Evolving synthetic images for the control of affective experience

Darius Valevicius (darius.valevicius@umontreal.ca)
UdeM Neuroscience, 2960 chemin de la Tour
Montreal, Quebec H3T 1J4 Canada

Celine Haddad (celine.haddad@umontreal.ca)
UdeM Psychologie, 90 avenue Vincent d'Indy
Montreal, Quebec H2V 2S9 Canada

Vincent Taschereau-Dumouchel (vincent.taschereau-dumouchel@umontreal.ca)
UdeM Psychiatrie et addictologie, 2900 boul. Édouard-Montpetit
Montreal, Quebec H3T 1J4 Canada

Abstract:

Creating stimuli that reliably produce targeted alterations in subjective emotional experience would be a useful tool for understanding the neural correlates of conscious feelings. In this study, we develop a method to create images that maximize a target emotion (e.g. fear) by using a combination of generative artificial intelligence and an evolutionary algorithm. We show that image evolution conditioned on fear ratings can generate images that strongly drive fear responses, while conditioning image evolution on a physiological proxy of fear (i.e. electrodermal activity) does not create images rated as fearful. However, the latter still converges on generally feared insectoid categories. This method may be used in the context of decoded neurofeedback to study the contribution of different brain regions to subjective fear.

Keywords: fear; physiology; consciousness; artificial intelligence

Introduction

How does the brain construct the conscious experience of emotions? Historically, regions such as the amygdala have been considered central to the experience of emotions such as fear. However, it is possible that these regions are not directly responsible for the subjective sense of fear, but rather serve to coordinate related processes such as physiological arousal and sensorimotor responses to perceived threats (Taschereau-Dumouchel et al, 2022).

To understand the neural correlates subjective fear on the one hand, or physiological arousal on the other, it would be useful to have a method to drive these processes in a reliable way. One method of driving specific patterns of neural activity is closed-loop neurofeedback (Taschereau-Dumouchel et al, 2018), which can, for instance, use real-time functional magnetic resonance imaging (rt-fMRI) to drive a target pattern of brain activity. In these paradigms, fMRI data are analyzed as they are collected, and measures derived from these data are used to alter the experimental stimulus.

One recent example of a closed-loop paradigm uses generative AI image models to produce pictures that maximize a target brain measure (Ponce et al, 2019). By iteratively selecting, recombining, and mutating the latent vectors of the generated images based on neural measurements, synthetic images can be produced which maximize the desired pattern of neural activity. This paradigm has been successfully implemented using single-unit intracranial recordings in macaques to generate visual “prototypes” which maximally activate a target neuron in the anterior visual stream.

In this paper, we present a proof-of-concept study for producing affective images with human subjects. using self-reported and physiological measurements. We propose to apply this paradigm in the context of decoded neurofeedback to create images that maximize brain activity related to subjective fear, to see if the generated images are also perceived as fearful. We can also compare the relative efficacy of images produced using different ROI decoders, which could provide causal evidence of the contribution of these regions to subjective fear.

Methods

20 participants were recruited to a behavioral study where a series of AI-generated images were presented, following which participants were asked to make ratings of subjective fear on a 10-point continuous scale. Simultaneously, measurements of electrodermal activity (EDA) and skin conductance responses (SCRs) were collected using an EMOTIBIT device (Montgomery et al, 2023).

Images were evolved using an evolutionary algorithm (Hansen, 2016), where after every generation of eight images, the self-report or SCR measures were used to recombine their latent vectors and produce a new set of images. Ten generations were generated per condition, and 3 conditions were tested: Self-report, SCR, and a control condition where random fitness values were generated.

Images were generated using Stable UnCLIP (Ramesh et al, 2018). We found that, when attempting to evolve CLIP's 768-dimensional latent space vector, the generated images failed to converge on target categories. To constrain the problem space, we applied a principal components analysis (PCA) to the latent space embeddings of 2700 animal images. By isolating the first 80 dimensions and handling vectors in PCA-space before re-converting them to CLIP vectors, we were able to define a subspace of CLIP that exclusively represented animal categories and features. Using this subdomain, we were able to converge on target categories within about 6 generations.

Results

Repeated measures ANOVA revealed a significant interaction between condition and generation (F(29) = 26.77, p < 0.001) when predicting self-reported fear ratings (Figure 1). Post-hoc Tukey tests conducted on the last three generations showed a significant difference between the self-report and control conditions (MD = 1.67, p < 0.001) and the self-report and SCR conditions (MD = 1.72, p < 0.001).

The SCR-based evolution did not seem to increase fear ratings (MD = 0.05, p = 0.94). However, images in this condition mostly converge on insects, similar to the self-report condition, and in contrast to the control condition which tended to converge on mammals (Figure 2). This suggests an interesting dissociation between fear and SCR, where though the same animal categories are represented, SCR-evolved images are lacking qualities that would make them consciously frightening.

