognitive load, eliciting mixed emotional and verbal patterns (DePaulo et al. 2003). Interviewees may have attempted to mask these negative emotions with positive cues, creating the observed emotional and verbal inconsistency (Ekman et al. 1988). While the specificity of deception-related cues remains debated, elevated cognitive load, speech irregularities, and facial leakage cues are well-documented indicators of discomfort (DePaulo et al. 2003).

Conversely, applicants in the AVI-AITR condition exhibited significantly lower factor 4 scores compared to AVI-AI, nearly reverting to AVI-H levels. This reduction aligns with the lower deceptive IM, higher honest SP, and reduced stress observed in the transparent condition. By enhancing understandability, AVI-AITR appears to have mitigated the emotional dissonance and cognitive load experienced during algorithmic evaluation.

2.2.2. Validating Self-Reported Measures in Experiment I

We validated self-reported measures through replication and complementary approaches. In Experiment I, we measured IM and stress as key behavioral and affective reactions central to our examination. These measures were replicated in a quasi-field study conducted in a more ecologically relatable context, where participants engaged in a job-related interview process with realistic incentives and outcomes (Appendix: Section 2.1.3). The similar effects observed in this study provide robust support for the reliability of our self-reported measures. Specifically, for IM, evaluators in the quasi-field experiment assessed participants' baseline abilities based on an independently collected writing sample and reported perceived IM during subsequent interview evaluations. Additionally, Criteria-Based Content Analysis (CBCA) was employed to provide an objective validation of self-reported deceptive IM, further reinforcing the robustness of our findings.

2.3. Other Robustness Checks

2.3.1. Social Desirability Bias

Social desirability bias is defined as individuals' general tendency to portray a positive image of themselves while self-reporting (i.e., social desirability bias) (Larson 2019; Uziel 2010). We explicitly informed applicants that responses in the post-interview questionnaire would not affect compensation. Yet, dissimilarity in dispositional, i.e., invariant or personality-like, social desirability bias may explain heterogeneity in self-reported IM across the experiment groups (Bryan et al. 2013). We addressed the social desirability bias concern by measuring dispositional social desirability bias, i.e., the invariant tendency for interviewees to present themselves in a favorable light as per prevalent norms (King and Bruner 2000). We measured it before the manipulation presentation, and observed no significant differences across the three experiment groups ($F(2,463) = 0.22$, $p>.05$),[6] suggesting that our findings are unlikely to be contaminated by dispositional social desirability bias.

Furthermore, people often behave differently around AI relative to humans (Baird and Maruping 2021). Therefore, in addition to dispositional social desirability bias, applicants' comfort in confidential self-reporting of IM may vary due to evaluator type – human or AI. We call this evaluator bias, i.e., social desirability bias, after stimulus presentation. We adapted an item based on Gerber and Rogers (2009) to measure evaluator bias (

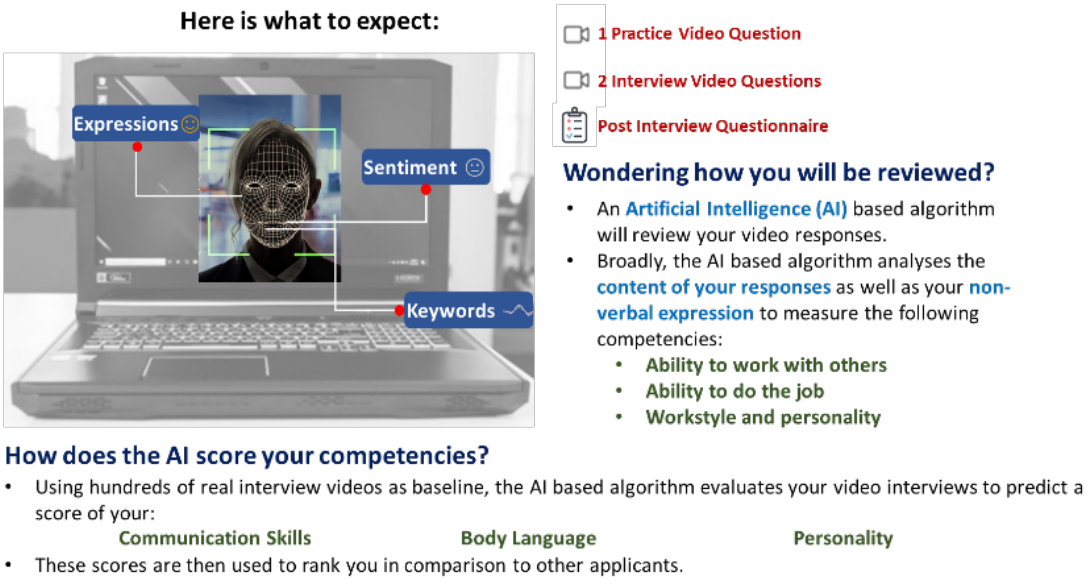
Table A5) at the end of the experiment and again observed no significant difference across the three experiment groups ($F(2,463) = 1.16$, $p>.05$). These tests cumulatively address concerns due to social desirability bias.

2.3.2. Salience of Manipulation

An alternate explanation of why AVI-AITR lowered and did not heighten deceptive EIC relative to AVI-AI is that the explanatory elements disclosed under AVI-AITR might not be sufficiently salient. We addressed this concern by developing an alternate operationalization of transparent AI (AVI-AITR2), where we append the explanatory elements by emphasizing that the AI algorithm uses hundreds of interviews as a baseline (what) to predict a score of communication skills, body language, and personality (how) that is ranked (why) for each applicant (Figure A8). We collected 160 observations with AVI-AITR2 (Table A31). The Kolmogorov–Smirnov test confirmed no significant difference in empirical distribution relative to data collected in Experiment I.

