When the brain meets AI: neural markers of face assessment

Evy van Weelden[1], Merel Hoskens[1], Emma Schippers[1], Pien Pieterse[1], Gijs Bekkers[1], Alfred Norén[1]

[1]Royal Netherlands Aerospace Centre, Amsterdam, the Netherlands

Synopsis

This study investigates neural activity associated with truth evaluations of real and AI-generated images using EEG and machine learning. Thirty-three participants viewed real and AI-generated faces while their brain activity was recorded and behavioral responses were collected. By exploring the neural correlates of truth evaluations, this research seeks to provide insights into how AI-generated disinformation affects people's beliefs and decision-making, and ultimately inform strategies for mitigating its impact.

Background

The rise of "slopaganda", a form of (false) information generated with the use of artificial intelligence (AI) that spreads online at an uncontrollable rate to influence public opinion, poses a threat to public safety (Klincewicz et al., 2025). Slopaganda exploits cognitive biases and can have long-lasting effects on neural representations, influencing beliefs and reasoning, highlighting the need for preventative measures on a neural level (Klincewicz et al., 2025). Fake images are a crucial component of "slopaganda", enabling malicious actors to create convincing and misleading visual content that can be used to manipulate public opinion and influence decision-making. For example, in 2022, a deepfake of Ukrainian president Volodymyr Zelensky surrendering to Russian forces was circulated on social media, sparking widespread concern and misinformation (Wakefield, 2022).

Previous studies have used neuroimaging to study effects of misinformation on true/false decisions (e.g., Moravec et al., 2019), but the impact of AI-generated images and videos on truth assessments and brain activity remains unknown. Faces elicit a recognizable neural waveform, i.e., the N170, which is not exclusive to *real* faces (Vaitonytė et al., 2024). However, differences in face perception between real and computer-generated faces may arise in the Late Positive Potential (LPP) (Vaitonytė et al., 2024). The current study aims to investigate neural activity associated with truth and false evaluations of real and AI-generated faces using EEG and ML, addressing two research questions: 1) what neural activity derived from event-related potentials (ERPs) is associated with truth vs. false assessments, specifically for real vs. AI-generated images?, 2) can we use ML to detect when information is interpreted to be true vs. false based on ERP amplitudes and latencies? This study seeks to provide insights into the neural mechanisms underlying truth evaluations and contribute to the development of preventative measures against AI-generated disinformation.

Methods

Participants

A total of 33 volunteers participated in this study (23 males, 10 females, age: M = 36.88, SD = 13.34, range [22, 61]). Prior to each experiment, participants provided their informed consent. This study was evaluated and approved by the NLR ethics review board, ref: CMO2025013.

Materials

Stimuli. Real faces were collected from the Chicago Face Database (Ma et al., 2015). The AI-generated faces (realistic and unrealistic) were generated with the use of DeepAI, an online and publically available AI-based content generator (Deep AI, Inc., 2025). Prompts for creating the images were drafted to mimic the descriptions of the photographs in the Chicago Face Database. All faces were cut into an oval, to minimize variations in hair, ears and background. Fifty images (25 male, 25 female) were collected for each category of faces, i.e., real, realistic AI-generated, and unrealistic AI-generated, resulting in a total of 150 images. Figure 1 provides examples of each category.

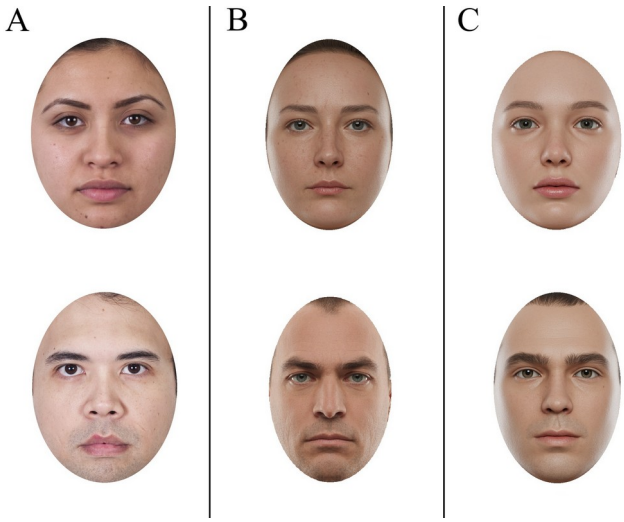

Figure 1. Examples of face categories: A) real faces, B) realistic AI-generated faces, C) unrealistic AI-generated faces (doll-like, with exaggerated smoothness and symmetry).

Questionnaires. The New Media Literacy Scale (Koc & Barut, 2016) was used to inquire about the participants' media literacy. A 5-point Likert scale question was added to inquire about the participants' familiarity with online AI-based tools for content generation.

EEG. The Unicorn BCI Core-8 (g.tec medical engineering GmbH, Austria) was used to collect EEG signals from participants at a 250 Hz sampling rate. Channels were placed according to the 10-20 system, and included: Fz, C3, Cz, C4, Pz, PO7, POz, PO8.

