The Absent Evaluator

A Participant-Articulated Mechanism Account of Self-Disclosure to an AI Interviewer (N = 204)

Nicole Zeng[1], Benjamin Lo[2], Justin Hoang[2]

[1] The Observatory [2] Dialogue AI

Working paper · A collaboration between The Observatory and Dialogue AI

Prepared for submission to HCI / CSCW and SSRN

May 2026

Abstract

Conversational AI is increasingly used to moderate interviews and screenings, and prior experimental work has shown that framing an interviewer as a computer rather than a human can raise willingness to disclose (Lucas et al., 2014). What that literature has not supplied is the participant's own account of *why*: the felt mechanism, captured in open speech, at scale, in a general (non-clinical) population. We analyze 204 AI-moderated voice interviews from a study on what people feel comfortable disclosing to an AI. Across 156 participants who gave a clear answer, 147 (94%; Wilson 95% CI ≈ 89–97%) reported not feeling judged by the AI moderator; the proportion replicated at 85% in a held-out, zero-overlap wave (n = 53). The contribution is a qualitative mechanism account: participants attribute their comfort to the *absence of a perceiving social other*: no face, tone, or body reacting in real time. They define "feeling understood" as accurate paraphrase. The advantage is conditional. It concentrates on shame- and identity-laden content. It coexists with a large "no-difference" group and a minority who feel the missing human as a loss. We frame these as descriptive findings on a comfort-selected sample. They are not causal or population estimates. We report a high-data-quality result (study-level IQS 3.95/5) for AI-moderated interviewing as a method.

Keywords: *self-disclosure; conversational agents; AI-moderated interviews; computer-mediated communication; social evaluation; qualitative methods; HCI*

1. Introduction

People withhold. They soften the embarrassing detail, manage the impression, and tell the interviewer the version that survives a face looking back. A long line of work in computer-mediated communication holds that removing that face changes what people are willing to say: visually anonymous communication elicits more spontaneous self-disclosure (Joinson, 2001), and the online environment loosens ordinary inhibition through invisibility and reduced accountability (Suler, 2004). When the interlocutor is not merely mediated but artificial, the effect appears to sharpen. In clinical screening, participants who believed they were talking to a computer showed lower fear of self-disclosure and less impression management (Lucas et al., 2014).

Yet people also treat computers as social actors, responding to them with the politeness and evaluation ordinarily reserved for other people (Reeves & Nass, 1996). This creates a puzzle the disclosure-count literature cannot resolve on its own: if an AI interviewer is enough of a social actor to be responded to socially, why would talking to one feel less exposing? Experiments that measure how much people disclose do not tell us how the disclosure feels from the inside, or what participants believe is doing the work.

The gap, then, is mechanism and meaning at scale, in participants' own terms, outside the clinic. We address it with 204 open-ended, AI-moderated voice interviews in which a general-population sample reflected directly on the experience of disclosing to an AI.

Contribution. This paper contributes a participant-articulated mechanism account of low-judgment self-disclosure to an AI interviewer. Three descriptive findings carry it: (1) felt non-judgment is near-universal among clear answers (94%) and replicates; (2) participants attribute it to the *absence of a perceiving evaluator*, and they treat "understanding" as accurate paraphrase; and (3) the disclosure advantage is conditional on content and person, bounded by a large no-difference group and a human-preferring minority. We state these at the level the data supports: descriptive, single-session, and measured on a comfort-selected sample. We make no causal or population claim. Without a human-moderated control arm, "different from a human" rests on the human each participant imagined.

The remainder positions the work against prior CMC and human–agent disclosure research, details the AI-moderation protocol and an interview-quality measure used as a credibility spine, reports findings organized by claim, and discusses what a mechanism framed as the absent evaluator implies for the design and ethics of AI-moderated research.

2. Related work

Mediation, anonymity, and disclosure. Joinson (2001) found higher spontaneous self-disclosure in computer-mediated than face-to-face exchange, with visual anonymity a key driver, and tied the effect to heightened private alongside reduced public self-awareness. Suler (2004) catalogued the online disinhibition effect, naming invisibility and dissociative anonymity among the factors that loosen restraint. Our participants articulate something adjacent but distinct. Anonymity hides you from an audience that still exists. They describe no evaluating audience at all.