[6] Confirmed using post-hoc tests with Bonferroni correction.

Figure A8. Stimuli for AVI-AITR2



In the left panels a), b), and c) of

Table A32, the mean deceptive EIC (1.51 vs. 1.66), honest SP (4.50 vs. 4.49), and stress (-0.03 vs. -0.01) for AVI-AITR and AVI-AITR2 is presented. In the right panels of

Table A32, we report a pairwise comparison of these reactions relative to AVI-AI and AVI-AITR. Between AVI-AITR and AVI-AITR2, we observe no significant difference in deceptive EIC (

Table A32(a)), honest SP (

Table A32(b)), and stress (

Table A32(c)). Between AVI-AI and AVI-AITR2, results remain consistent. Altogether, these results adequately address the concerns due to the salience of the operationalization of transparent AI.

Table A31. Descriptive Statistics: Experiment I – Robustness Check (Alternate Conceptualization of Transparent AI, i.e., AVI-AITR2)

Variable	Mean	Std. Dev	Min	Max
Age	34.15	11.91	18.00	66.00
Gender[a]	0.46	0.50	0.00	1.00
Ethnicity[b]	0.06	0.23	0.00	1.00
Level of Education[c]	0.78	0.41	0.00	1.00
CRT Score[d]	1.53	1.20	0.00	3.00
Annual Income (Thousand USDs)	70.50	34.22	5.00	115.00
Deceptive EIC	1.66	1.06	1.00	7.00
Honest SP	4.49	1.29	1.00	7.00
Stress[e]	0.00	1.16	-3.33	4.67

Notes: Number of observations = 160; [a]Coded as 1 for males and 0 for non-males; [b]Coded as 1 for black or African American and 0 otherwise; [c]Coded as 1 for at least college degree and 0 otherwise; [d]Calculated as sum of correct responses to three CRT questions; [e]Standardized for analysis.

Table A32. Robustness Check: Salience of Manipulation (AVI-AITR2)

a)

Table A32a: Deceptive EIC[a]

	(1)	*(2)*
AVI-AI (1)	-	
AVI-AITR (2)	-0.54** (0.13)	-
AVI-AITR2 (3)	-0.40** (0.14)	0.15 (0.11)

b)

Table A32b: Honest SP[a]

	(1)	*(2)*
AVI-AI (1)	-	
AVI-AITR (2)	0.42** (0.15)	-
AVI-AITR2 (3)	0.40** (0.15)	-0.01 (0.15)

c)

Table A32c: Stress[a]

	(1)	*(2)*
AVI-AI (1)	-	
AVI-AITR (2)	-0.37** (0.14)	-
AVI-AITR2 (3)	-0.34** (0.13)	0.03 (0.14)

Notes: [a]Cells indicate row minus column means compared using t-test; * p<.05, ** p<.01, standard errors in parentheses.

3. Theoretical Contributions

Transparency in Information Systems: From Disclosure to Understandability

Transparency has long occupied a central position in Information Systems (IS) and organizational research as both a governance mechanism and a design principle. Early work on social translucence (Erickson and Kellogg 2000) highlighted that making social information visible, such as the presence or activities of others, can foster mutual awareness, coordination, and norm-guided behavior in digital systems. Transparency, in this foundational sense, serves to reduce uncertainty and promote trust by allowing actors to interpret others' actions. Schnackenberg and Tomlinson (2016) later formalized transparency as a multidimensional construct encompassing disclosure, clarity, and accuracy, emphasizing that its effectiveness depends on interpretability rather than volume: transparency succeeds when it enables sensemaking.

Nevertheless, IS research also cautions that transparency is a double-edged construct. Excessive or poorly designed transparency can overwhelm, distort, or even inhibit performance. Bernstein's (2012) transparency paradox vividly illustrates this: when factory workers were subjected to constant observation, they engaged in concealment behaviors, hindering learning and productivity. Similarly, Bannister and Connolly (2011) observed that in e-government settings, unbounded transparency eroded decision quality and raised privacy risks, calling for moderation and contextual calibration. More recent research reframes transparency as bidirectional (Gierlich-Joas et al. 2024)—granting visibility not only to organizational observers but also to employees about what data are collected about them—thereby rebalancing power and fostering mutual accountability. Collectively, this literature (Table A33) converges on a consistent theme: transparency must be designed for understandability, not maximal visibility. It is beneficial when it restores interpretive control and predictability but counterproductive when it induces surveillance anxiety or cognitive overload.

Table A33. Review of Prior work on Transparency in Related Contexts

Citation	Title	Journal	Key Findings
(Bannister and Connolly 2011)	The trouble with transparency: A critical review of openness in e-government	Policy & Internet	Transparency in e-government can sometimes harm good governance, so its scope and expectations must be carefully managed in the digital age.
(Bauer et al. 2023)	Expl(AI)ned: The impact of explainable artificial intelligence on users' information processing	Information Systems Research	Explanations in AI systems reshape users' mental models, which can both aid understanding and introduce confirmation bias, manipulation, and biased decisions.
(Bernstein 2012)	The transparency paradox: A role for privacy in organizational learning and operational control	Administrative Science Quarterly	Greater workplace transparency can paradoxically reduce performance by encouraging concealment, whereas limited privacy enhances productivity and innovation.

