🌻 People are often more candid with machines than with other people. Why?

📅 23 Jun 2025

(Gambino et al. 2020)

People are often more candid with an AI interviewer than with a human interviewer, especially about sensitive or embarrassing topics. This is not a magic property of “AI truth serum”; it’s a predictable consequence of how people manage impressions, respond to perceived judgement, and adapt their social scripts to machines. In practice, increased candour can be a **feature** (less socially desirable responding) and a **risk** (over-disclosure, misplaced trust, and biased measurement).

Mechanisms (why candour often increases)

Reduced evaluation pressure (less impression management)

When the “other” is perceived as non-judgemental (or at least less socially consequential), respondents can feel less need to perform competence, morality, status, or emotional stability. One way to phrase it: the *cost of looking bad* is lower, so people spend less effort on impression management and can answer more directly. Evidence from virtual-human interviewing suggests that simply believing the interviewer is automated can reduce fear of self-disclosure and impression management, and increase willingness to disclose in health-screening contexts (Lucas et al. 2014). Relatedly, in a hiring context, chatbot-based personality assessment was found to be **less susceptible to social desirability bias** than traditional psychometric tests (though with trade-offs in predictive validity) (Dukanovic & Krpan 2025).

Different “social script” for machines (CASA, updated)

The CASA paradigm argues that people often apply social rules to computers (e.g., politeness, reciprocity), even when they know the system is not a person (Gambino et al. 2020). The important twist for candour is that the script can be *social-but-not-socially-risky*: respondents may treat the AI as a conversational partner while simultaneously perceiving reduced judgement, reduced gossip risk, and reduced reputational spillovers.

Greater control over pacing and phrasing

Many AI interview modalities (especially text chat) give respondents more control: time to think, revise, and choose precise wording. That can reduce anxiety and improve articulation, which can look like "candour" because the answer is clearer and less hedged. (This can also enable strategic responding—see risks below.)

Psychological distance and "as-if anonymity"

Even when the system is not truly anonymous, the *felt* experience can be closer to anonymous disclosure: no facial expressions, fewer micro-judgements, and less immediate social feedback. That psychological distance can be particularly salient when the alternative is a high-status interviewer or a context with strong power asymmetries (e.g., workplace, clinic, immigration, legal).

Differential effects (who becomes more candid, when)

Topic sensitivity matters

The effect is strongest when questions touch stigma, shame, taboo behaviours, or norm violations (e.g., mental health, sexual behaviour, substance use, unethical conduct, politically sensitive views). If the question is neutral and low-stakes, you should not expect large candour gains—people already answer candidly.

Trust, literacy, and perceived surveillance can flip the effect

Candour can *decrease* if respondents suspect monitoring, future consequences, or data misuse ("this is going on my record"). In those cases, an AI interviewer can feel like a **surveillance interface** rather than a safe partner, and respondents may become more guarded than they would with a trusted human.

The interface design changes the psychology

More humanlike embodiments (voice, face, avatar) can increase rapport, but may also increase evaluation pressure. Conversely, a plain text interface can reduce social pressure, but might reduce emotional safety for some people. CASA-style responses can be triggered by subtle cues, so small design changes can shift disclosure dynamics (Gambino et al. 2020).

Population differences are real (and can become bias)

People differ in comfort with technology, prior experiences with institutions, cultural norms around authority, and baseline social anxiety. That means "AI increases candour" is a

distributional claim, not a universal one: you can get systematic differences in who discloses what, which then affects your dataset.

Risks and things to think about

Over-disclosure and informed consent

If respondents feel unusually safe, they may disclose more than they later feel comfortable with. Make consent and data-handling salient at the moment of disclosure (not just in a long preamble), and be explicit about whether a human will read the transcript.

Strategic responding and “gaming” the system

Lower social desirability pressure does not imply higher truthfulness. Some respondents may optimize for what they think the algorithm rewards (or what they think the organization wants). The hiring study is a useful reminder: reduced social desirability bias can coexist with weaker predictive validity and new failure modes (Dukanovic & Krpan 2025).

AI introduces its own interviewer effects

AI systems can inadvertently lead respondents (tone, follow-up choices, perceived empathy), and can behave inconsistently across demographic groups or writing styles. Even if the interview is consistent, downstream AI-assisted analysis can introduce non-random errors: work on LLMs for qualitative analysis shows serious risks of biased annotation and misleading inferences when errors correlate with respondent characteristics (Ashwin et al. 2023).

Safety boundaries

If the topic includes trauma, self-harm, abuse, or crisis, an “always-on, friendly” conversational agent can create ethical and duty-of-care hazards. Decide up front what the AI should and should not do in those scenarios, and ensure participants are not misled about the system’s capabilities.

We do not recommend using any AI interviewer for employment interviews or in a clinical setting without human oversight and significant safeguards.

Practical takeaway

Treat “AI increases candour” as a **design variable**: you can dial it up or down via framing (“AI” vs “automated”), embodiment, reminders about audience (private vs reviewed), and question style. The key is to choose the level of candour you want, *and* to document the conditions so you can interpret the data responsibly.

1. my comment ↩

References

Ashwin, Chhabra, & Rao (2023). *Using Large Language Models for Qualitative Analysis Can Introduce Serious Bias*. https://doi.org/10.48550/arXiv.2309.17147.

Dukanovic, & Krpan (2025). *Comparing Chatbots to Psychometric Tests in Hiring: Reduced Social Desirability Bias, but Lower Predictive Validity*. Frontiers. https://doi.org/10.3389/fpsyg.2025.1564979.

Gambino, Fox, & Ratan (2020). *Building a Stronger CASA: Extending the Computers Are Social Actors Paradigm*. Communication and Social Robotics Labs. https://doi.org/10.3316/INFORMIT.097034846749023.

Lucas, Gratch, King, & Morency (2014). *It's Only a Computer: Virtual Humans Increase Willingness to Disclose*. https://doi.org/10.1016/j.chb.2014.04.043.