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 rely on deep learning, a type of machine learning that uses artificial neural networks (ANNs), which are inspired by the human brain's structure and function. Like the brain, an ANN is made up of layers of "neurons" with higher levels performing more complex tasks. Artificial neurons, much like their biological namesake, adjust their connection strength to neighbouring neurons, enabling the ANN to learn complex patterns from data inputs. Like an artist learning to paint a face, an ANN's first attempts are rudimentary, and it "learns" to portray a realistic face by continuously comparing its work to original source photos. With enough practice, and enough training material, a person's likeness can be made to do anything – or, with post-production lip-syncing, say anything (Korshunova et al., 2016). As we will discuss in the following, the ability for ANNs to replicate real world imagery – including but not limited to faces – may be as exciting as it is concerning. We will argue that this technology presents novel opportunities for generating diverse stimuli, and advancing our understanding of visual systems.

Unlike costly visual effects technology used in Hollywood films, deepfake technology is open-source and highly accessible. For the first time, convincing fake videos could be generated by individuals, at home, on their personal computers (Zucconi, 2018). The potential for societal harm became clear, as many individuals rushed to circulate their home-made deepfake pornography (Cole, 2017). The likeness of celebrities, public figures, and members of the public were shared without consent, sometimes 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). To counteract these negative consequences, entire fields have been created that focus on automatic deepfake detection methods (Rana et al., 2022).

But perhaps we got off to the wrong start. The term "deepfake" now carries negative connotations, potentially causing hesitancy or scepticism when discussing its legitimate research applications. 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. It is also possible to transfer the dynamics of one person to another, which researchers have demonstrated by transforming an inexperienced grad student into a stunning ballerina performer (Chan et al., 2019). Using dynamic face 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). Before deepfakes, manipulation of facial dynamism was only possible using dynamic 3D models. These clearly artificial faces are not ideal, as face realism (Mustafa et al., 2017; Urgen et al., 2018) and realistic facial motion (Skiba & Vuilleumier, 2020) has been shown to elicit distinct neural responses.

Deepfake ANNs typically use autoencoders, designed to maintain temporal coherence in videos, and are trained on one identity at a time. Generative adversarial networks (GANs) employ two ANNs, with one dedicated to evaluating the quality of the other's generated content. Although GANs have limitations in producing temporally consistent videos, their refined feedback loop allows them to generate diverse and high-quality static content. When trained on thousands of different identities, GANs can learn the essential elements of human facial features. This has led to viral social media posts and news articles showcasing high-quality portraits of people who do not exist (e.g., Hill & White, 2020). In fact, GANs can be trained on any object category – numerous websites showcase "photographs" of people,

places, and even cats, that have never existed. In academia, some have highlighted the potential threat to research integrity – it may now be easier to modify or falsify scientific images (e.g., in cellular neuroscience) that can deceive the judgement of human experts (Gu et al., 2022).

Others have demonstrated the utility of GANs; like dynamic deepfakes, they can be easily manipulated for a variety of experimental purposes. Yang et al. (2021) demonstrated how AI can create scenes with specific parameters, such as room layout, objects, and clutter. Zhang et al. (2019) demonstrated the ability to believably modify parts of an image, such as changing a person's hair colour and style. Recent advancements in text-to-image technology like Midjourney (Midjourney Inc., 2022) and DALL-E 2 (Marcus et al., 2022) have made it easy for anyone to generate and modify photo-realistic images using simple sentences. This technology has not only been entertaining for users creating imaginative content, but may also be valuable for researchers, who can now effectively modulate model parameters for research purposes without extensive coding experience.

The growing sophistication of AI may offer a solution to a problem impacting psychological and neuroscience research: overreliance on limited stimulus sets, which impacts generalisability and ecological validity (Dawel et al., 2022; Grootswagers & Robinson, 2021). Researchers can now easily create and modify diverse image databases tailored to their research needs. This prompts the question – is AI-generated content ecologically valid? Behavioural research confirms that it is difficult for humans to detect high-quality AI-generated images, suggesting they accurately portray real people, places, and things (e.g., Bray et al., 2022; Korshunov & Marcel, 2020; Lu et al., 2023; Shen et al., 2021). The ultimate test might be whether real and AI-generated faces elicit similar neural responses, given human expertise in face perception (Haxby et al., 2000). So far, research has shown that face-swap deepfakes elicit similar neural responses to real videos, unless we know one of the people being portrayed (Tauscher et al., 2021). Another study revealed that while neural responses were largely similar, consciously undetectable fake faces can still elicit slightly different neural responses compared to photos (Moshel et al., 2022). Even this minor difference may not last long, however – as the authors note, recent advancements in AI-generated images challenge human perception to a greater extent.

The utility of GANs extends beyond their ability to mimic photographs of reality. Neural responses in the macaque visual cortex can be used to guide the evolution of GANs, allowing them to synthesise "super stimuli" that activate neurons beyond their typical maximal activation levels (Bashivan et al., 2019; Ponce et al., 2019). These images are being used to study the visual system, as they vary depending on the neurons that guide them.