Citation	Title	Journal	Key Findings
(Binns et al. 2018)	"It's reducing a human being to a percentage": Perceptions of justice in algorithmic decisions	Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems	People's justice perceptions of algorithmic decisions depend more on exposure to multiple explanation styles than on any one explanation type.
(Choung et al. 2024)	When AI is perceived to be fairer than a human: Understanding perceptions of algorithmic decisions in a job application context	International Journal of Human-Computer Interaction	People perceive AI hiring decisions as fairer, more competent, and more trustworthy than human ones, especially when outcomes are negative.
(Dargnies et al. 2026)	Aversion to hiring algorithms: Transparency, gender profiling, and self-confidence	Management Science	Both workers and managers tend to prefer human over algorithmic hiring, though confidence feedback and fairness cues can increase algorithm acceptance.
(Dietvorst and Bartels 2022)	Consumers object to algorithms making morally relevant tradeoffs because of algorithms' consequentialist decision strategies	Journal of Consumer Psychology	Consumers reject algorithms making morally charged tradeoffs because they see algorithmic maximization as inappropriate in ethical contexts.
(Ehsan et al. 2021)	Expanding explainability: Towards social transparency in AI systems	Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems	Embedding social and organizational context into AI explanations fosters trust, better decisions, and organizational accountability.
(Erickson and Kellogg 2000)	Social translucence: An approach to designing systems that support social processes	ACM Transactions on Computer-Human Interaction	Designing digital systems that make users' actions visible promotes awareness, accountability, and effective collaboration.
(Gierlich-Joas et al. 2024)	Inverse transparency and the quest for empowerment through the design of digital workplace technologies	Journal of the Association for Information Systems	A "stewardship" approach to bidirectional transparency can balance employee empowerment with privacy in digital work environments.
(Gilliland 1993)	The perceived fairness of selection systems: An organizational justice perspective	Academy of Management Review	Fairness in hiring stems from procedural and distributive justice principles, which shape applicant and organizational outcomes.
(Jussupow et al. 2024)	An integrative perspective on algorithm aversion and	MIS Quarterly	The paper clarifies definitions and measurement of algorithm

Citation	Title	Journal	Key Findings
	appreciation in decision-making		aversion and appreciation to guide consistent, theory-driven future research.
(Kempf et al. 2024)	Transparency of algorithmic control systems and worker judgments	International Conference on Wirtschaftsinformatik	Transparency about an algorithm's transformation process enhances workers' acceptance and positive judgments of algorithmic control systems.
(Kizilcec 2016)	How much information? Effects of transparency on trust in an algorithmic interface	Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems	Moderate transparency fosters user trust, but too little or too much transparency can both undermine it.
(Klehe et al. 2008)	Transparency in structured interviews: Consequences for construct and criterion-related validity	Human Performance	Transparent interviews improve applicant performance and validity without compromising fairness or predictive power.
(Kordzadeh and Ghasemaghaei 2022)	Algorithmic bias: Review, synthesis, and future research directions	European Journal of Information Systems	Algorithmic bias affects fairness perceptions and technology adoption, but empirical research is lacking on the mechanisms linking bias to behavior.
(Kronblad et al. 2024)	When justice is blind to algorithms: Multilayered blackboxing of algorithmic decision making in the public sector	MIS Quarterly	Public institutions often ignore or obscure algorithmic harms, leading to legal and social injustices through "organizational blackboxing."
(Langer and König 2023)	Introducing a multi-stakeholder perspective on opacity, transparency, and strategies to reduce opacity in algorithm-based human resource management	Human Resource Management Review	Algorithmic opacity can both help and harm different HR stakeholders, requiring multi-stakeholder approaches to transparency management.
(Lavanchy et al. 2023)	Applicants' fairness perceptions of algorithm-driven hiring procedures	Journal of Business Ethics	Applicants perceive algorithmic recruitment as less fair than human or hybrid systems, fearing loss of individuality.
(Möhlmann et al. 2023)	Algorithm Sensemaking: How Platform Workers Make Sense of Algorithmic Management	Journal of the Association for Information Systems	Scholars urge cross-disciplinary research on algorithmic management that addresses both benefits and ethical concerns.

Citation	Title	Journal	Key Findings
(Ochmann et al. 2024)	Perceived algorithmic fairness: An empirical study of transparency and anthropomorphism in algorithmic recruiting	Information Systems Journal	Transparency and anthropomorphism interventions shape perceived fairness, especially in interpersonal and informational justice dimensions.
(Oostrom et al. 2024)	Applicant reactions to algorithm- versus recruiter-based evaluations of an asynchronous video interview and a personality inventory	Journal of Occupational and Organizational Psychology	Applicants react more negatively to algorithmic interview evaluations, perceiving them as less fair and less valid.
(Ostinelli et al. 2025)	Unintended effects of algorithmic transparency: The mere prospect of an explanation can foster the illusion of understanding how an algorithm works	Journal of Consumer Psychology	Knowing explanations exist creates a false sense of comprehension that leads to unwarranted trust in algorithmic recommendations.
(Rai 2020)	Explainable AI: From black box to glass box	Journal of the Academy of Marketing Science	Explainable AI can enhance trust, fairness, and privacy by balancing accuracy and explainability to meet stakeholder needs.
(Nagtegaal 2021)	The impact of using algorithms for managerial decisions on public employees' procedural justice	Government Information Quarterly	Algorithms improve perceived fairness in simple administrative tasks but reduce it in complex ones unless paired with human input.
(Schnackenberg and Tomlinson 2016)	Organizational transparency: A new perspective on managing trust in organization–stakeholder relationships	Journal of Management	Transparency builds stakeholder trust through information disclosure, clarity, and accuracy mechanisms.
(Suen and Hung 2024)	Revealing the influence of AI and its interfaces on job candidates' honest and deceptive impression management in asynchronous video interviews	Technological Forecasting and Social Change	The design of AI interview interfaces influences candidates' honest and deceptive impression management and anxiety levels.
(Tutun et al. 2024)	Explainable artificial intelligence for mental disorder screening: A computational design science approach	Journal of Management Information Systems	A transparent, explainable AI tool for mental disorder screening improves diagnostic accuracy and clinician trust.
(Wang and Benbasat 2007)	Recommendation agents for electronic commerce: Effects of explanation	Journal of Management Information Systems	Providing how, why, and trade-off explanations in online recommendation agents strengthens

