Embracing Deepfakes and AI-generated images in Neuroscience Research

Casey Becker[1]

Robin Laycock[1]

[1]RMIT University, Melbourne, Australia

Keywords: artificial neural networks, research methods, dynamic stimuli, vision, perception

Embracing Deepfakes and AI-generated images in Neuroscience Research

In 2017, a revolutionary type of video went viral: the 'deepfake.' This technology convincingly replicates a person's likeness, making it indistinguishable from a video recording. Deepfake algorithms are modelled off the human brain, and "learn" by continuously assessing their ability to create new versions of the face against original photographs of a person. With enough time and enough source photos, the likeness can be made to do anything – or, with post-production lip-syncing, say anything (Korshunova et al., 2016). Unlike costly visual effects technology used in Hollywood films, deepfake technology is open-source and highly accessible. Convincing fake media could now be generated by individuals, at home, on their personal computers (Zucconi, 2018).

The potential for societal harm soon became clear, as many individuals rushed to circulate their home-made deepfake pornography (Cole, 2017). The likeness of celebrities, public figures, and member of the public were shared without consent, often with life-ruining effects (Santana, 2022). Deepfakes went viral again in 2018, when media circulated a video that apparently showed former U.S. president Barack Obama saying that then President Donald Trump is a "total dipshit" (BuzzFeedVideo, 2018). The deepfake served as a public service announcement against the dangers of manipulated media, highlighting its potential to influence public opinion (Silverman, 2018). Indeed, research has shown that political deepfakes can increase negative views of a politician (Dobber et al., 2021), even when the viewer recognises the media is faked (Vaccari & Chadwick, 2020). Other research shows that deepfakes might contribute to distorting memories (Murphy & Flynn, 2021), fooling even those with high cognitive ability (Ahmed, 2021). To counteract these negative consequences, entire fields have been created that focus on automatic deepfake detection methods (Rana et al., 2022).

Despite the valid concerns surrounding the misuse of deepfakes, there is an emerging discussion of their constructive uses (Lin & Parvataneni, 2021; Mahmud & Sharmin, 2021). For example, deepfakes have been used to recreate celebrity's humanitarian messages in multiple languages (Die, 2019), create interactive art and museum installations (Mihailova, 2021; Wynn et al., 2021), generate hyper-realistic videogame characters of actors or players (Vejay et al., 2022), and change the age of actors in films (Loock, 2021). Importantly, some researchers have recognised the potential in deepfakes to improve our understanding of social perception: Deepfakes offer accessible, realistic, and customisable dynamic face stimuli (Barabanschikov & Marinova, 2021; Dobs et al., 2018; Haut et al., 2021). For example, Vijay et al. (2021) used deepfake technology to manipulate the presence of eye-contact, smiling and nodding, thus isolating their impact on observers' perceptions. Barabanschikov and Marinova (2021) created dynamic face illusions using deepfakes, including the Thatcher effect (an illusion where features like the mouth and eyes are inverted, making the face appear grotesque when upside down but normal when viewed right-side up) and dynamic chimeras (illusory stimuli created by combining different facial features or expressions from multiple individuals). Deepfakes have also been used to manipulate race (Haut et al., 2021) and physical attractiveness (Eberl et al., 2022) without disrupting the dynamic features of speaker or facial expression. Previously, such manipulations were only possible using static stimuli or 3D models. Using dynamic stimuli is important, as research increasingly identifies that dynamic face perception is distinct from static (Krumhuber et al., 2023; Pitcher et al., 2011; Pitcher et al., 2014). Manipulating dynamic stimuli realistically is also important, as face realism (Mustafa et al., 2017; Urgen et al., 2018) and realistic facial motion (Skiba & Vuilleumier, 2020) have been shown to elicit distinct neural responses. Researchers have demonstrated deepfakes' ability to transfer body motions, transforming an inexperienced grad student into a stunning ballerina performer (Chan et al., 2019).

AI's capacity to generate entirely new static images has also attracted the attention of the public. Recently, text-to-image models like Midjourney (Midjourney Inc., 2022) and DALL-E 2 (Marcus et al., 2022) have advanced to the point where they can produce hyper-realistic images from simple sentences, enabling individuals without coding expertise to create visually impressive content. Users have eagerly produced everything from highly realistic artwork to creative reinterpretations of fictional historical events. Researchers have also harnessed this power, increasingly incorporating AI-generated images as stimuli in visual neuroscience research. For example, Yang et al. (2021) demonstrated how AI can create scenes with specific parameters, such as room layout, objects, and clutter. Other researchers have synthesized abstract images, termed "nightmare fuel" (Fan, 2019), which activate primate brain regions beyond their typical maximal activation levels (Bashivan et al., 2019; Ponce et al., 2019), and help to uncover the neural mechanisms underlying visual perception. Like deepfakes, AI-generated images can be easily manipulated for various experimental purposes, providing exciting opportunities to create custom visual stimuli designed for specific experiments.

