Splitting Source Evaluation and Compliance: AI Avatars and Behavioral Intention in Health Communication

Alessio Sartore*, Paola Signori**

*Adjunct Professor, University of Verona, alessio.sartore@univr.it
**Full Professor, Department of Management, University of Verona, paola.signori@univr.it

Abstract

Artificial intelligence (AI) avatars are moving into preventive and mental-health communication faster than the research needed to evaluate them. A central assumption in this literature is that source credibility drives compliance, but it has rarely been tested against a transparently disclosed AI source across health domains of contrasting emotional intensity. This study asks whether the negative perceptions typically associated with AI sources (greater eeriness, lower trustworthiness, lower perceived accuracy) actually translate into reduced behavioral intention. We ran a 2 (source: human vs disclosed AI) × 2 (domain: mammography vs psychotherapy) between-subjects experiment with 338 women aged 40–80, recruited through Prolific. Stimuli presented an identical health recommendation attributed either to a human physician or to a hyper-realistic AI avatar explicitly identified as artificial; there was no concealment. Two-way ANOVAs confirmed that participants perceived the AI source as significantly more eerie, less trustworthy, and modestly less accurate. Yet the source had no effect at all on behavioral intention. Domain, by contrast, mattered considerably: mammography screening generated much higher compliance intention than psychotherapy. The dissociation between how the source was evaluated and how participants intended to act suggests that in health contexts the personal stakes carried by a message can override source-credibility judgments entirely. The findings challenge the classical credibility-to-compliance chain and indicate that disclosed AI avatars can function as effective channels for preventive health communication, without any need for artificial humanization. Limitations include self-reported intention and a female 40–80 sample; field experiments are needed to verify whether the paradox holds for actual behavior.

Keywords: AI avatars; health communication; source credibility; uncanny valley; behavioral intention

1. Introduction

Hyper-realistic AI avatars are appearing with increasing frequency in preventive health campaigns, patient education, and mental-health support (Gomez-Cabello et al., 2025; Longoni et al., 2019). Once produced, AI avatars can in principle deliver health messages at scale and without the costs associated with professional recording sessions. This is a practical advantage that partly explains their growing adoption. Physician avatars have already been piloted for postoperative education with acceptable usability results (Gomez-Cabello et al., 2025), and conversational AI in mental-health care has expanded to the point of warranting systematic scoping reviews (Ni & Jia, 2025). The practical question organizations now face is whether to disclose the artificial nature of these sources, and, if they do, what the disclosure costs them. The dominant answer from marketing and persuasion research is pessimistic: source credibility theory holds that trustworthiness and expertise are the primary drivers of message acceptance (Ohanian, 1990), and both are reliably damaged by AI disclosure (Wortel et al., 2024; Longoni

et al., 2019). Studies of AI-generated advertising content find that labelling a source as artificial reduces perceived credibility and consumer attitudes (Wortel et al., 2024). Consumers also show a specific resistance to AI in medical contexts, preferring human providers for consequential health decisions (Longoni et al., 2019). Taken together, these findings predict a clear penalty for disclosed AI sources: lower credibility should yield lower compliance. We test that prediction directly. Using a controlled experiment, we compare a fully disclosed AI avatar with a human physician across two health domains (mammography screening and psychotherapy) chosen to sample health domains of contrasting emotional intensity. Our outcome is behavioral intention, a measure one step closer to action than general attitude, though we acknowledge it remains self-reported. The result is a source-compliance split: even when the AI source is evaluated less favorably across every perceptual dimension we measured, behavioral intention is statistically equivalent to that generated by the human source. Source evaluation and compliance intention appear to operate independently in personally consequential health contexts. This is a finding with direct implications for both theory and practice.

2. Theoretical Framework

The starting point is Mori's (1970/2012) uncanny valley hypothesis, which predicts that as an artificial agent approaches but fails to perfectly match human appearance, familiarity collapses into eeriness. Ho and MacDorman (2010) provided the psychometric infrastructure to measure this reaction, and subsequent marketing research has confirmed the effect in commercial settings (Song & Shin, 2022; Kim et al., 2024). Song and Shin (2022) found that chatbot uncanniness depresses trust and purchase intention in e-commerce, while Kim, Ryoo, and Choi (2024) extended the phenomenon to what they call an 'uncanny valley of mind': anthropomorphic AI agents can disrupt advertising effectiveness when users feel unsettled by the agent's apparent mental capacities. The basic tension identified by the uncanny valley literature is that greater realism can simultaneously increase engagement and amplify eeriness, leaving the behavioral consequences of realistic AI avatars empirically open.