Macaques find these images very interesting (Rose et al., 2021), as do humans, who have described them as "nightmare fuel" (Fan, 2019). In humans, similar GANs can be trained on images that are accompanied by neuroimaging data elicited by those images (e.g., EEG; Singh et al., 2023; fMRI Huang et al., 2021). This allows the ANN to learn which features of the neural data correspond with which features of a stimulus, until it can reconstruct key visual elements from neural data alone with surprising accuracy. This innovative work advances our understanding of cortical representations in natural vision, and promotes the development of advanced brain-computer interfaces (BCIs). Such enhancements could benefit people with disabilities by providing improved communication aids, intuitive prosthetic control, and advancements in visual prosthetics.

Rapid advancements in technology often raise concerns about their impact on society and science. The introduction of photoshop provoked alarm over scientific fraud (Richardson et al., 1995) – an issue that is still being managed today (Bik, 2022). However, as researchers adapt to these new technologies, their potential benefits become more apparent; the ability for photoshopped images to be realistic has improved their utility in vision science, cognitive neuroscience, and psychological research. We believe the same is true for AI-generated content, which offer a wealth of easily generated visual stimuli, and hold great promise for advancing our understanding of visual perception and brain function. By embracing these advancements and addressing the challenges they pose, AI-generated stimuli may allow us to learn more about the human brain and visual system than was previously possible.

References

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 [research-article]. *Science*, *364*(6439). https://doi.org/10.1126/science.aav9436

Bik, E. (2022). Science has a nasty photoshopping problem. *The New York Times*. https://www.nytimes.com/interactive/2022/10/29/opinion/science-fraud-image-manipulation-photoshop.html

Bray, S. D., Johnson, S. D., & Kleinberg, B. (2022). Testing Human Ability To Detect Deepfake Images of Human Faces. *arXiv preprint arXiv:2212.05056*.

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

Dawel, A., Miller, E. J., Horsburgh, A., & Ford, P. (2022). A systematic survey of face stimuli used in psychological research 2000-2020. *Behav Res Methods*, *54*(4), 1889-1901. https://doi.org/10.3758/s13428-021-01705-3

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.

Grootswagers, T., & Robinson, A. K. (2021). Overfitting the Literature to One Set of Stimuli and Data [Perspective]. *Frontiers in human neuroscience*, *15*. https://doi.org/10.3389/fnhum.2021.682661

Gu, J., Wang, X., Li, C., Zhao, J., Fu, W., Liang, G., & Qiu, J. (2022). AI-enabled image fraud in scientific publications. *Patterns*, *3*(7), 100511. https://doi.org/https://doi.org/10.1016/j.patter.2022.100511

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

Haxby, J. V., Hoffman, E. A., & Gobbini, M. I. (2000). The distributed human neural system for face perception. *Trends in Cognitive Sciences*, *4*(6), 223-233. https://doi.org/https://doi.org/10.1016/S1364-6613(00)01482-0

Hill, K., & White, J. (2020). Do These A.I.-Created Fake People Look Real to You? *The New York Times*. https://www.nytimes.com/interactive/2020/11/21/science/artificial-intelligence-fake-people-faces.html

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

Korshunov, P., & Marcel, S. (2020). Deepfake detection: humans vs. machines. *arXiv preprint arXiv:2009.03155*.

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

Lin, Y., & Parvataneni, K. (2021). Deepfake Generation, Detection, and Use Cases: A Review Paper. *International Journal of Computational and Biological Intelligent Systems*, *3*(2). https://doi.org/Retrieved from https://ijcbis.org/index.php/ijcbis/article/view/1642

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

Lu, Z., Huang, D., Bai, L., Liu, X., Qu, J., & Ouyang, W. (2023). Seeing is not always believing: A Quantitative Study on Human Perception of AI-Generated Images. *arXiv e-prints*, arXiv: 2304.13023. https://doi.org/10.48550/arXiv.2304.13023

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

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

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

Richardson, M. L., Frank, M. S., & Stern, E. J. (1995). Digital image manipulation: what constitutes acceptable alteration of a radiologic image. https://doi.org/10.2214/ajr.164.1.7998545

Rose, O., Johnson, J., Wang, B., & Ponce, C. R. (2021). Visual prototypes in the ventral stream are attuned to complexity and gaze behavior. *Nature communications*, *12*(1), 6723. https://doi.org/10.1038/s41467-021-27027-8

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

Shen, B., RichardWebster, B., Toole, A. O., Bowyer, K., & Scheirer, W. J. (2021, 15-18 Dec. 2021). A Study of the Human Perception of Synthetic Faces. 2021 16th IEEE International Conference on Automatic Face and Gesture Recognition (FG 2021), https://doi.org/10.1109/FG52635.2021.9667066

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

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

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

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, T., Ren, P., Xie, X., & Zhang, L. (2021). Gan prior embedded network for blind face restoration in the wild. Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition,

Zhang, J., Li, A., Liu, Y., & Wang, M. (2019). Adversarially Regularized U-Net-based GANs for Facial Attribute Modification and Generation. *IEEE Access*, *7*, 86453-86462. https://doi.org/10.1109/ACCESS.2019.2926633

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