In clinical neuroscience, AI can be used to synthesise neuroimaging scans (Jeong et al., 2022; Laino et al., 2022; Qu et al., 2021; Sorin et al., 2020; Wang et al., 2023; Yi et al., 2019), which can improve AI classification of neurological phenomena where examples are rare, improving diagnoses (Sims, 2022) and enhancing our understanding of brain diseases and function (Wang et al., 2023). AI can be used to convert different imaging modalities like MRI to CT (Kearney et al., 2020), helpful for diagnoses, and also comparing across different types of research. Emerging research has demonstrated the ability for AI to read minds: AI can reconstruct key features of visual stimuli by studying participant fMRI (Huang et al., 2021; Wang et al., 2022) and EEG data (Singh et al., 2023). This innovative work deepens our understanding of cortical representations in natural vision and promotes advancements in brain-computer interface technology, potentially benefiting those with disabilities.

In the wake of the recent explosion of interest in chatGPT and the concomitant concerns raised about the implications for education and scientific writing (Hill-Yardin et al., 2023), some have made comparisons with the alarm generated from the introduction of calculators (Dickson, 2023). Similarly, Photoshop techniques initially raised concerns about the ability to distinguish real from fake (Farid, 2006), an issue that has improved its utility in vision research. As large language models (LLMs), text-to-image generation tools, and deepfake technology improves, researchers will have increased access to a wealth of easily generated visual stimuli, which holds great promise for advancing our understanding of perception. We argue that experimental psychologists and cognitive neuroscientists should stay updated on emerging tools, embracing the potential benefits these technologies bring to visual neuroscience.

Ahmed, S. (2021). Fooled by the fakes: Cognitive differences in perceived claim accuracy and sharing intention of non-political deepfakes. *Personality and Individual Differences*, *182*, 111074. https://doi.org/10.1016/j.paid.2021.111074

Barabanschikov, V. A., & Marinova, M. M. (2021). Deepfake in Face Perception Research. *Experimental Psychology (Russia)*, *14*(1), 4-19. https://doi.org/10.17759/exppsy.2021000001

Bashivan, P., Kar, K., & DiCarlo, J. J. (2019). Neural population control via deep image synthesis. *Science*, *364*(6439), eaav9436. https://doi.org/10.1126/science.aav9436

BuzzFeedVideo. (2018). *You Won't Believe What Obama Says In This Video!* ?

Chan, C., Ginosar, S., Zhou, T., & Efros, A. A. (2019). Everybody dance now. Proceedings of the IEEE/CVF international conference on computer vision, Seoul, Korea. https://doi.org/10.48550/arXiv.1808.07371

Cole, S. (2017). AI-Assisted Fake Porn Is Here and We're All Fucked. *Vice*. https://www.vice.com/en/article/gydydm/gal-gadot-fake-ai-porn

Dickson, B. (2023). Will ChatGPT get your job? These jobs can be threatened by artificial intelligence. https://socialbites.ca/latest-news/217040.html

Die, M. M. (2019). David Beckham speaks nine languages to launch malaria must die voice petition. *You Tube. Retrieved June*, *9*, 2022.

Dobber, T., Metoui, N., Trilling, D., Helberger, N., & de Vreese, C. (2021). Do (microtargeted) deepfakes have real effects on political attitudes? *The International Journal of Press/Politics*, *26*(1), 69-91. https://doi.org/10.1177/1940161220944364

Dobs, K., Bulthoff, I., & Schultz, J. (2018). Use and Usefulness of Dynamic Face Stimuli for Face Perception Studies-a Review of Behavioral Findings and Methodology. *Front Psychol*, *9*, 1355. https://doi.org/10.3389/fpsyg.2018.01355

Eberl, A., Kuhn, J., & Wolbring, T. (2022). Using deepfakes for experiments in the social sciences - A pilot study. *Front Sociol*, *7*, 907199. https://doi.org/10.3389/fsoc.2022.907199

Fan, S. (2019). How Researchers Used AI to Better Understand Biological Vision.

