Using Generative AI as an Artistic Material: A Hacker's Guide

TERENCE BROAD, Creative Computing Institute, University of the Arts London, United Kingdom

Hacking is an approach to working with technology aimed at pushing it beyond it's intended design, use it in unconventional ways in order to get it to do things it is not supposed to do. This paper documents a number of artistic experiments that take a hacker's approach to subverting and intervening in the normal functioning of generative AI systems, treating generative AI as an artistic material. This paper argues this hacking ethos offers a critical approach to explainable AI in the arts.

Additional Key Words and Phrases: Generative AI, Creative Computing, Computational Arts, Explainable AI, Hacking

ACM Reference Format:
Terence Broad. 2024. Using Generative AI as an Artistic Material: A Hacker's Guide. In *Proceedings of Explainable AI for the Arts Workshop 2024 (XAIxArts 2024).* ACM, New York, NY, USA, 4 pages.

1 INTRODUCTION

An increasingly large amount of the media we consume is now being created using generative AI algorithms, with large parts of the internet now increasingly being populated with AI generated spam [13]. This media has been generated by deep neural networks that have being trained on (often very large) datasets of existing media (which is normally scraped from large parts of the internet). These algorithms are then used to regurgitate visual, textual and auditory data into 'new' forms, in a mechanised and mass-produced fashion [20]. This has brought forth a new, algorithmically complex and opaque form cultural production.

This paper will show how generative AI can itself be used as an artistic material, in ways that go beyond simply mashing up and regurgitating of existing training data. By adopting *hacking* (in it's original technical sense that emerged out of the MIT hacker culture of the 1960's and 70's), which refers to "exploring the limits of what is possible, in a spirit of playful cleverness" [22], artists have found many ways of furthering our understanding of the operations of generative AI. Through targeted interventions in the data used to condition and represent these networks, as well as interventions into the computational processes to train and sample from these networks, it is possible to expose the underlying processes underpinning these algorithms and develop new forms of algorithmic expression.

This paper will give examples of artistic projects (from myself and other artists) that take this kind of hacking approach towards generative AI to push these networks beyond their originally intended functioning and to make interventions to expose the workings of these underlying computational processes. These will be grouped into four categories of intervention: into the inputs of the network, to the learned parameters of a network, to the training of networks, and to the computational graph of a network in inference. This paper will show that generative AI can be used as an artistic material, and in doing so, can lead to new ways of understand and exposing the nature of the underlying algorithms, providing a critical approach to explainable AI (XAI) in the arts.

Author's Contact Information: Terence Broad, t.broad@arts.ac.uk, Creative Computing Institute, University of the Arts London, United Kingdom.

XAIxArts 2024, June 23, 2024, Chicago, IL, United States

2 SUBVERTING A NETWORKS INPUTS

One way of pushing generative networks outside their comfort zone, towards revealing aspects of their inner nature, is to feed in data samples it was not intended to handle. Feeding in data either at the limits of, or completely outside of, the expected ranges, is one way of subverting a networks generative processes in order to reveal aspects of its inner workings.

In *Introspections* [18], the artist Philipp Schmitt took off-the-shelf image translation models, designed to translate photographs into line drawings and vice-versa and fed into them blank images. At first, the images returned were themselves blank, but after the outputs were repeatedly fed back into the same model many times, detailed artefacts emerged, showing complex hallucinations from the model's internal operations.

