Self-Disclosure of Mental Health via Deepfakes: Testing the Effects of Self-Deepfakes on Affective Resistance and Intentions to Seek Mental Health Support

Keywords: Synthetic media, deepfake, self-disclosure, mental health, affective resistance

Extended Abstract

The Impact of Work

The advancement of artificial intelligence (AI)-enabled deepfake technology, which digitally manipulates a person's appearance, becomes a double-edged sword. The ability of deepfake technology to create highly realistic content has raised concerns among scholars due to the heightened risk of deception and false information (e.g., Lee and Shin, 2022; Lee et al., 2023; Pantserev, 2020), yet this same realism holds promising potential for positive applications, such as in healthcare. For example, personalized visualizations through deepfake technology can motivate users to pursue physical improvements and enhance their performance in tailored training programs (Clarke et al., 2023). These applications underscore the potential of deepfake technology as an effective tool for addressing health-related challenges, a potential that can be explained through the self-referencing effect. According to this perspective, information becomes easier to process and more persuasive when it is personally relevant, as referred to the self-referencing effect (Burnkrant and Unnava, 1989; Rogers et al., 1977).

Despite the potential of deepfake technology to be used in healthcare applications, the current scholarship calls for more nuanced research on the use of deepfake technology, as its application has been noted to evoke feelings of eeriness or discomfort rooted in uncanny valley perceptions (Weisman and Peña, 2021). These unsettling emotions can trigger affective resistance, where negative emotional reactions—such as discomfort or distrust—lead individuals to disengage from the health message (Appel, 2022), which can diminish the effectiveness of health interventions. Against this backdrop, this study investigates the emotional barriers linked to uncanny valley effects in health messaging by various messengers, including deepfake versions of oneself (herein referred to as self-deepfakes), deepfakes of well-known celebrities, and virtual agents, for self-disclosure in mental health messages, frequently referred to as the "talking cure" (Corcoran, 2000). In the context of mental health, self-disclosure is essential for fostering self-efficacy and enhancing coping strategies, including promoting intentions to seek mental health support (Rüsch et al., 2011; Wu et al., 2022).

This study addresses the boundaries where individuals may emotionally resist mental health messages that incorporate self-disclosure, potentially limiting their effectiveness in fostering intentions to seek mental health support as a coping strategy. In light of existing research demonstrating the positive impact of celebrity and virtual agent endorsements in raising public awareness of mental health through online help-seeking interventions (Gronholm and Thornicroft, 2022; Kim, 2024), this lab-based experiment provides evidence on the potential effects of using deepfakes in such interventions, highlighting their personalized nature and potential downsides, in comparison to more traditional approaches to addressing self-disclosure about mental health.

Main Theoretical Contribution

Theoretically, the current study revisits traditional video self-modeling approaches, which leverage the self-referencing effect to enhance engagement, within the context of deepfake technology by integrating the self-referencing effect and the uncanny valley effect. This study reveals that the synthesized nature of self-representations in deepfakes introduces artificiality that triggers discomfort, thereby increasing resistance to mental health self-disclosure messages. This discomfort underscores a significant limitation of deepfake technology in sensitive contexts, as individuals—especially those with higher baseline levels of mental health who find greater relevance to the topic—may be reluctant to engage with messages that present uncanny or distorted self-representations. Our findings emphasize the importance of future research to systematically investigate the boundaries of self-referencing in AI-driven synthetic media, focusing on how the degree of resemblance influences perceptions of personal relevance, evokes emotional resistance, and varies across individual differences. Furthermore, as deepfake technology finds its way into the healthcare sector, practitioners must remain mindful that while it offers innovative possibilities, it may also stir emotional resistance.

Data and Methods

This lab-based study employed a one-way within-subjects experimental design with three conditions: self-deepfake video, celebrity deepfake video, and virtual agent video. The results of a prior power analysis conducted using G*Power, assuming a large effect size ($f = 0.4$), 80% power, and a two-sided 5% significance level for a within-subjects analysis, indicated a minimum required sample size of $N = 12$. Participants were recruited from a private university in South Korea and were required to complete both a pre-test (Wave 1) and a post-test (Wave 2) survey, with a one-week interval to allow AI experts to produce high-quality deepfake stimuli personalized for each participant. Given the resource-intensive method of creating personalized deepfakes for each participant, we relied on a higher-powered within-subjects design to allow for a smaller, more manageable sample size. Of the 24 participants who initially enrolled, 21 completed both phases of the study.