Farid, H. (2006). Digital doctoring: how to tell the real from the fake. *Significance*, *3*(4), 162-166. https://doi.org/10.1111/j.1740-9713.2006.00197.x

Haut, K., Wohn, C., Antony, V., Goldfarb, A., Welsh, M., Sumanthiran, D., Jang, J.-z., Rafayet Ali, M., & Hoque, E. (2021). Could you become more credible by being White? Assessing Impact of Race on Credibility with Deepfakes. *arXiv e-prints*, arXiv: 2102.08054. https://doi.org/10.48550/arXiv.2102.08054

Hill-Yardin, E. L., Hutchinson, M. R., Laycock, R., & Spencer, S. J. (2023). A Chat (GPT) about the future of scientific publishing. *Brain, behavior, and immunity*, S0889-1591 (0823) 00053-00053. https://doi.org/10.1016/j.bbi.2023.02.022

Huang, W., Yan, H., Wang, C., Yang, X., Li, J., Zuo, Z., Zhang, J., & Chen, H. (2021). Deep Natural Image Reconstruction from Human Brain Activity Based on Conditional Progressively Growing Generative Adversarial Networks. *Neurosci Bull*, *37*(3), 369-379. https://doi.org/10.1007/s12264-020-00613-4

Jeong, J. J., Tariq, A., Adejumo, T., Trivedi, H., Gichoya, J. W., & Banerjee, I. (2022). Systematic Review of Generative Adversarial Networks (GANs) for Medical Image Classification and Segmentation. *J Digit Imaging*, *35*(2), 137-152. https://doi.org/10.1007/s10278-021-00556-w

Kearney, V., Ziemer, B. P., Perry, A., Wang, T., Chan, J. W., Ma, L., Morin, O., Yom, S. S., & Solberg, T. D. (2020). Attention-Aware Discrimination for MR-to-CT Image Translation Using Cycle-Consistent Generative Adversarial Networks. *Radiol Artif Intell*, *2*(2), e190027. https://doi.org/10.1148/ryai.2020190027

Korshunova, I., Shi, W., Dambre, J., & Theis, L. (2016). Fast Face-Swap Using Convolutional Neural Networks. *2017 IEEE International Conference on Computer Vision (ICCV)*, 3697-3705. https://doi.org/10.1109/ICCV.2017.397

Krumhuber, E. G., Skora, L. I., Hill, H. C. H., & Lander, K. (2023). The role of facial movements in emotion recognition. *Nature Reviews Psychology*. https://doi.org/10.1038/s44159-023-00172-1

Laino, M. E., Cancian, P., Politi, L. S., Della Porta, M. G., Saba, L., & Savevski, V. (2022). Generative Adversarial Networks in Brain Imaging: A Narrative Review. *J Imaging*, *8*(4), 83. https://doi.org/10.3390/jimaging8040083

Lin, Y., & Parvataneni, K. (2021). Deepfake Generation, Detection, and Use Cases: A Review Paper. *International Journal of Computational and Biological Intelligent Systems*, *3*(2).

Loock, K. (2021). On the realist aesthetics of digital de-aging in contemporary Hollywood cinema. *Orbis Litterarum*, *76*(4), 214-225. https://doi.org/10.1111/oli.12302

Mahmud, B. U., & Sharmin, A. (2021). Deep insights of deepfake technology: A review. *arXiv preprint arXiv:2105.00192*. https://doi.org/10.48550/arXiv.2105.00192

Marcus, G., Davis, E., & Aaronson, S. (2022). A very preliminary analysis of dall-e 2. *arXiv preprint arXiv:2204.13807*. https://doi.org/10.48550/arXiv.2204.13807

Midjourney Inc. (2022). *Midjourney*. In https://midjourney.com

Mihailova, M. (2021). To dally with Dalí: Deepfake (Inter) faces in the art museum. *Convergence*, *27*(4), 882-898. https://doi.org/10.1177/13548565211029401

Murphy, G., & Flynn, E. (2021). Deepfake false memories. *Memory*, 1-13. https://doi.org/10.1080/09658211.2021.1919715

Mustafa, M., Guthe, S., Tauscher, J.-P., Goesele, M., & Magnor, M. (2017). How Human Am I? EEG-based Evaluation of Virtual Characters. Proceedings of the 2017 CHI Conference on Human Factors in Computing Systems, https://doi.org/10.1145/3025453.3026043

Pitcher, D., Dilks, D. D., Saxe, R. R., Triantafyllou, C., & Kanwisher, N. (2011). Differential selectivity for dynamic versus static information in face-selective cortical regions. *Neuroimage*, *56*(4), 2356-2363. https://doi.org/10.1016/j.neuroimage.2011.03.067

Pitcher, D., Duchaine, B., & Walsh, V. (2014). Combined TMS and FMRI reveal dissociable cortical pathways for dynamic and static face perception. *Curr Biol*, *24*(17), 2066-2070. https://doi.org/10.1016/j.cub.2014.07.060

Ponce, C. R., Xiao, W., Schade, P. F., Hartmann, T. S., Kreiman, G., & Livingstone, M. S. (2019). Evolving Images for Visual Neurons Using a Deep Generative Network Reveals Coding Principles and Neuronal Preferences. *Cell*, *177*(4), 999-1009 e1010. https://doi.org/10.1016/j.cell.2019.04.005

Qu, C., Zou, Y., Dai, Q., Ma, Y., He, J., Liu, Q., Kuang, W., Jia, Z., Chen, T., & Gong, Q. (2021). Advancing diagnostic performance and clinical applicability of deep learning-driven generative adversarial networks for Alzheimer's disease. *Psychoradiology*, *1*(4), 225-248. https://doi.org/10.1093/psyrad/kkab017

Rana, M. S., Nobi, M. N., Murali, B., & Sung, A. H. (2022). Deepfake Detection: A Systematic Literature Review. *IEEE Access*. https://doi.org/10.1109/ACCESS.2022.3154404

Santana, M. S. (2022). *Justice for Women: Deep fakes and Revenge Porn* Global Conference on Women's Studies, Rotterdam, The Netherlands.