A different prediction comes from the Computers Are Social Actors (CASA) paradigm (Nass & Moon, 2000). CASA holds that people apply social heuristics to computers mindlessly, attributing social competence to non-human agents even when they know the agent is artificial. This implies that disclosure alone need not produce the credibility penalty source-credibility theory anticipates because some degree of social transfer to the machine will persist regardless. The qualification is that disclosure can activate skeptical, persuasion-knowledge-based processing that partially offsets this transfer (Wortel et al., 2024), and whether social attribution or skepticism wins out is likely to depend on the stakes of the communication context.

The Stereotype Content Model (Fiske, Cuddy, Glick, & Xu, 2002) offers a way to think about why these two processes need not move together. The model distinguishes warmth from competence as independent dimensions of social perception: in this study mapped into trustworthiness and perceived accuracy. An entity can be judged competent but cold, or warm but unreliable. The two dimensions vary independently. This independence is the structural mechanism behind the paradox we investigate. Recent evidence supports the dissociation: Sun et al. (2025) found that LLM-generated health information can attract higher accuracy ratings while receiving lower trust scores than human-generated content: this is a pattern that anticipates our own results.

Source credibility theory (Ohanian, 1990) synthesizes these dimensions into a persuasion model positioning trustworthiness and expertise as the principal antecedents of compliance. Applied straightforwardly to a disclosed AI source, the theory predicts a double penalty: both the warmth dimension (trustworthiness) and the competence dimension (accuracy) should be reduced, and compliance should follow downward. The empirical record in health settings is

consistent with this prediction at the attitudinal level, meaning that consumers resist medical AI and prefer human providers (Longoni et al., 2019). What the literature has not adequately addressed is whether attitudinal penalties translate into behavioral ones in health domains where personal stakes are presumed to be high. That gap is where our argument lies. The Elaboration Likelihood Model (Petty & Cacioppo, 1986) provides the integrating mechanism: under the central route, recipients process message content deeply when personal relevance is high; under the peripheral route, they rely on heuristic cues (for example source identity) when involvement is low.

When a health message concerns a potentially life-altering risk (for instance the possibility of an undetected cancer, or the prospect of meaningful relief from psychological distress) the content may carry sufficient personal relevance to dominate how people process it. Under conditions of high personal relevance, message substance tends to receive deeper scrutiny than source characteristics (Petty & Cacioppo, 1986): when what is being said carries direct implications for a person's own health, who is saying it may matter considerably less. Under this account, source identity functions as a peripheral cue when involvement is low. When stakes are high, it is subordinate to message substance. In health domains such as those studied here, source evaluation appears to have less influence on what people decide to do. This is the theoretical core of the source-compliance split: message stakes can separate source evaluation from behavioral intention. This asymmetry is reflected in the data as the domain effect on behavioral intention ($\eta^2 = .169$) is substantially larger than the source effect ($\eta^2 = .001$), and that is consistent with central-route processing driven by message content rather than source identity. Three hypotheses follow.

H1 predicts that a disclosed AI source will be perceived as more eerie and less trustworthy than a human source.

H2 predicts that the AI source will be perceived as less accurate.

H3, that is the core of this study, predicts that despite H1 and H2, behavioral intention will not differ between the two sources.

3. Methodology

The study used a 2 (source: human vs disclosed AI) x 2 (domain: mammography vs psychotherapy) between-subjects design. Participants were 338 women aged 40–80 recruited through Prolific. The age range was chosen deliberately: it broadly covers the age range targeted by European mammography screening guidelines (45–74 years; European Commission Initiative on Breast Cancer), while including adjacent age groups for whom opportunistic screening is clinically relevant. Psychotherapy was selected as a contrasting domain because mental health care represents a distinct and increasingly prominent component of preventive health, recognised as integral to overall wellbeing (World Health Organization, 2022). An initial 346 respondents participated, 8 were excluded for failing the source-recognition manipulation check (they identified the source incorrectly or responded “do not remember”), leaving 174 participants in the AI conditions and 164 in the human conditions. Cell sizes were: AI-mammography n=83, AI-psychotherapy n=91, human-mammography n=78, human-psychotherapy n=86.

Stimuli held the persuasive text constant across all four conditions and varied only the source and the health domain. In the human conditions, the message was attributed to a female physician (Dr. Elena Ferrini) and a female psychotherapist (Dr. Mariasole Conti). In the AI conditions, they were attributed to the hyper-realistic avatar version of the same human physician (named Medai) and of the same psychotherapist (named Luna). The AI versions were explicitly identified as artificial intelligence assistants. There was no concealment and no attempt to soften or humanize the source beyond its visual appearance. The domain manipulation consisted of message content focused either on mammography screening or on

beginning psychotherapy, matched in length, syntactic structure, and the intensity of the persuasive appeal.