Taking an interdisciplinary methodological approach, authors with AI engineering backgrounds created the stimuli using deep learning algorithms. To produce the deepfake video, we utilized the open-source FaceFusion tool, which facilitates seamless facial exchanges within video content. For the celebrity video, we created a deepfake of a South Korean female celebrity (Song Hye-kyo) delivering a self-disclosure message about mental health. To achieve this, we generated additional audio using voice cloning with Speechify, which recreated the celebrity's voice for script reading through text-to-speech conversion. Diff2lip (Mukhopadhyay et al., 2024) was employed to synchronize lip movements with the voice-cloned audio. In the virtual agent video condition, participants viewed a video featuring a virtual human developed using DeepBrainAI, designed to closely resemble a real person. In this condition, as with the other groups, the virtual human discussed their own experiences with mental health issues, as well as their seeking out help from mental health professionals. The video duration was approximately 90 seconds. See Figure 1 for the screenshots of the stimuli.

Findings

The findings show that self-deepfakes elicited greater affective resistance than celebrity videos, reducing help-seeking intentions, while no significant differences emerged between self-deepfakes and virtual agents. Furthermore, analysis showed that participants with poorer baseline mental health demonstrated significantly greater affective resistance to self-disclosing videos featuring deepfaked versions of themselves, whereas no such difference was observed in responses to virtual agent videos, regardless of mental health levels. These findings suggest that using deepfakes of oneself in personal health self-disclosure messaging may provoke resistance, warranting a more cautious and thoughtful approach is recommended when applying deepfake technology for health-related purposes.

References

Appel M (2022) Affective resistance to narrative persuasion. *Journal of Business Research* 149: 850–859. https://doi.org/10.1016/j.jbusres.2022.05.001

Burnkrant RE and Unnava HR (1989) Self-referencing: A strategy for increasing processing of message content. *Personality and Social Psychology Bulletin* 15(4): 628–638. https://doi.org/10.1177/0146167289154015

Clarke C, Xu J, Zhu Y, Dharamshi K, McGill H, Black S and Lutteroth C (2023) FakeForward: Using deepfake technology for feedforward learning. In: *Proceedings of the 2023 CHI Conference on Human Factors in Computing Systems*. Hamburg, DE, 23–28 April 2023, pp.1–17. New York: Association for Computing Machinery.

Corcoran P (2000) Therapeutic self-disclosure: "The talking cure" and "silence." In: Corcoran P and Spencer V (eds) *Disclosures*. London: Ashgate Publishing, pp.118–148.

Gronholm PC and Thornicroft G (2022) Impact of celebrity disclosure on mental health-related stigma. *Epidemiology and Psychiatric Sciences 31*: e62. https://doi.org/10.1017/S2045796022000488

Kim Y (2024) AI health counselor's self-disclosure: Its impact on user responses based on individual and cultural variation. *International Journal of Human–Computer Interaction*: 1–14. https://doi.org/10.1080/10447318.2024.2316373

Lee J, Hameleers M and Shin SY (2024) The emotional effects of multimodal disinformation: How multimodality, issue relevance, and anxiety affect misperceptions about the flu vaccine. *New Media and Society* 26(12): 6838–6860. https://doi.org/10.1177/14614448231153959

Lee J and Shin SY (2022) Something that they never said: Multimodal disinformation and source vividness in understanding the power of AI-enabled deepfake news. *Media Psychology 25*(4): 531–546. https://doi.org/10.1080/15213269.2021.2007489

Mukhopadhyay S, Suri S, Gadde RT and Shrivastava A (2024) 'Diff2lip: Audio conditioned diffusion models for lip-synchronization', *Proceedings of the IEEE/CVF Winter Conference on Applications of Computer Vision*: 5292–5302.