The Algorithmic Resistance Research Group (ARRG!), which consists of the artists Eryk Salvaggio, Caroline Sinders, and Steph Maj Swanson, produce artworks critically technology with the goal of the 'creative misuse' of generative AI and other technologies [15]. In 2023 ARRG! hosted a workshop at the hacker convention DEFCON 31, where they invited professional hackers to try bypass the guardrails around LLMs [16]. In one of their artistic experiments, Salvaggio prompted denoising diffusion models [9] to generate images of *'Gaussian noise'*, something that they are ironically very bad at doing (Fig. 1a) [17]. In another experiment, Swanson presented the discovered fictional character *LOAB* (Fig. 1b) that appeared as a persistent and hidden presence in the generations from text-to-image diffusion models. LOAB was discovered through the use of negative prompts, where text prompts are given in order to directly to distance the generated images from that representation [23]. In the case of *LOAB*, the image was originally found by only generating images with the negative prompt *'DIGITA PNTICS skyline logo::-1'*, which in turn was found by search for the opposite of the image of the actor Marlon Brando with the negative prompt *'Brando::-1'*. The emergence of this character, found through solely conditioning models with negative prompts, is a provoking example of hidden representations buried deep within the latent space of these generative AI models.

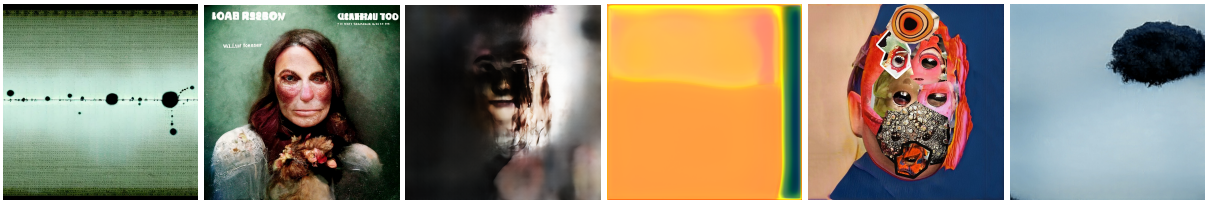
3 CORRUPTING A NETWORKS WEIGHTS

The goal of training a generative neural network is to encode features from the training data into the learned parameters (weights and biases) of the network. After training, the data is no longer needed and the model can be used to generate new data that resembles the training set. One way of intervening in the normal functioning of generation is to intentionally alter or corrupt the learned parameters (the weights and biases) after training. In the series of works *Neural Glitch* (Fig. 1c) the artist Mario Klingemann randomly altered, deleted or exchanged the trained weights of pretrained networks [12]. Through this process of corruption of the weights of the network, Klingemann deliberately amplified the artefacts of the model in the generated outputs and make visible the underlying computational process behind it's generation.

4 UPENDING A NETWORKS TRAINING

In the standard approaches to training a generative neural network, the goal is to minimise the divergence of the generated data distribution from the training data distribution (aka maximising likelihood). However, this is not the only way of configuring a loss function to train a generative neural network, alternative ways of configuring training provide methods for achieving *active divergence* from the original training dataset [2, 3].

There have been two artistic experiments where I upended the normal approach to training generative adversarial networks (GANs). In the standard GAN training regime, a generator network imitates a dataset, and a discriminator tries to spot 'real' images from 'fake' [8]. In creating the series of works *(un)stable equilibrium* (Fig. 1d) [4], I replaced the training data with another generator network, and got both generator networks to imitate each other. This allowed me to train both of the generator networks from scratch (starting with randomly initialised parameters) without any data whatsoever. In making the work *Being Foiled* I took a single pretrained



(a) (b) (c) (d) (e) (f)

Fig. 1. Images of various artworks. (a) Eryk Salvaggio's *Gaussian noise* [17]. (b) Steph Maj Swanson's *LOAB* [23]. (c) Mario Klingemann's *Neural Glitch* [12]. (d) Terence Broad's *(un)stable equilibrium 1:4* [4] (e) Mal Som's *Strange Fruit* [21]. (f) Jen Sykes' *The Offing* [24]. All images are reproduced with permission from the respective artists.

generator and discriminator pair and inverted the adversarial loss function, optimising the generator to produce images that the discriminator network predicted as being increasingly more 'fake', which quickly led to to the generation of very uncanny images [5].