Figure 1. *Experimental stimuli across the four conditions.*

Condition 1: Human source: Mammography (Dr. Elena Ferrini)
Condition 2: AI source: Mammography (Medai)

Condition 3: Human source: Psychotherapy (Dr. Mariasole Conti)
Condition 4: AI source: Psychotherapy (Luna)

Note. All stimuli presented identical persuasive text; source identity was manipulated through visual presentation and explicit labelling. AI sources were explicitly identified as artificial intelligence assistants.

Four constructs were measured using validated scales adapted from established instruments in the literature. Behavioral intention and perceived accuracy were measured on seven-point Likert scales. Trustworthiness and eeriness were measured using seven-point semantic-differential scales. Behavioral intention was assessed with three items developed for this study following Ajzen's (1991) conceptualization of domain-specific intention, measuring the likelihood of booking an appointment, seeking further information, and recommending the service to others ($\alpha = .90$). Trustworthiness was measured with Ohanian's (1990) seven-item semantic-differential subscale ($\alpha = .95$). Perceived accuracy was assessed with three items from Appelman and Sundar (2016): accurate, authentic, believable ($\alpha = .92$). Eeriness was measured with Ho and MacDorman's (2010) five-item semantic-differential scale ($\alpha = .94$). Participants were recruited via Prolific, completed the questionnaire on SurveyMonkey, and data were analyzed in Jamovi using two-way ANOVAs for each dependent variable.

4. Results

The manipulation was successful: 174 participants correctly identified the AI source and 164 the human source; 8 were excluded for failing the check. Table 1 reports means and standard deviations by source and domain; Table 2 presents the full ANOVA results.

Table 1. Means and standard deviations by source and by domain

Variable	AI M (SD)	Human M (SD)	Mammography M (SD)	Psychotherapy M (SD)
Behavioral intention	4.82 (1.64)	4.72 (1.78)	5.50 (1.43)	4.11 (1.67)
Trustworthiness	5.08 (1.29)	5.73 (1.31)	5.73 (1.31)	5.09 (1.29)
Accuracy	5.25 (1.37)	5.64 (1.36)	5.82 (1.26)	5.09 (1.39)
Eeriness	4.38 (1.45)	5.25 (1.34)	4.92 (1.56)	4.71 (1.37)

Note. N = 338. All scales 1–7. SD in parentheses.

Table 2. Two-way ANOVA effects (F, p, η^2) for each dependent variable

Dependent variable	Source	Domain	Source × Domain
Behavioral intention	0.21, $p = .650$, $\eta^2 = .001$	68.94, $p < .001$, $\eta^2 = .169$	4.74, $p = .030$, $\eta^2 = .012$
Trustworthiness	22.98, $p < .001$, $\eta^2 = .061$	22.11, $p < .001$, $\eta^2 = .059$	3.89, $p = .049$, $\eta^2 = .010$
Accuracy	7.81, $p = .005$, $\eta^2 = .021$	26.43, $p < .001$, $\eta^2 = .071$	1.66, $p = .198$, $\eta^2 = .004$
Eeriness	34.62, $p < .001$, $\eta^2 = .092$	2.41, $p = .121$, $\eta^2 = .006$	7.65, $p = .006$, $\eta^2 = .020$

Note. N = 338. df = (1, 334) for behavioral intention and accuracy; (1, 326) for trustworthiness; (1, 331) for eeriness, due to missing item responses.

Source effects

H1 was supported. The AI source was perceived as significantly more eerie than the human source, $F(1, 331) = 34.62$, $p < .001$, $\eta^2 = .092$, and significantly less trustworthy, $F(1, 326) = 22.98$, $p < .001$, $\eta^2 = .061$. H2 was also supported: accuracy ratings were lower for the AI, $F(1, 334) = 7.81$, $p = .005$, $\eta^2 = .021$, though this effect was small. The AI source thus carried a perceptual penalty across all three dimensions.