Citation	Title	Journal	Key Findings
	facilities on trusting beliefs		consumers' trust by enhancing perceptions of competence, benevolence, and integrity.
(Xiong and Kim 2025)	Who wants to be hired by AI? How message frames and AI transparency impact individuals' attitudes and behaviors toward companies using AI in hiring	Computers in Human Behavior: Artificial Humans	Including AI transparency information in job advertisements improves applicants' trust, attitudes, and word-of-mouth toward companies using AI in hiring.
(Yeomans et al. 2019)	Making sense of recommendations	Journal of Behavioral Decision Making	Although recommender systems outperform humans at predicting preferences, people resist relying on them because they view human recommendations as more understandable.
Note: The set of papers summarized here reflects representative work across information systems and related fields addressing issues of transparency in algorithmic decision-making, recruitment or related contexts. The selection is illustrative rather than exhaustive and is intended to highlight salient conceptual developments rather than provide a systematic review.			

Transparency in Algorithmic and AI-Mediated Contexts

As decision-making increasingly relies on AI systems, IS and HCI research has examined how algorithmic opacity influences user behavior, perceptions, and outcomes. Algorithmic management studies (Kellogg et al. 2020; Möhlmann et al. 2023) reveal that opaque control systems heighten worker uncertainty, diminish autonomy, and prompt the formation of "folk theories" to explain algorithmic behavior. In such environments, transparency is sought not for curiosity but as a psychological safeguard against unpredictability.

Empirical studies demonstrate that algorithmic transparency yields mixed results. Xu et al. (2014) found that moderate transparency in recommender systems improved perceived decision quality, but too much information decreased satisfaction—revealing an inverted-U relationship between transparency and trust. Kizilcec (2016) observed that short, plain-language explanations restored user trust following adverse algorithmic outcomes, whereas excessive technical detail produced confusion and skepticism. Similarly, research on algorithm aversion shows that revealing an algorithm's reasoning or past performance can reduce resistance to automated systems (Dietvorst et al. 2015), but transparency must be comprehensible and relevant to user concerns. Conversely, Ostinelli et al. (2025) demonstrated that superficial or token transparency can create an illusion of understanding, leading to overconfidence and automation bias.

These findings have prompted a conceptual shift toward social or sociocentric transparency (Ehsan et al. 2021), which emphasizes contextualizing the algorithm within its social use environment: who created it, why it exists, and how its decisions are applied. This reorientation expands transparency from a technical property (explainability) to a relational construct that fosters social legibility. Langer and König (2023) further argue that transparency in algorithmic human resource management must balance stakeholders' goals—providing enough information to reassure candidates and regulators without enabling strategic gaming. Together, this literature underscores that transparency is not a fixed property of

systems but a design intervention whose efficacy depends on the audience, purpose, and social framing (see Table A33 for representative studies).

Transparency in AI-Based Recruitment and Selection

Algorithmic transparency becomes particularly consequential in high-stakes evaluative settings such as recruitment, where AI increasingly mediates screening, interviewing, and selection. Research documents a growing tension between efficiency and perceived fairness. Lavanchy et al. (2023) show that candidates view AI-only selection processes as less fair and more dehumanizing than human evaluations, even when outcomes are favorable, because algorithms are seen as incapable of appreciating individuality. Conversely, Choung et al. (2024) find evidence of algorithmic appreciation: under certain conditions, applicants consider algorithmic judgments more objective than human ones, particularly when transparency signals impartiality and bias mitigation. These opposing findings highlight that how algorithmic evaluation is communicated is as important as whether it is used. Xiong and Kim (2025) corroborate this, showing that transparency about how AI systems are audited and governed substantially improves candidates' trust and willingness to engage.

In the specific context of asynchronous video interviews (AVIs), transparency directly influences affective and behavioral responses. Lukacik et al. (2022) describe AVIs as creating a "void" of feedback where candidates perform for an unseen, algorithmic audience without cues or reciprocity, amplifying stress and defensive behavior. Oostrom et al. (2024) experimentally confirmed that merely disclosing that an AI would evaluate the interview heightened anxiety and lowered perceived fairness. Conversely, Suen and Hung (2024) found that providing feedback or clarifying evaluative criteria can affect stress.

Theoretical Gaps and Contributions of the Present Study

Despite these advances, the extant literature remains limited in three ways. First, most IS and algorithmic transparency studies emphasize perceptual outcomes (trust, fairness, acceptance) over behavioral and affective ones. We lack causal evidence on how transparency shapes concrete reactions, such as stress and impression management, of individuals being evaluated by algorithms. Second, much existing research treats transparency as a technocentric property of algorithms or interfaces, neglecting its sociocentric function as a bridge that restores interpretive symmetry between human evaluators and those being evaluated. Third, there is limited understanding of how transparency's benefits may erode when algorithmic scoring models themselves are misaligned with authentic performance signals. Our study addresses these gaps by theorizing and empirically testing transparency, as conceptualized from a sociocentric perspective, i.e., a form of transparency that restores understandability.

Notably, our study advances three domains of IS and management research. First, we extend IS theory by reinterpreting transparency as a structural solution to loss of user understandability, a mechanism that reduces ambiguity not by disclosing more data, but by making sociotechnical systems more legible and contextually intelligible. We introduce and validate sociocentric transparency as an evolution of social transparency (Ehsan et al. 2021), emphasizing the relational context (who/what/how) over technical exposition. This conceptual move shifts transparency from disclosure to understandability, aligning with calls for interpretability that centers the user's situated cognition.