Procedure

All participants started with the questionnaires inquiring their demographics, media literacy and familiarity with AI-based online tools. Participants were then seated approximately 50 cm from a 24 inch monitor (1920 x 1080) with a 60 Hz refreshing rate. The experiment was presented on the monitor with the use of PsychoPy (version 2025.1.1) (Peirce et al., 2019). Each trial consisted of a fixation cross (700-2000 ms), followed by a face stimulus (1200 ms) on a white background. Immediately after the stimulus presentation, participants classified the face as real or AI-generated via keyboard and then rated their confidence in this decision (0–100%) using a mouse-controlled slider. Each participant performed a training phase of six trials before continuing with the actual experiment, consisting of 150 trials (one trial per face stimulus).

Results

This study is work-in-progress. Figure 2 shows averaged ERPs per condition for channels PO7, PO8, Cz and Pz, using cleaned EEG data. Based on the figure, we cannot identify differences between conditions at the N170 (negative peak at 140-200ms post-stimulus). However, there

seems to be a difference at the LPP (positive peak at 400-600ms post-stimulus). Statistical analyses need to be performed in order to confirm this observation.

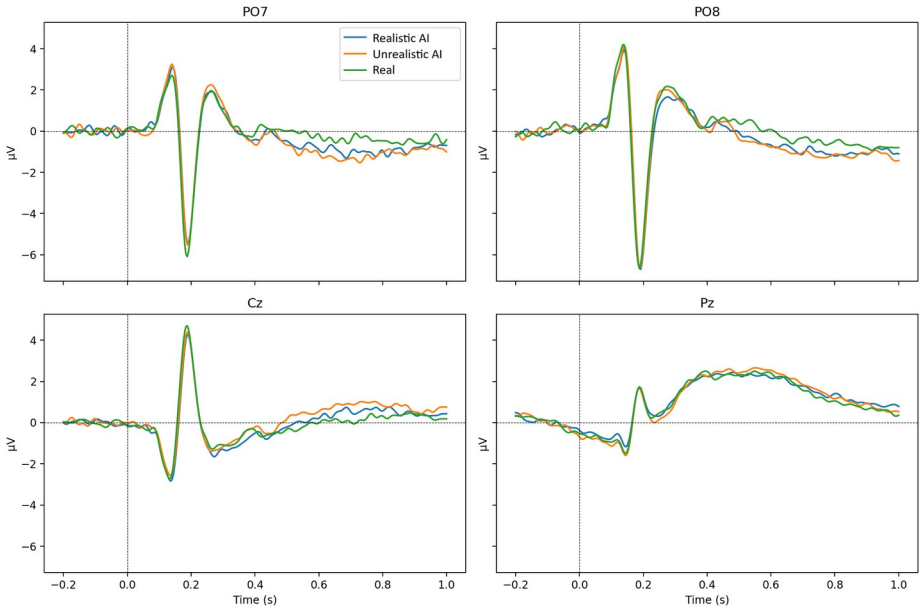

Figure 2. Averaged ERPs of all participants per condition for PO7, PO8, Cz and Pz.

Discussion

Our initial observations of the ERPs align with those in Vaitonytė et al. (2024). In further analyses, we will study differences in ERPs based on subjective responses and use supervised ML for feature analysis and classification of the ERP data. The findings of this study will provide foundational insights into the relationship between neural activity and decision-making, which could be extended to other types of information (text, audio) in a (simulated) social media feed. With this knowledge, we can begin to build a shield against AI-generated disinformation.

References

Deep AI, Inc. (2025). DeepAI [AI Image Generator]. Available at: https://deepai.org/machine-learning-model/text2img

Klincewicz, M., Alfano, M., & Fard, A. E. (2025). Slopaganda: The interaction between propaganda and generative AI. *arXiv preprint arXiv:2503.01560*.

Koc, M., & Barut, E. (2016). Development and validation of New Media Literacy Scale (NMLS) for university students. *Comput Hum Behav*, 63:834-843. doi: 10.1016/j.chb.2016.06.035

Ma, D. S., Correll, J., & Wittenbrink, B. (2015). The chicago face database: A free stimulus set of faces and norming data. *Behav Res Methods*, 47:1122–1135. doi: 10.3758/s13428-014-0532-5

Moravec, P. L., Minas, R. K., & Dennis, A. R. (2019). Fake News on Social Media. *MIS quarterly*, 43(4):1343-1360. doi: 10.25300/MISQ/2019/15505

Peirce, J. W., Gray, J. R., Simpson, S., MacAskill, M. R., Höchenberger, R., Sogo, H., Kastman, E., Lindeløv, J. (2019). PsychoPy2: experiments in behavior made easy. *Behav Res Methods.* doi: 10.3758/s13428-018-01193-y

Vaitonytė, J., Alimardani, M., & Louwerse, M. (2024). Face Processing in Real and Virtual Faces: An EEG Study. In *Proceedings of the Annual Meeting of the Cognitive Science Society*, Vol. 46. Retrieved from: https://escholarship.org/uc/item/6rj20960

Wakefield, J. (2022, March 18). *Deepfake presidents used in Russia-Ukraine war*. BBC News. https://www.bbc.com/news/technology-60780142