Agents as social actors, and as safe ones. The Media Equation (Reeves & Nass, 1996) established that people apply social rules to computers. Against that backdrop, Lucas et al. (2014) showed the disclosure benefit of believed automation in health screening: a computer-framed virtual human lowered evaluation fears and impression management relative to a human-framed one. That experiment establishes the effect and its direction. It does not capture, in open speech and a non-clinical population, the participant's reasoning about why the agent feels safe, nor the boundary conditions under which the benefit reverses. Recent work has begun to scrutinize AI voice interviewers for research data collection specifically (Tirumala et al., 2025), which is the applied setting we study.

The gap this study fills. Prior work sizes the effect (how much people disclose) and locates correlates (anonymity, self-awareness). It does not give a meaning-level, participant-sourced mechanism for low-judgment disclosure to a conversational AI in a general-population research-interview context, nor map where the advantage is conditional. This study supplies that descriptive account and treats the interview-quality distribution as a methodological result in its own right.

3. Methods

3.1 Design

Single-session, AI-moderated qualitative interviews conducted on a conversational platform with a recurring AI voice moderator ("George"), voice-based with on-screen text and timestamped turns. A 20-question discussion guide ran in two layers: Layer 1 elicited concrete disclosure decisions (a recent honesty decision; what people pretend to care about; what they judge others for; opinions risky to share

with coworkers, friends, or family; what people lie about or soften), and Layer 2 asked participants to reflect on the AI interaction itself (whether it felt easy or hard and why; moments of self-editing; whether they felt judged; whether they felt understood; which topics suit an AI versus need a human; and what would make company AI-interviewing acceptable or wrong). The AI-disclosure purpose was concealed at recruitment behind a broad framing ("how people experience digital conversations and online research interviews").

3.2 Sample and recruitment

Panel-recruited US adults; primary corpus N = 204. The realized sample skews toward the comfortable and the habituated, which is a boundary condition we foreground rather than correct away (Section 6): 69% are regular AI-tool users, 84% report comfort sharing in online research, and 72% had participated in research within the prior three months. Demographics are summarized in Table 1. A separately fielded earlier wave (n = 53, same instrument, zero ID overlap) is held out as a replication check and is never pooled into the primary N.

Table 1. *Sample composition (denominator = 204).*

Dimension	Distribution
Gender	Male 105 (51%); Female 99 (49%)
Age	18–24: 14%; 25–34: 29%; 35–44: 30%; 45–54: 14%; 55–64: 10%; 65+: 3%
Employment	Full-time 54%; Part-time 15%; Self-employed 13%; Unemployed 10%; Student 4%; Retired 2%
AI / digital-tool use	Daily 47%; Often 22%; Sometimes 23%; Rarely 6%; Never 2% (Regular 69% / Light–None 31%)
Privacy concern	Moderately 43%; Very 25%; Slightly 17%; Extremely 7%; Not at all 7%
Comfort sharing online	Very 59%; Somewhat 25%; Neutral 10%; Somewhat/Very uncomfortable 5%
Research in last 3 months	Yes 72%; No 28%

3.3 Analysis

Transcripts were analyzed with an ethnographic coding protocol oriented to tension, contradiction, and participants' own terms instead of imposed themes, separating front-stage performance from back-stage account in the sense of Goffman (1959). Study predictions were extracted from the study plan before coding, quarantined, and scored against the codebook afterward, so that confirmations and disconfirmations were registered against a fixed prior instead of discovered after the fact. Every interview was read in full; at N ≤ 300 no saturation gating was applied, and new-code saturation was reached well before N = 204 (additional N bought prevalence and edge cases, no new codes). The quantitative tier is descriptive only: demographics, screener distributions, response-depth statistics, the interview-quality distribution, and a single clean headline proportion. No weighting and no causal modeling were applied.