We also redirect attention from model-level explainability to the user's experience in situ, i.e., examining how algorithmic audiences affect human affect and behavior. By demonstrating that transparency alleviates stress and discourages defensive impression management, our work expands the algorithmic decision-making literature to include affective and behavioral adaptation as core outcomes. We reconcile mixed findings across prior studies by showing that transparency's benefits depend on

context: what mitigates algorithm aversion (e.g., feature-based explanations) may differ from what reduces stress (e.g., evaluative clarity).

Furthermore, we contribute actionable implications for organizations adopting AI in hiring. Our results show that providing candidates with sociocentric transparency, explaining who evaluates them, what competencies are measured, and how the AI operates, enhances fairness perceptions, reduces anxiety, and fosters authentic responses. This improves both candidate experience and assessment integrity. However, we also highlight a caveat: transparency alone cannot compensate for flawed or biased algorithms. Transparency and algorithmic validity must coevolve; together, they form the foundation for ethical, trustworthy, and human-centered AI-HR systems. Altogether, our study bridges the micro-behavioral mechanisms of candidate reactions with the macro-design challenge of trustworthy algorithmic governance. We empirically demonstrate that transparency works best when it reintroduces legibility—when individuals can understand and anticipate how they will be evaluated. Yet, we also delineate its boundary conditions: transparency is necessary but not sufficient. It must operate in tandem with algorithmic accountability to achieve both fairness and psychological acceptability.

References

Baird, A., and Maruping, L. M. (2021) The Next Generation of Research on IS Use: A Theoretical Framework of Delegation to and from Agentic IS Artifacts *MIS Quarterly* 45(1):315–341.

Bannister, F., and Connolly, R. (2011) The Trouble with Transparency: A Critical Review of Openness in E-Government. *Policy & Internet* 3(1):1–30.

Bauer, K., von Zahn, M., and Hinz, O. (2023) Expl(AI)Ned: The Impact of Explainable Artificial Intelligence on Users' Information Processing. *Information Systems Research* 34(4):1582–1602.

Bernstein, E. S. (2012) The Transparency Paradox: A Role for Privacy in Organizational Learning and Operational Control. *Administrative Science Quarterly* 57(2):181–216.

Bill, B., Melchers, K. G., Steuer, J., and Eisele, E. (2024) Are Traditional Interviews More Prone to Effects of Impression Management Than Structured Interviews? *Applied Psychology* 73(3):1309–1330.

Binns, R., Van Kleek, M., Veale, M., Lyngs, U., Zhao, J., and Shadbolt, N. 2018. "'It's Reducing a Human Being to a Percentage': Perceptions of Justice in Algorithmic Decisions," *Proceedings of the 2018 CHI Conference on Human Factors in Computing Systems*, pp. 1–14.

Bourdage, J. S., Roulin, N., and Tarraf, R. (2018) "I (Might Be) Just That Good": Honest and Deceptive Impression Management in Employment Interviews. *Personnel Psychology* 71(4):597–632.

Bryan, C. J., Adams, G. S., and Monin, B. (2013) When Cheating Would Make You a Cheater: Implicating the Self Prevents Unethical Behavior. *Journal of Experimental Psychology: General* 142(4):1001.

Brynjolfsson, E., and McAfee, A. 2017. "Artificial Intelligence, for Real," in: *Harvard Business Review*.

Cho, J., Mirzaei, S., Oberg, J., and Kastner, R. 2009. "FPGA-Based Face Detection System Using Haar Classifiers," *Proceedings of the ACM/SIGDA international symposium on Field programmable gate arrays*, pp. 103–112.

Choudhury, P., Wang, D., Carlson, N. A., and Khanna, T. (2019) Machine Learning Approaches to Facial and Text Analysis: Discovering CEO Oral Communication Styles. *Strategic Management Journal* 40(11):1705–1732.

Choung, H., Seberger, J. S., and David, P. (2024) When AI Is Perceived to Be Fairer Than a Human: Understanding Perceptions of Algorithmic Decisions in a Job Application Context. *International Journal of Human–Computer Interaction* 40(22):7451–7468.

Coffman, K., and Klinowski, D. (2025) Gender and Preferences for Performance Feedback. *Management Science* 71(4):3497–3516.

Coffman, K. B., Collis, M. R., and Kulkarni, L. (2024) Whether to Apply. *Management Science* 70(7):4649–4669.

Dargnies, M.-P., Hakimov, R., and Kübler, D. (2026) Aversion to Hiring Algorithms: Transparency, Gender Profiling, and Self-Confidence. *Management Science* 72(1):285–301.

Dennis, A., Lakhiwal, A., and Sachdeva, A. (2023) AI Agents as Team Members: Effects on Satisfaction, Conflict, Trustworthiness, and Willingness to Work With. *Journal of Management Information Systems* 40(2):307–337.

DePaulo, B. M., Lindsay, J. J., Malone, B. E., Muhlenbruck, L., Charlton, K., and Cooper, H. (2003) Cues to Deception. *Psychological Bulletin* 129(1):74–118.

Dietvorst, B. J., and Bartels, D. M. (2022) Consumers Object to Algorithms Making Morally Relevant Tradeoffs Because of Algorithms' Consequentialist Decision Strategies. *Journal of Consumer Psychology* 32(3):406–424.

Dietvorst, B. J., Simmons, J. P., and Massey, C. (2015) Algorithm Aversion: People Erroneously Avoid Algorithms after Seeing Them Err. *Journal of Experimental Psychology: General* 144(1):114–126.

Ehsan, U., Liao, Q. V., Muller, M., Riedl, M. O., and Weisz, J. D. 2021. "Expanding Explainability: Towards Social Transparency in AI Systems," *Proceedings of the 2021 CHI Conference on Human Factors in Computing Systems*, pp. 1–19.