Silverman, C. (2018). How To Spot A Deepfake Like The Barack Obama–Jordan Peele Video. https://www.buzzfeed.com/craigsilverman/obama-jordan-peele-deepfake-video-debunk-buzzfeed

Sims, C. (2022). Highly Accurate FMRI ADHD Classification using time distributed multi modal 3D CNNs. *arXiv preprint arXiv:2205.11993*. https://doi.org/10.48550/arXiv.2205.11993

Singh, P., Pandey, P., Miyapuram, K., & Raman, S. (2023). EEG2IMAGE: Image Reconstruction from EEG Brain Signals. *arXiv preprint arXiv:2302.10121*. https://doi.org/10.48550/arXiv.2302.10121

Skiba, R. M., & Vuilleumier, P. (2020). Brain Networks Processing Temporal Information in Dynamic Facial Expressions. *Cereb Cortex*, *30*(11), 6021-6038. https://doi.org/10.1093/cercor/bhaa176

Sorin, V., Barash, Y., Konen, E., & Klang, E. (2020). Creating artificial images for radiology applications using generative adversarial networks (GANs)–a systematic review. *Academic radiology*, *27*(8), 1175-1185. https://doi.org/10.1016/j.acra.2019.12.024

Urgen, B. A., Kutas, M., & Saygin, A. P. (2018). Uncanny valley as a window into predictive processing in the social brain. *Neuropsychologia*, *114*, 181-185. https://doi.org/10.1016/j.neuropsychologia.2018.04.027

Vaccari, C., & Chadwick, A. (2020). Deepfakes and disinformation: Exploring the impact of synthetic political video on deception, uncertainty, and trust in news. *Social Media+ Society*, *6*(1), 2056305120903408. https://doi.org/10.1177/2056305120903408

Vejay, L., Adine, M., & Zach, H. (2022). Artificial intelligence: deepfakes in the entertainment industry. https://www.wipo.int/wipo_magazine/en/2022/02/article_0003.html

Vijay, R. S., Shubham, K., Renier, L. A., Kleinlogel, E. P., Mast, M. S., & Jayagopi, D. B. (2021). An Opportunity to Investigate the Role of Specific Nonverbal Cues and First Impression in Interviews using Deepfake Based Controlled Video Generation. Companion Publication of the 2021 International Conference on Multimodal Interaction, https://doi.org/10.1145/3461615.3485397

Wang, C., Yan, H., Huang, W., Li, J., Wang, Y., Fan, Y. S., Sheng, W., Liu, T., Li, R., & Chen, H. (2022). Reconstructing rapid natural vision with fMRI-conditional video generative adversarial network. *Cereb Cortex*, *32*(20), 4502-4511. https://doi.org/10.1093/cercor/bhab498

Wang, R., Bashyam, V., Yang, Z., Yu, F., Tassopoulou, V., Chintapalli, S. S., Skampardoni, I., Sreepada, L. P., Sahoo, D., Nikita, K., Abdulkadir, A., Wen, J., & Davatzikos, C. (2023). Applications of generative adversarial networks in neuroimaging and clinical neuroscience. *Neuroimage*, *269*, 119898. https://doi.org/10.1016/j.neuroimage.2023.119898

Wynn, N., Johnsen, K., & Gonzalez, N. (2021). Deepfake Portraits in Augmented Reality for Museum Exhibits. 2021 IEEE International Symposium on Mixed and Augmented Reality Adjunct (ISMAR-Adjunct), https://doi.org/10.1109/ISMAR-Adjunct54149.2021.00125

Yang, C., Shen, Y., & Zhou, B. (2021). Semantic hierarchy emerges in deep generative representations for scene synthesis. *International Journal of Computer Vision*, *129*, 1451-1466. https://doi.org/10.1007/s11263-020-01429-5

Yi, X., Walia, E., & Babyn, P. (2019). Generative adversarial network in medical imaging: A review. *Med Image Anal*, *58*, 101552. https://doi.org/10.1016/j.media.2019.101552

Zucconi, A. (2018). How To Create The Perfect DeepFakes. *Tutorial*. https://www.alanzucconi.com/2018/03/14/create-perfect-deepfakes/