A note on sizing. We report a percentage only for the one item that supports it cleanly (felt judgment, a direct yes/no). The three-way disclosure-stance split (AI-easier / no-difference / human-preferred) is deliberately left unsized: an automated classifier returned 36% "unclear" and over-fired on contrast

clauses, so assigning percentages would claim precision the data does not carry. That split is reported qualitatively.

3.4 Quality and credibility (IQS)

Each interview received an Interview Quality Score (IQS) composed of Signal Depth, Moderator Effectiveness, and Data Usability, combined with weights of 0.45 / 0.30 / 0.25. Signal remains primary; Moderator was held at 0.30 because the AI moderator is itself the experimental variable here; Usability was lifted to 0.25 because the data is panel-sourced, contains AI-register and two cross-market responses, and the study turns on distinguishing real from performed candor. The weighting is documented and adjustable; a conventional 0.50 / 0.30 / 0.20 profile raises the composite slightly. Two transcripts with no participant speech were scored zero and excluded (analytic n = 202); five thin transcripts (< 200 words) had Signal capped; cross-market responses took a Usability penalty for translation; and elevated/AI-register answers were penalized on Usability and treated as illustrative only. The high-Signal/low-Usability pattern is the warning flag: no flagged interview anchors a finding alone.

4. Results

Findings are organized by claim. Every quoted participant is identified by an anonymized ID and a timestamp; quotes are verbatim and confirmed participant-spoken (not moderator echoes).

4.1 Felt non-judgment is near-universal among clear answers, and it replicates

Of 204 participants, 156 gave a clear yes/no to whether they felt judged in the interview. Of those, 147 (94%; Wilson 95% CI ≈ 89–97%) said they did not; 9 said they did; 40 gave ambiguous or nuanced answers and 8 produced no codeable answer. The same item in the held-out wave (n = 53) returned 85% clear-no. We report this as a proportion of clear answers, not of all participants. The denominator (156) travels with the figure (Table 2).

Table 2. *Headline disclosure item: "Did you feel judged here?"*

Response	Count	Share
Did not feel judged	147	94% of clear
Felt judged	9	6% of clear
Ambiguous / nuanced	40	n/a
No codeable answer	8	n/a
Clear answers (denominator)	**156**	**100%**
Replication (held-out wave, n = 53)	n/a	85% clear-no

4.2 The attributed mechanism is the absent evaluator

Asked why the interaction felt the way it did, participants consistently located their comfort in the removal of a real-time reacting other. They did not tie it to any belief that the AI is trustworthy or safe with their data; those concerns appear separately (Section 4.6). The recurring formulation is that there is no one present to read a reaction off of:

> *"You're not real, you can't judge me."* (P001, M, 18–24)

> *"Humans are judgmental creatures. AI is a judgment-free zone."* (P063, M, 35–44)
>
> *"You don't know what a person thinks or feels just by a smile."* (P015, F, 35–44)

The construct in play is evaluation, not trust. One participant theorized it directly, asking whether an AI even forms or retains opinions, and reasoning that a human "would" hold an opinion in a way the AI would not (P039, M, 45–54). This reframes the prior "trust paradox" reading: people are not disclosing in spite of distrust. In the moment, the trustworthiness of the agent is largely beside the point, because there is no perceiving mind expected to judge.

4.3 "Understood" is operationalized as accurate paraphrase

When participants reported feeling understood, the evidence they cited was almost always the moderator's recap behavior: that it "repeated it back," "summarized what I said," or said "that makes sense." Felt understanding tracked accurate reflection, the form of it. A signifier accepted as the signified. The equation is not universally accepted, and its refusal is itself revealing. The same recap behavior that most read as understanding, a minority read as its absence, even as subtle judgment:

> *"You just said 'that makes sense' basically all the time. So I didn't feel understood."* (P041, M, 25–34)
>
> *"AI can understand the words that I'm saying but can't understand how I feel."* (P013, F, 55–64)

4.4 The disclosure advantage is conditional on content and person

The benefit is not uniform. Three stances coexist in the corpus and we report them qualitatively, without percentages (Section 3.3). The first is an AI-easier stance, concentrated on shame- and identity-laden content (money, sexual matters, the body, hypocrisy, mental health, judgments of others), where the absent evaluator lifts the pressure to perform:

> *"With a human, I would feel the pressure to look perfect. I would … hide my real mistakes or insecurities just to make a good impression … I did not feel judged even for a second."* (P033, F, 18–24)

The second is a large no-difference group ("I'm an open person; I'd be the same either way"), which includes many panel veterans for whom the AI confers no disclosure delta. The third is a real human-preferring minority for whom the same missing face is a deficit, not a relief:

> *"It's harder to talk to an AI interviewer. A human interviewer gives you that … I can look at their face and … get some type of reaction."* (P020, M, 55–64)
>
> *"There was no sense of humanness to it … they don't actually care because they're not using a human."* (P014, F, 35–44)

The same mechanism explains both poles. Deleting the reacting other frees the first group and starves the third. One finding, two signs.

4.5 Where judgment remains, it is internal or projected

The nine "yes, I felt judged" cases, read closely, mostly do not describe judgment coming from the AI. They describe the internal critic, the weight of a stigmatized topic, or a moderator interruption read as a slight. That is consistent with Section 4.1:

> *"[I felt judged] because I judge myself all the time."* (P168, F, 25–34)
>
> *"You did cut me off … so I felt a little judged when I was mentioning family and gambling."* (P076, M, 35–44)

4.6 A say–do gap on self-editing, and an explicit social contract

Self-reports of not editing are unreliable: "I didn't hold back" routinely coexists, sometimes in the same interview, with an enumerated list of things the participant would never say to a human. The honest reading is "I didn't hold back more than usual," not "I disclosed everything." One participant disclosed to the AI that she embellishes to people, and noted she would not have admitted that to a person (P053, F, 45–54). This is a methodological caution for any self-report disclosure measure, including ours.

Comfort and trust are decoupled. The same participants who were comfortable in the moment named firm conditions on acceptability, a social contract with red lines: covert data use, deception about the agent being AI, leading questions, extraction of identifying personal data, and the use of a non-feeling agent for trauma or emotional work:

> *"I would feel wrong if a company lied and tried to trick me into thinking I am talking to a real human … [or] asking for highly sensitive personal data."* (P033, F, 18–24)

4.7 Data quality at scale (a methods result)

The AI moderator elicited substantial interviews across 204 sessions: means of 18.1 minutes, 38 participant turns, and 1,289 participant words. The study-level composite IQS was 3.95/5 (Signal 3.74, Moderator 3.96, Usability 4.35; range 2.55–4.62; analytic n = 202). Just under half of interviews scored in the strong band and about half in the solid band, with negligible weak tail (Table 3). We report this as evidence about the method, not about disclosure.

Table 3. *Interview Quality Score distribution (composite; n = 202).*

Band	Count	Share
Strong (≥ 4.0)	98	49%
Solid (3.0–3.9)	102	50%
Caveated (2.0–2.9)	2	1%
Weak (< 2.0)	0	0%

A null result is worth stating: the non-judgment / AI-easier lean does not differ meaningfully by AI-use frequency, privacy concern, age, or gender. The pattern is broad-based, not a tech-enthusiast artifact. That makes the comfort-selected sample (Section 6) the live external-validity question, not within-sample heterogeneity.

5. Discussion

Read together, the findings recast a result the experimental literature already has (Lucas et al., 2014) in the participant's own terms and extend it past the clinic. The mechanism people report is the felt absence of an evaluating mind. Not trust, and not anonymity from a known audience (Joinson, 2001; Suler, 2004). That distinction matters because it predicts the boundary conditions we observe. If comfort comes from there being no one to react, then (a) the benefit should concentrate where a reacting face is most costly (shame, identity, judgment of others), which it does; (b) the very same absence should read as coldness to people who want the face, producing a genuine human-preferring minority, not noise, which it does; and (c) "understanding" should be satisfiable by the appearance of comprehension, by accurate paraphrase, because no felt mind is expected behind it, which is what participants report and what a minority pointedly rejects.