Ekman, P. (1997) Should We Call It Expression or Communication? *Innovation: The European Journal of Social Science Research* 10(4):333–344.
Ekman, P., Friesen, W. V., and O'Sullivan, M. (1988) Smiles When Lying. *Journal of Personality and Social Psychology* 54(3):414–420.
Ekman, P., and O'Sullivan, M. (2006) From Flawed Self-Assessment to Blatant Whoppers: The Utility of Voluntary and Involuntary Behavior in Detecting Deception. *Behavioral Sciences & the Law* 24(5):673–686.
Ellis, A. P., West, B. J., Ryan, A. M., and DeShon, R. P. (2002) The Use of Impression Management Tactics in Structured Interviews: A Function of Question Type? *Journal of Applied Psychology* 87(6):1200.
Erickson, T., and Kellogg, W. A. (2000) Social Translucence: An Approach to Designing Systems That Support Social Processes. *ACM Transactions on Computer-Human Interaction* 7(1):59–83.
Feng, S., Zhou, H. Y., and Dong, H. B. (2019) Using Deep Neural Network with Small Dataset to Predict Material Defects. *Materials & Design* 162:300–310.
Frid-Adar, M., Diamant, I., Klang, E., Amitai, M., Goldberger, J., and Greenspan, H. (2018) GAN-Based Synthetic Medical Image Augmentation for Increased CNN Performance in Liver Lesion Classification. *Neurocomputing* 321:321–331.
Fügener, A., Grahl, J., Gupta, A., and Ketter, W. (2021) Will Humans-in-the-Loop Become Borgs? Merits and Pitfalls of Working with AI. *MIS Quarterly* 45:1527–1556.
Gerber, A. S., and Rogers, T. (2009) Descriptive Social Norms and Motivation to Vote: Everybody's Voting and So Should You. *Journal of Politics* 71(1):178–191.
Gierlich-Joas, M., Baiyere, A., and Hess, T. (2024) Inverse Transparency and the Quest for Empowerment through the Design of Digital Workplace Technologies. *Journal of the Association for Information Systems* 25(5):1212–1239.
Gilliland, S. W. (1993) The Perceived Fairness of Selection Systems: An Organizational Justice Perspective. *Academy of Management Review* 18(4):694–734.
Gioaba, I., and Krings, F. (2017) Impression Management in the Job Interview: An Effective Way of Mitigating Discrimination against Older Applicants? *Frontiers in Psychology* 8:770.
Guo, W., Sengul, M., and Yu, T. (2021) The Impact of Executive Verbal Communication on the Convergence of Investors' Opinions. *Academy of Management Journal* 64(6):1763–1792.
Helfat, C. E., and Peteraf, M. A. (2015) Managerial Cognitive Capabilities and the Microfoundations of Dynamic Capabilities. *Strategic Management Journal* 36(6):831–850.
Ioffe, S., and Szegedy, C. 2015. "Batch Normalization: Accelerating Deep Network Training by Reducing Internal Covariate Shift," *International Conference on Machine Learning*, pp. 448–456.
Jackson, D. A. (1993) Stopping Rules in Principal Components Analysis: A Comparison of Heuristical and Statistical Approaches. *Ecology* 74(8):2204–2214.
Jiang, L., Yin, D., and Liu, D. (2019) Can Joy Buy You Money? The Impact of the Strength, Duration, and Phases of an Entrepreneur's Peak Displayed Joy on Funding Performance. *Academy of Management Journal* 62(6):1848–1871.
Jussupow, E., Benbasat, I., and Heinzl, A. (2024) An Integrative Perspective on Algorithm Aversion and Appreciation in Decision-Making. *MIS Quarterly* 48(4):1575–1590.
Kamath, U., Liu, J., and Whitaker, J. 2019. *Deep Learning for NLP and Speech Recognition*. Springer.
Kellogg, K. C., Valentine, M. A., and Christin, A. (2020) Algorithms at Work: The New Contested Terrain of Control. *Academy of Management Annals* 14(1):366–410.
Kempf, M., Simić, F., Alizadeh, A., and Benlian, A. 2024. "Transparency of Algorithmic Control Systems and Worker Judgments," in: *International Conference on Wirtschaftsinformatik*.
King, M. F., and Bruner, G. C. (2000) Social Desirability Bias: A Neglected Aspect of Validity Testing. *Psychology & Marketing* 17(2):79–103.
Kizilcec, R. F. 2016. "How Much Information? Effects of Transparency on Trust in an Algorithmic Interface," *Proceedings of the 2016 CHI Conference on Human Factors in Computing Systems*, pp. 2390–2395.