To create the work *Strange Fruits* (Fig. 1e) [21] the artist Mal Som deliberately fed in generated outputs of the model back into the training dataset to induce mode collapse, a common failure state of GANs [3]. Som would later find model checkpoints in the transition state of collapse. Som's artistic approach was a precursor to more recent research highlight the issues of mode collapse when training generative models on their own synthetic outputs [1, 19].

5 HACKING THE COMPUTATIONAL GRAPH

The computational graph is the term given for the chain of computations, as defined by the input data, learned parameters, network topology, and computational functions that define the forward pass of a neural network (aka inference). In the *network bending* framework [6, 7], I developed a toolkit to allow for interventions into the computational flow of a model during inference. This approach allowed for a flexible and direct way of artistic manipulation of the internal representation of a generative model, using deterministically controlled filters that are inserted as their own layers into a generative model, to allow artists to make expressive changes to the flow of computation within the models themselves. Since it's introduction, a number of user-interfaces have been developed allowing interactive control of this approach [10, 14, 25].

In the process of the developing the network bending framework, I made a number of artworks with the framework. In *Teratome* I introduced aggressive filters into the earlier layers in the network, which acted as a major intervention into the formation of the images at their higher level structure, but not their lower-level fidelity. In the work *Fragments of Self* I used a latent vector that generated an image approximating my own image (a self portrait of sorts), and used network bending to again interrupt the generation of my own likeness, in a turbulent and constantly evolving fashion [7]. Another artist who used network bending in their work was Jen Sykes. In creation of the work *The Offing* (Fig. 1f), a model was trained on archival imagery of landscapes, which was then intentionally isolate and then manipulate images of the horizon [24].

6 DISCUSSION AND CONCLUSION

This paper argues that approaches to: subverting, corrupting, upending and hacking generative neural networks in their inputs, weights, training and inference; allows artists to use generative AI as an artistic material and to make critical works that reveal to us otherwise unseen aspects of these models generation. This hackers ethos provides a critical apporach for explainable AI (XAI) in the arts. These approaches are not dissimilar to the *glitch*

art and *databending* movements that were likewise seeking to reveal, through imperfection, otherwise hidden aspects and material functionality of digital media [11].

Generative AI produces media through a complex fabric of computation, contingent on large scraped datasets, where features and representations get encoded into the weights of unfathomably large data arrays, which in turn is enmeshed through complex chains of computation. The ease and realism through which this generated media is mass produced and it's almost uncanny flawlessness [20] makes it easy to forget the complex computational contingencies that produce it. Rather than simply using generative AI as a tool, treating it critically as an artistic material can help bring this complex fabric of computation to the fore.

REFERENCES

[1] Sina Alemohammad, Josue Casco-Rodriguez, Lorenzo Luzi, Ahmed Imtiaz Humayun, Hossein Babaei, Daniel LeJeune, Ali Siahkoohi, and Richard G Baraniuk. 2023. Self-consuming generative models go mad. *arXiv preprint arXiv:2307.01850* (2023).

[2] Sebastian Berns and Simon Colton. 2020. Bridging Generative Deep Learning and Computational Creativity. In *Proc. 11th International Conference on Computational Creativity*.

[3] Terence Broad, Sebastian Berns, Simon Colton, and Mick Grierson. 2021. Active Divergence with Generative Deep Learning - A Survey and Taxonomy. In *Proc. 12th International Conference on Computational Creativity*.

[4] Terence Broad and Mick Grierson. 2019. Searching for an (un)stable equilibrium: experiments in training generative models without data. *NeurIPS 2019 Workshop on Machine Learning for Creativity and Design* (2019).

[5] Terence Broad, Frederic Fol Leymarie, and Mick Grierson. 2020. Amplifying The Uncanny. *Proc. 8th Conference on Computation, Communication, Aesthetics and X (xCoAx)* (2020).