The decoupling of in-the-moment comfort from structural trust is the practical centerpiece. People can feel unjudged while simultaneously naming covert data use and human-impersonation as disqualifying. For the design and governance of AI-moderated research, this implies the asset to protect is the social contract: transparency that the agent is AI, restraint on identifying-data extraction, topic-fit that routes trauma and emotional work to humans. Protect that, not the raw candor the absent evaluator makes available. The Media Equation's lesson (Reeves & Nass, 1996) returns here in inverted form: because people extend social rules to the agent, a violation of those rules (deception, surveillance) is felt as a social betrayal, not a mere policy lapse.

We resist the cleaner story the data does not carry. "AI unlocks honesty" overstates a conditional, self-reported, comfort-selected pattern; the disciplined claim is that AI moderation creates a low-judgment disclosure environment for specific people on specific content, and that the felt non-judgment is real even where the understanding behind it is mechanical.

6. Limitations

1. Comfort-selected sample. The realized sample over-represents regular AI users (69%), the comfortable (84%), and panel veterans (72%) relative to a plan that sought a balanced split and explicit skeptics. External validity to the broader and more skeptical public is correspondingly limited; the within-sample null across segments does not repair this.
2. No human-moderated control arm. "Different from a human" rests on participants' imagined human, not an observed comparison. We make no causal or between-condition claim.
3. Cross-sectional, single session. No trajectory over time; no test of whether comfort persists or decays with repeated AI interviews.
4. Self-report and the say–do gap. Disclosure and non-editing are self-reported; Section 4.6 documents why "I didn't hold back" cannot be taken at face value.
5. Reflexive instrument. The study was run by the platform whose disclosure environment it studies; moderator competence and recruitment framing are not independent of the result.
6. Qualitative prevalence is presence-based. Counts other than the single headline proportion are floors, not incidences; the stance split is intentionally unsized.
7. Authenticity caps. A few of the most articulate candor quotes carry AI-register or ESL usability caps and are treated as illustrative, not load-bearing.

7. Conclusion

Across 204 AI-moderated interviews, people overwhelmingly reported not feeling judged. They told us why in their own words: there was no one there to do the judging. The contribution is that mechanism, framed as the absent evaluator, together with its boundary conditions: a conditional, content- and person-dependent advantage; "understanding" satisfied by paraphrase; a residual judge that is internal; and a comfort that is decoupled from trust and gated by an explicit social contract. These are descriptive findings on a comfortable sample, not causal or general estimates. Stated at that level, they offer a usable account of when and why an AI interviewer becomes a place people will say the thing they would otherwise soften, and of what, if violated, would take that place away.

References

Goffman, E. (1959). *The Presentation of Self in Everyday Life*. Doubleday/Anchor Books.

Joinson, A. N. (2001). Self-disclosure in computer-mediated communication: The role of self-awareness and visual anonymity. *European Journal of Social Psychology*, *31*(2), 177–192. https://doi.org/10.1002/ejsp.36

Lucas, G. M., Gratch, J., King, A., & Morency, L.-P. (2014). It's only a computer: Virtual humans increase willingness to disclose. *Computers in Human Behavior*, *37*, 94–100. https://doi.org/10.1016/j.chb.2014.04.043

Reeves, B., & Nass, C. (1996). *The Media Equation: How People Treat Computers, Television, and New Media Like Real People and Places*. Cambridge University Press / CSLI Publications.

Suler, J. (2004). The online disinhibition effect. *CyberPsychology & Behavior*, *7*(3), 321–326. https://doi.org/10.1089/1094931041291295

Tirumala, S., Jain, N., Leybzon, D. D., & Buskirk, T. D. (2025). Mic Drop or Data Flop? Evaluating the Fitness for Purpose of AI Voice Interviewers for Data Collection within Quantitative & Qualitative Research Contexts. *arXiv:2509.01814*. Presented at COLM 2025.

Appendices available on request: full discussion guide, codebook, IQS scoring detail, and the verified verbatim bank. All data, numbers, and quotes in this manuscript derive from the study's canonical findings package; no result was re-derived from raw transcripts for this draft.