Klehe, U.-C., König, C. J., Richter, G. M., Kleinmann, M., and Melchers, K. G. (2008) Transparency in Structured Interviews: Consequences for Construct and Criterion-Related Validity. *Human Performance* 21(2):107–137.
Kordzadeh, N., and Ghasemaghaei, M. (2022) Algorithmic Bias: Review, Synthesis, and Future Research Directions. *European Journal of Information Systems* 31(3):388–409.
Krieger, N., Smith, K., Naishadham, D., Hartman, C., and Barbeau, E. M. (2005) Experiences of Discrimination: Validity and Reliability of a Self-Report Measure for Population Health Research on Racism and Health. *Social Science & Medicine* 61(7):1576–1596.
Kronblad, C., Essén, A., and Mähring, M. (2024) When Justice Is Blind to Algorithms: Multilayered Blackboxing of Algorithmic Decision Making in the Public Sector. *MIS Quarterly* 48(4):1637–1662.
Lakhiwal, A., Bala, H., and Léger, P.-M. (2023) Ambivalence Is Better Than Indifference: Behavioral and Neurophysiological Assessment of Ambivalence in Online Environments. *MIS Quarterly* 47(2):705–732.
Langer, M., Baum, K., König, C. J., Hähne, V., Oster, D., and Speith, T. (2021) Spare Me the Details: How the Type of Information About Automated Interviews Influences Applicant Reactions. *International Journal of Selection and Assessment* 29(2):154–169.
Langer, M., and König, C. J. (2023) Introducing a Multi-Stakeholder Perspective on Opacity, Transparency and Strategies to Reduce Opacity in Algorithm-Based Human Resource Management. *Human Resource Management Review* 33(1):100881.
Langer, M., König, C. J., and Fitili, A. (2018) Information as a Double-Edged Sword: The Role of Computer Experience and Information on Applicant Reactions Towards Novel Technologies for Personnel Selection. *Computers in Human Behavior* 81:19–30.
Larson, R. B. (2019) Controlling Social Desirability Bias. *International Journal of Market Research* 61(5):534–547.
Lavanchy, M., Reichert, P., Narayanan, J., and Savani, K. (2023) Applicants' Fairness Perceptions of Algorithm-Driven Hiring Procedures. *Journal of Business Ethics* 188(1):125–150.
LeCun, Y., Bengio, Y., and Hinton, G. (2015) Deep Learning. *Nature* 521(7553):436–444.
Lemaitre, G., Nogueira, F., and Aridas, C. K. (2017) Imbalanced-Learn: A Python Toolbox to Tackle the Curse of Imbalanced Datasets in Machine Learning. *Journal of Machine Learning Research* 18(1):559–563.
Lever, J., Krzywinski, M., and Altman, N. (2016) Points of Significance: Model Selection and Overfitting. 13(9):703–704.
Liu, X., Zhang, B., Susarla, A., and Padman, R. (2020) Go to YouTube and Call Me in the Morning: Use of Social Media for Chronic Conditions. *MIS Quarterly* 44(1):257–283.
Lucey, P., Cohn, J. F., Kanade, T., Saragih, J., Ambadar, Z., and Matthews, I. 2010. "The Extended Cohn-Kanade Dataset (Ck+): A Complete Dataset for Action Unit and Emotion-Specified Expression," *2010 IEEE Computer Society Conference on Computer Vision and Pattern Recognition-Workshops*: IEEE, pp. 94–101.
Lukacik, E.-R., Bourdage, J. S., and Roulin, N. (2022) Into the Void: A Conceptual Model and Research Agenda for the Design and Use of Asynchronous Video Interviews. *Human Resource Management Review* 32(1):100789.
McCarthy, J., and Goffin, R. (2004) Measuring Job Interview Anxiety: Beyond Weak Knees and Sweaty Palms. *Personnel Psychology* 57(3):607–637.
Mendoza, S. A., Gollwitzer, P. M., and Amodio, D. M. (2010) Reducing the Expression of Implicit Stereotypes: Reflexive Control through Implementation Intentions. *Personality and Social Psychology Bulletin* 36(4):512–523.
Mirowska, A., and Mesnet, L. (2022) Preferring the Devil You Know: Potential Applicant Reactions to Artificial Intelligence Evaluation of Interviews. *Human Resource Management Journal* 32(2):364–383.

Möhlmann, M., Alves de Lima Salge, C., and Marabelli, M. (2023) Algorithm Sensemaking: How Platform Workers Make Sense of Algorithmic Management. *Journal of the Association for Information Systems* 24(1):35–64.
Nagtegaal, R. (2021) The Impact of Using Algorithms for Managerial Decisions on Public Employees' Procedural Justice. *Government Information Quarterly* 38(1):101536.
Newman, A., Bavik, Y. L., Mount, M., and Shao, B. (2021) Data Collection via Online Platforms: Challenges and Recommendations for Future Research. *Applied Psychology* 70(3):1380–1402.
Ng, H.-W., Nguyen, V. D., Vonikakis, V., and Winkler, S. 2015. "Deep Learning for Emotion Recognition on Small Datasets Using Transfer Learning," *Proceedings of the 2015 ACM on International Conference on Multimodal Interaction*, pp. 443–449.
O'Sullivan, M., Ekman, P., and Friesen, W. V. (1988) The Effect of Comparisons on Detecting Deceit. *Journal of Nonverbal Behavior* 12(3):203–215.
Ochmann, J., Michels, L., Tiefenbeck, V., Maier, C., and Laumer, S. (2024) Perceived Algorithmic Fairness: An Empirical Study of Transparency and Anthropomorphism in Algorithmic Recruiting. *Information Systems Journal* 34(2):384–414.
Ong, A. D., Bergeman, C. S., Bisconti, T. L., and Wallace, K. A. (2006) Psychological Resilience, Positive Emotions, and Successful Adaptation to Stress in Later Life. *Journal of Personality and Social Psychology* 91(4):730–749.
Oostrom, J. K., Holtrop, D., Koutsoumpis, A., van Breda, W., Ghassemi, S., and de Vries, R. E. (2024) Applicant Reactions to Algorithm-Versus Recruiter-Based Evaluations of an Asynchronous Video Interview and a Personality Inventory. *Journal of Occupational and Organizational Psychology* 97(1):160–189.
Ostinelli, M., Bonezzi, A., and Lisjak, M. (2025) Unintended Effects of Algorithmic Transparency: The Mere Prospect of an Explanation Can Foster the Illusion of Understanding How an Algorithm Works. *Journal of Consumer Psychology* 35(2):203–219.
Palan, S., and Schitter, C. (2018) Prolific.ac—a Subject Pool for Online Experiments. *Journal of Behavioral and Experimental Finance* 17:22–27.
Porter, S., Ten Brinke, L., and Wallace, B. (2012) Secrets and Lies: Involuntary Leakage in Deceptive Facial Expressions as a Function of Emotional Intensity. *Journal of Nonverbal Behavior* 36(1):23–37.
Powell, D. M., Bourdage, J. S., and Bonaccio, S. (2021) Shake and Fake: The Role of Interview Anxiety in Deceptive Impression Management. *Journal of Business and Psychology* 36(5):829–840.
Rai, A. (2020) Explainable AI: From Black Box to Glass Box. *Journal of the Academy of Marketing Science* 48(1):137–141.
Rizi, M. S., and Roulin, N. (2024) Does Media Richness Influence Job Applicants' Experience in Asynchronous Video Interviews? Examining Social Presence, Impression Management, Anxiety, and Performance. *International Journal of Selection and Assessment* 32(1):54–68.
Roth, L., Klehe, U.-C., and Willhardt, G. (2021) Liar, Liar, Pants on Fire: How Verbal Deception Cues Signal Deceptive Versus Honest Impression Management and Influence Interview Ratings. *Personnel Assessment and Decisions* 7(1):7.
Roulin, N., Bangerter, A., and Levashina, J. (2014) Interviewers' Perceptions of Impression Management in Employment Interviews. *Journal of Managerial Psychology* 29(2):141–163.
Roulin, N., Bangerter, A., and Levashina, J. (2015) Honest and Deceptive Impression Management in the Employment Interview: Can It Be Detected and How Does It Impact Evaluations? *Personnel Psychology* 68(2):395–444.
Roulin, N., and Powell, D. M. (2018) Identifying Applicant Faking in Job Interviews. *Journal of Personnel Psychology* 17(3):143–154.
Santurkar, S., Tsipras, D., Ilyas, A., and Madry, A. (2018) How Does Batch Normalization Help Optimization? *Advances in Neural Information Processing Systems* 31.
Schnackenberg, A. K., and Tomlinson, E. C. (2016) Organizational Transparency: A New Perspective on Managing Trust in Organization-Stakeholder Relationships. *Journal of Management* 42(7):1784–1810.