[6] Terence Broad, Frederic Fol Leymarie, and Mick Grierson. 2021. Network Bending: Expressive Manipulation of Deep Generative Models. *Proc. 10th International Conference on Artificial Intelligence in Music, Sound, Art and Design (EvoMUSART).* (2021).

[7] Terence Broad, Frederic Fol Leymarie, and Mick Grierson. 2022. Network Bending: Expressive Manipulation of Generative Models in Multiple Domains. *Entropy* 24, 1 (2022), 28.

[8] Ian Goodfellow, Jean Pouget-Abadie, Mehdi Mirza, Bing Xu, David Warde-Farley, Sherjil Ozair, Aaron Courville, and Yoshua Bengio. 2014. Generative Adversarial Nets. In *Advances in neural information processing systems*. 2672–2680.

[9] Jonathan Ho, Ajay Jain, and Pieter Abbeel. 2020. Denoising diffusion probabilistic models. *Advances in neural information processing systems* 33 (2020), 6840–6851.

[10] Tero Karras, Miika Aittala, Samuli Laine, Erik Härkönen, Janne Hellsten, Jaakko Lehtinen, and Timo Aila. 2021. Alias-Free Generative Adversarial Networks. *Advances in Neural Information Processing Systems* (2021).

[11] Jakko Kemper. 2023. Glitch, the Post-Digital Aesthetic of Failure and Twenty-First-Century Media. *European Journal of Cultural Studies* 26, 1 (2023), 47–63.

[12] Mario Klingemann. 2018. Neural Glitch / Mistaken Identity. https://underdestruction.com/2018/10/28/neural-glitch/. Accessed: 2024-05-02.

[13] Jason Koebler. 2024. Facebook's Algorithm Is Boosting AI Spam That Links to AI-Generated, Ad-Laden Click Farms. *404 Media* (2024).

[14] Jonas Kraasch. 2023. Autolume-Live: An interface for live visual performances using GANs. (2023).

[15] Eryk Salvaggio. 2023. The Algorithmic Resistance Research Group (ARRG!). https://cyberneticforests.substack.com/p/the-algorithmic-resistance-research. Accessed: 2024-05-02.

[16] Eryk Salvaggio. 2023. Cultural Red Teaming. https://cyberneticforests.substack.com/p/cultural-red-teaming. Accessed: 2024-05-02.

[17] Eryk Salvaggio. 2023. Writing Noise into Noise. https://cyberneticforests.substack.com/p/writing-noise-into-noise. Accessed: 2024-05-02.

[18] Philipp Schimtt. 2019. Introspections. https://philippschmitt.com/work/introspections. Accessed: 2024-05-02.

[19] Ilia Shumailov, Zakhar Shumaylov, Yiren Zhao, Yarin Gal, Nicolas Papernot, and Ross Anderson. 2023. The curse of recursion: Training on generated data makes models forget. *arXiv preprint arXiv:2305.17493* (2023).

[20] Amy Smith and Michael Cook. 2023. AI-Generated Imagery: A New Era for the Readymade'. In *SIGGRAPH Asia 2023 Art Papers*. 1–4.

[21] Mal Som. 2020. Strange Fruit. http://www.aiartonline.com/highlights-2020/mal-som-errthangisalive/. Accessed: 2021-02-05.

[22] Richard Stallman. 2002. On Hacking. https://stallman.org/articles/on-hacking.html. Accessed: 2023-08-03.

[23] Steph Maj Swason. 2022. I discovered this woman, who I call Loab [...]. https://twitter.com/supercomposite/status/1567162288087470081. Accessed: 2024-05-02.

[24] Jen Sykes. 2022. The Offing. https://j3nsykes.github.io/TheOffing/. Accessed: 2024-05-02.

[25] Shuoyang Zheng. 2023. Stylegan-canvas: Augmenting stylegan3 for real-time human-ai co-creation. In *Joint Proceedings of the ACM IUI Workshops*.