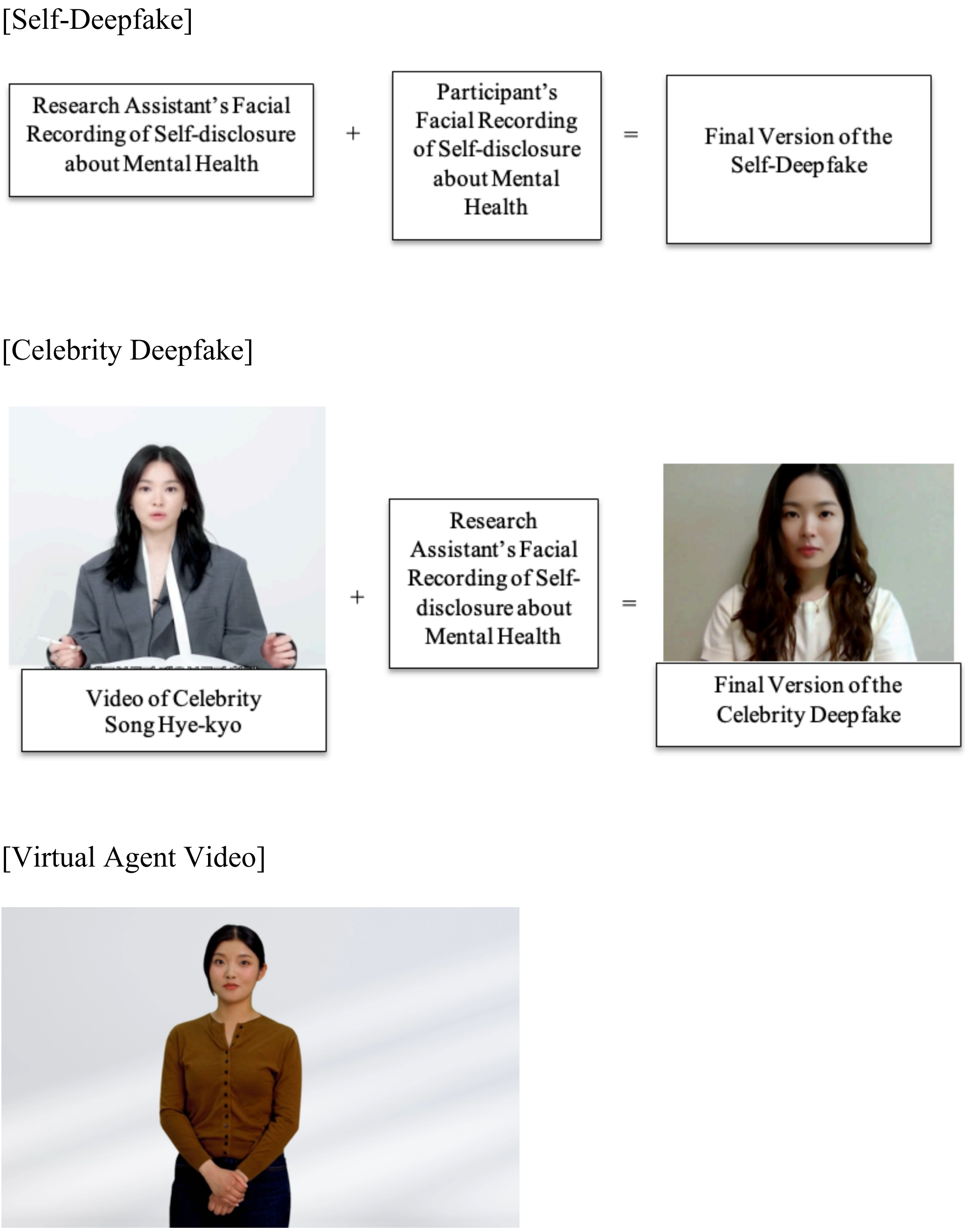
Pantserev KA (2020) The malicious use of AI-based deepfake technology as the new threat to psychological security and political stability. *Cyber Defence in the Age of AI, Smart Societies and Augmented Humanity*: 37–55. https://doi.org/10.1007/978-3-030-35746-7_3

Rogers TB, Kuiper NA and Kirker WS (1977) Self-reference and the encoding of personal information. *Journal of Personality and Social Psychology 35*(9): 677–688. https://doi.org/10.1037/0022-3514.35.9.677

Rüsch N, Evans-Lacko SE, Henderson C, Flach C and Thornicroft G (2011) Knowledge and attitudes as predictors of intentions to seek help for and disclose a mental illness. *Psychiatric Services 62*(6): 675–678. https://doi.org/10.1176/ps.62.6.pss6206_0675

Weisman WD and Peña JF (2021) Face the uncanny: The effects of doppelganger talking head avatars on affect-based trust toward artificial intelligence technology are mediated by uncanny valley perceptions. *Cyberpsychology, Behavior, and Social Networking* 24(3): 182–187. https://doi.org/10.1089/cyber.2020.0175

Wu Y, Shao J, Zhang D, Wang Y, Wang S, Wang Z, ... and Gu J (2022) Pathways from self-disclosure to medical coping strategy among adolescents with moderate and major depression during the COVID-19 pandemic: A mediation of self-efficacy. *Frontiers in Psychiatry 13,* 976386. https://doi.org/10.3389/fpsyt.2022.976386



[Self-Deepfake]

Research Assistant's Facial Recording of Self-disclosure about Mental Health
+
Participant's Facial Recording of Self-disclosure about Mental Health
=
Final Version of the Self-Deepfake

[Celebrity Deepfake]

Video of Celebrity Song Hye-kyo
+
Research Assistant's Facial Recording of Self-disclosure about Mental Health
=
Final Version of the Celebrity Deepfake

[Virtual Agent Video]

Figure 1. Three types of video representations of self-disclosure about mental health