Schudlik, K., Reinhard, M.-A., and Müller, P. (2023) The Relationship between Preparation, Impression Management, and Interview Performance in High-Stakes Personnel Selection: A Field Study of Airline Pilot Applicants. *The International Journal of Aerospace Psychology* 33(2):120–138.
Shaha, M., and Pawar, M. 2018. "Transfer Learning for Image Classification," *2018 Second International Conference on Electronics, Communication and Aerospace Technology (ICECA)*: IEEE, pp. 656–660.
Srinivasa-Desikan, B. 2018. *Natural Language Processing and Computational Linguistics: A Practical Guide to Text Analysis with Python, Gensim, spaCy, and Keras*. Packt Publishing Ltd.
Stevens, K., Kegelmeyer, P., Andrzejewski, D., and Buttler, D. 2012. "Exploring Topic Coherence over Many Models and Many Topics," *Proceedings of the 2012 Joint Conference on Empirical Methods in Natural Language Processing and Computational Natural Language Learning*, pp. 952–961.
Suen, H.-Y., and Hung, K.-E. (2024) Revealing the Influence of AI and Its Interfaces on Job Candidates' Honest and Deceptive Impression Management in Asynchronous Video Interviews. *Technological Forecasting and Social Change* 198:123011.
Torrey, L., and Shavlik, J. 2010. "Transfer Learning," in *Handbook of Research on Machine Learning Applications and Trends: Algorithms, Methods, and Techniques*. IGI global, pp. 242–264.
Tutun, S., Topuz, K., Tosyali, A., Bhattacherjee, A., and Li, G. (2024) Explainable Artificial Intelligence for Mental Disorder Screening: A Computational Design Science Approach. *Journal of Management Information Systems* 41(4):958–981.
Uziel, L. (2010) Rethinking Social Desirability Scales: From Impression Management to Interpersonally Oriented Self-Control. *Perspectives on Psychological Science* 5(3):243–262.
Vrij, A. (2005) Criteria-Based Content Analysis: A Qualitative Review of the First 37 Studies. *Psychology, Public Policy, and Law* 11(1):3.
Vrij, A., and Mann, S. (2006) Criteria-Based Content Analysis: An Empirical Test of Its Underlying Processes. *Psychology, Crime & Law* 12(4):337–349.
Wang, F., Cheng, J., Liu, W. Y., and Liu, H. J. (2018) Additive Margin Softmax for Face Verification. *IEEE Signal Processing Letters* 25(7):926–930.
Wang, W., and Benbasat, I. (2007) Recommendation Agents for Electronic Commerce: Effects of Explanation Facilities on Trusting Beliefs. *Journal of Management Information Systems* 23(4):217–246.
Weiss, B., and Feldman, R. S. (2006) Looking Good and Lying to Do It: Deception as an Impression Management Strategy in Job Interviews. *Journal of Applied Social Psychology* 36(4):1070–1086.
Wen, L., Li, X., Li, X., and Gao, L. 2019. "A New Transfer Learning Based on VGG-19 Network for Fault Diagnosis," *2019 IEEE 23rd International Conference on Computer Supported cooperative Work in Design (CSCWD)*: IEEE, pp. 205–209.
Xiong, Y., and Kim, J. K. (2025) Who Wants to Be Hired by AI? How Message Frames and AI Transparency Impact Individuals' Attitudes and Behaviors toward Companies Using AI in Hiring. *Computers in Human Behavior: Artificial Humans* 3:100120.
Xu, J., Benbasat, I., and Cenfetelli, R. T. (2014) The Nature and Consequences of Trade-Off Transparency in the Context of Recommendation Agents. *MIS Quarterly* 38(2):379–406.
Yeomans, M., Shah, A., Mullainathan, S., and Kleinberg, J. (2019) Making Sense of Recommendations. *Journal of Behavioral Decision Making* 32(3):403–414.
Zhang, Z., and Sabuncu, M. 2018. "Generalized Cross Entropy Loss for Training Deep Neural Networks with Noisy Labels," *Advances in Neural Information Processing Systems*, pp. 8778–8788.