Behavioral intention: the Source-Compliance Split

H3 was supported. Source had no effect on behavioral intention, $F(1, 334) = 0.21$, $p = .650$, $\eta^2 = .001$. Intentions toward the AI source ($M = 4.82$) and the human source ($M = 4.72$) were statistically indistinguishable. Domain produced the largest effect in the design, $F(1, 334) = 68.94$, $p < .001$, $\eta^2 = .169$, with mammography ($M = 5.50$) generating considerably higher intention than psychotherapy ($M = 4.11$). The accuracy effect reaching significance with $N = 338$ is worth noting: the AI was penalized on every perceptual dimension, yet behavioral intention did not move. Small but significant source x domain interactions appeared for behavioral intention, $F(1, 334) = 4.74$, $p = .030$, $\eta^2 = .012$, for trustworthiness, $F(1, 326) = 3.89$, $p = .049$, $\eta^2 = .010$, and for eeriness, $F(1, 331) = 7.65$, $p = .006$, $\eta^2 = .020$. Accuracy showed no interaction, $F(1, 334) = 1.66$, $p = .198$, $\eta^2 = .004$.

5. Discussion

The data tell a clean story. Participants in the AI conditions found the source more eerie, trusted it less, and rated its content as slightly less accurate, yet they reported the same intention to act as participants who saw a human physician or psychotherapist. Source evaluation and behavioral intention, in other words, appear to have gone separate ways.

The pattern is not what a straightforward application of source credibility theory would predict. Ohanian's (1990) model, and the substantial literature that has built on it, treats trustworthiness and expertise as the gatekeepers of persuasion: damage the source, and compliance should follow. In low-stakes or attitude-only designs, that logic holds well enough (Wortel et al., 2024; Longoni et al., 2019). Here, in a higher stakes setting, it does not. One likely reason is that the health messages in this study carried personal consequences, such as the possibility of an undetected cancer, or the prospect of meaningful psychological relief. Personal relevance of this kind is known to shift persuasion toward message-centered rather than source-centered processing (Petty & Cacioppo, 1986), and the current data are consistent with that account.

The result also stands apart from recent work on AI avatars in cancer-screening contexts, where AI sources have been found to lower behavioral intention relative to human ones (Wang et al., 2026). A key difference between those studies and ours is disclosure. Where prior designs conceal or only partially signal the AI source, our study disclosed it fully. Transparency did not recover the perceptual damage (participants still found the AI eerie and less trustworthy) but it did not generate a behavioral cost either. This suggests a possible reframing: full disclosure may shift the processing of the AI cue from a vague unease to an explicit acknowledgment that participants can recognize, accept, and move past when the message content is sufficiently involving.

The domain effect deserves its own reading. Mammography generated substantially higher compliance intention than psychotherapy ($\eta^2 = .169$, by far the largest effect in the design). The significant source x domain interactions for behavioral intention, trustworthiness, and eeriness

indicate that the relationship between source evaluation and compliance does vary by domain: the override of source by message stakes is not uniform across health contexts. The psychotherapy domain showed lower compliance intention overall, and the source x domain interactions suggest that the relationship between source evaluation and behavioral intention may vary across health contexts in ways worth exploring with larger and more diverse samples The implications for marketers in health communication are more direct. Disclosed AI avatars appear to be a viable instrument for preventive health communication without any behavioral cost from transparency. There is no persuasive case for concealing the AI or for investing in artificial humanization to compensate for an eeriness penalty that, in health contexts at least, does not reach behavior. This finding is particularly relevant in the current regulatory environment: the EU AI Act's requirements for disclosure of AI-generated content (European Parliament, 2024) turn out to be compatible with persuasive effectiveness, at least in the domains studied here.

6. Limitations

The outcome measure was self-reported behavioral intention, which is at best a proxy for actual behavior. The intention-behavior gap is well documented (Sheeran & Webb, 2016), and the absence of a source effect on intention does not guarantee a similar absence in actual behavior. Another limitation concerns the sample: women aged 40-80 were chosen deliberately for their relevance to both health domains, but findings may not generalize to other age groups or cultural contexts. The study also assumed that both health domains engaged high personal stakes for participants, but this was not directly measured, so we think that future research should include explicit measures of perceived personal relevance to test whether the source-compliance split depends on message stakes. A related assumption concerns the high-stakes framing of the health context itself is that the study includes no low-stakes control condition, so the claim that personal relevance drives central-route processing remains theoretically grounded but empirically unverified.

7. Future Research

The most pressing extension is a field experiment: a genuine A/B test in a real health communication setting would establish whether the behavioral equivalence we observed for intention holds for actual appointment-booking or treatment-seeking. Longitudinal designs could examine whether the eeriness response to AI sources weakens with familiarity, which matters considerably for health services that patients return to repeatedly. Message framing represents a third direction: it remains unclear whether AI sources benefit from particular communicative styles that capitalize on perceived competence while managing the warmth deficit. Replication on age-diverse samples and across health domains with different levels of perceived barriers would help clarify the boundary conditions of the source-compliance split.