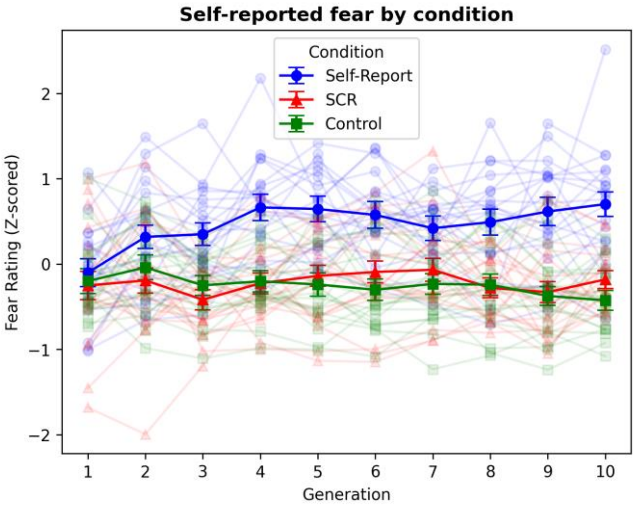

Figure 1: Fear ratings by generation for the three conditions. Error bars = 95% CI.

Conclusion

We introduce a paradigm for producing affective stimuli, which uses a combination of generative AI models and an evolutionary algorithm. Images evolved using self-reported fear as a fitness measure produced significantly more fear after evolution. We propose to apply this paradigm in the context of decoded neurofeedback, to see if fear predictions from different ROIs can be used to produce fear-inducing images.

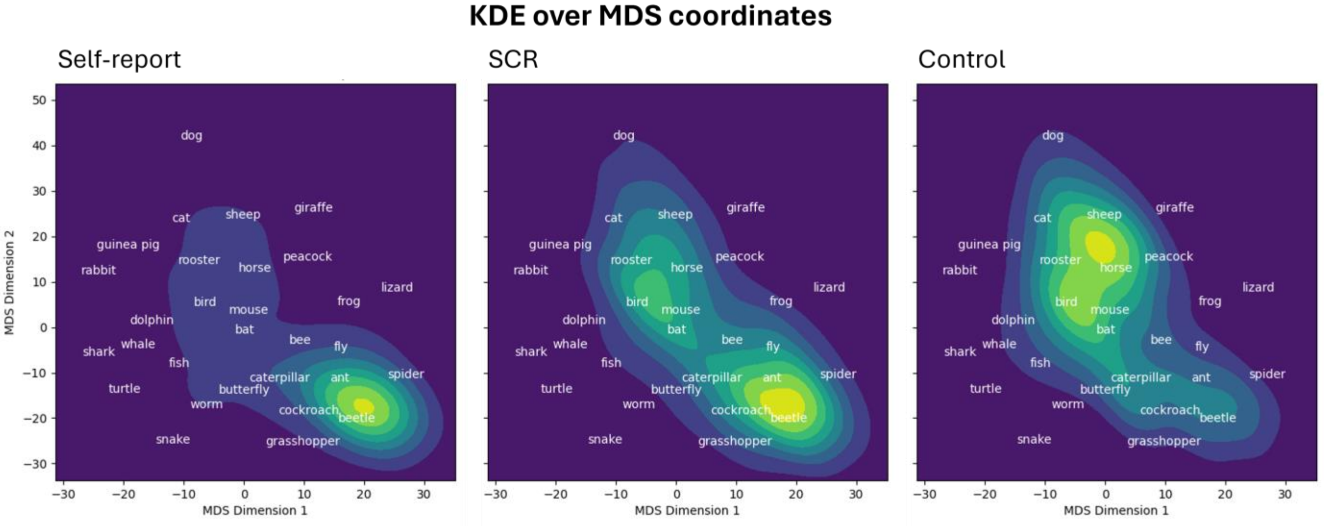

Figure 2: The average locations of end-of-evolution images in a multidimensional scaling space for the three conditions. For Self-report and SCR, images tend to converge on insectoid categories.

Acknowledgments

This research was funded by an NSERC Discovery Grant.

References

Hansen, N. (2016). The CMA Evolution Strategy: A Tutorial.

Montgomery, S. M., Nair, N., Chen, P., & Dikker, S. (2023). Introducing EmotiBit, an open-source multi-modal sensor for measuring research-grade physiological signals. Science Talks, 6, 100181.

Polák, J., Rádlová, S., Janovcová, M., Flegr, J., Landová, E., & Frynta, D. (2020). Scary and nasty beasts: Self-reported fear and disgust of common phobic animals. British Journal of Psychology, 111(2), 297–321.

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.e10.

Ramesh, A., Dhariwal, P., Nichol, A., Chu, C., & Chen, M. (2022). Hierarchical Text-Conditional Image Generation with CLIP Latents.

Taschereau-Dumouchel, V., Cortese, A., Chiba, T., Knotts, J. D., Kawato, M., & Lau, H. (2018). Towards an unconscious neural reinforcement intervention for common fears. Proceedings of the National Academy of Sciences, 115(13), 3470–3475.

Taschereau-Dumouchel, V., Kawato, M., & Lau, H. (2020). Multivoxel pattern analysis reveals dissociations between subjective fear and its physiological correlates. Molecular Psychiatry, 25(10), 2342–2354.