8. References

Ajzen, I. (1991). The theory of planned behavior. *Organizational Behavior and Human Decision Processes*, 50(2), 179–211. https://doi.org/10.1016/0749-5978(91)90020-T

Appelman, A., & Sundar, S. S. (2016). Measuring message credibility: Construction and validation of an exclusive scale. *Journalism & Mass Communication Quarterly*, 93(1), 59–79. https://doi.org/10.1177/1077699015606057

European Commission Initiative on Breast Cancer. (2022). *European guidelines on breast cancer screening and diagnosis*. Publications Office of the European Union.

European Parliament. (2024). Regulation (EU) 2024/1689 of the European Parliament and of the Council laying down harmonised rules on artificial intelligence (Artificial Intelligence Act). *Official Journal of the European Union*.

Fiske, S. T., Cuddy, A. J. C., Glick, P., & Xu, J. (2002). A model of (often mixed) stereotype content: Competence and warmth respectively follow from perceived status and competition. *Journal of Personality and Social Psychology*, 82(6), 878–902. https://doi.org/10.1037/0022-3514.82.6.878

Gomez-Cabello, C. A., Borna, S., Pressman, S., Haider, S. A., Haider, C. R., & Forte, A. J. (2025). Artificial intelligence physician avatars for patient education: A pilot study. *Journal of Clinical Medicine*, 14(23), 8595. https://doi.org/10.3390/jcm14238595

Ho, C.-C., & MacDorman, K. F. (2010). Revisiting the uncanny valley theory: Developing and validating an alternative to the Godspeed indices. *Computers in Human Behavior*, 26(6), 1508–1518. https://doi.org/10.1016/j.chb.2010.05.015

Kim, J., Ryoo, Y., & Choi, Y. (2024). That uncanny valley of mind: When anthropomorphic AI agents disrupt personalized advertising. *International Journal of Advertising*, 44(8), 1684–1713. https://doi.org/10.1080/02650487.2024.2411669

Longoni, C., Bonezzi, A., & Morewedge, C. K. (2019). Resistance to medical artificial intelligence. *Journal of Consumer Research*, 46(4), 629–650. https://doi.org/10.1093/jcr/ucz013

Mori, M. (2012). The uncanny valley (K. F. MacDorman & N. Kageki, Trans.). *IEEE Robotics & Automation Magazine*, 19(2), 98–100. https://doi.org/10.1109/MRA.2012.2192811 (Original work published 1970)

Nass, C., & Moon, Y. (2000). Machines and mindlessness: Social responses to computers. *Journal of Social Issues*, 56(1), 81–103. https://doi.org/10.1111/0022-4537.00153

Ni, Y., & Jia, F. (2025). A scoping review of AI-driven digital interventions in mental health care: Mapping applications across screening, support, monitoring, prevention, and clinical education. *Healthcare*, 13(10), 1205. https://doi.org/10.3390/healthcare13101205

Ohanian, R. (1990). Construction and validation of a scale to measure celebrity endorsers' perceived expertise, trustworthiness, and attractiveness. *Journal of Advertising*, 19(3), 39–52. https://doi.org/10.1080/00913367.1990.10673191

Petty, R. E., & Cacioppo, J. T. (1986). *Communication and persuasion: Central and peripheral routes to attitude change*. Springer.

Sheeran, P., & Webb, T. L. (2016). The intention–behavior gap. *Social and Personality Psychology Compass*, 10(9), 503–518. https://doi.org/10.1111/spc3.12265

Song, S. W., & Shin, M. (2022). Uncanny valley effects on chatbot trust, purchase intention, and adoption. *International Journal of Human–Computer Interaction*, 38(7), 595–607. https://doi.org/10.1080/10447318.2022.2121038

Sun, X., Ma, R., Wei, S., Cesar, P., Bosch, J. A., & El Ali, A. (2025). Understanding trust toward human versus AI-generated health information through behavioral and physiological sensing. arXiv, 2512.12348.

Wang, Y., Wu, F., & Ni, B. (2026). The detachment of responsibility: How and when AI-powered digital avatars undermine behavioral intentions in cancer screening persuasion. *Frontiers in Communication*, 11, 1767975. https://doi.org/10.3389/fcomm.2026.1767975

World Health Organization. (2022). World mental health report: Transforming mental health for all. WHO Press.

Wortel, C., Vanwesenbeeck, I., & Tomas, F. (2024). Made with artificial intelligence: The effect of AI disclosures in Instagram advertisements on consumer attitudes. *Emerging Media*, 2(3). https://doi.org/10.1177/27523543241292096