Demiurge
A Music of Babel

ROBERTO ALONSO TRILLO AND MAREK POLIKS

The authors explore the unfinished, attempting a fragmentary and blurred snapshot of an ongoing research project on the musical application of generative adversarial networks, entitled Demiurge. The project was initiated in July 2020 as a collaboration between an international team of interdisciplinary artists and researchers led by Marek Poliks (instrument design) and Roberto Alonso Trillo (violin). In keeping with its namesake, Demiurge asks questions about creation, specifically in the era of machine learning, and works to formalize a new creative process endemic to and natural within a world of machine collaborators.

We are here to explore the unfinished—attempting a fragmentary and blurred snapshot of an ongoing creative research project called Demiurge. The project was initiated in July 2020 as a collaboration between an international team of interdisciplinary artists and researchers led by Marek Poliks (instrument designer ∈ creative vector) and Roberto Alonso Trillo (violin ∈ creative vector). While one might define the project as an interactive synthesis engine built out of multiple neural networks, the spirit of Demiurge exceeds its technical scope. Instead, we have chosen to identify the project as a generative performance ecosystem, a self-documenting multimedia archive, and a subjective paradigm, independent of its creators, constructed out of networked brains and bodies. Befitting its namesake, Demiurge asks questions about creation, specifically in the era of machine learning, and works to formalize a creative process endemic to and natural within a world of machine collaborators.

REFLECTION 1: GANS

Demiurge’s machine core consists of three networked neural architectures, two of which are generative adversarial networks (GANs) [1]. Each network structure imparts its own complex relationship to musicality onto the project and its own communication style. The two GANs (una-GAN + mel-GAN) work together to generate sounds while a third neural network (a custom sequencer) interprets and sequences these sounds into larger-scale structures. After an initial development phase that explored a unidirectional sound-generation-to-sequencing process, our interests have refocused on exploring recursive relationships emerging among the various neural networks in play. These are available in their present, still-evolving state in a GitHub repository [2]. Let us consider each neural network separately.

Demiurge’s sound generation engine employs a modified version of the una-GAN architecture developed by Taiwan’s Academia Sinica Music and AI Lab [3]. Una-GAN is an unconditional GAN that generates noise-vector-based mel-spectrograms [4]. Una-GAN operates by interpolating sequences of noise information into recognizable audio features via a Boundary-Equilibrium GAN (BEGAN) [5]. Unlike prior approaches that operate on a single noise vector, una-GAN operates on a sequence of noise vectors whose aggregate is proportional in length to the output audio feature sequence. This operating principle enables una-GAN to generate sequences of any length, allowing our team to populate large databases of generated audio material. Given this lack of restriction in the time domain, una-GAN takes special precautions against feature incoherency and mode collapse, adopting a hierarchical generator that upsamples from “coarse” to “more refined” audio features and implementing a cycle-regularization process that correlates each noise vector frame to an output frame (promoting a diverse result) [6].

Una-GAN uses a linked vocoder, mel-GAN [7], to translate the generated mel-spectrograms into realistic-sounding waveforms. While mel-spectrograms are data-efficient for computation purposes, they are notoriously poor resolution sources for realistic audio. Trained mel-GAN models upsample their input mel-spectrograms 256x through convolutional
layers. The discriminator [8] against which these mel-GAN models have trained consists of a deconvolutional stack taking raw audio waveforms as input. Una-GAN’s output, as processed through mel-GAN, consists of variable-length audio, typified by relatively consistent behavioral profiles per model.

Our experimentation with una-GAN began with three priorities: audio morphology, quality, and diversity. How does one corral an unconditional GAN into producing output plausibly related to the input source without noticeable audio artifacts but with the variety required to feed our sequencer engine? We began by running una-GAN independently on several distinct collections containing different spectral registers and sound types performed by coauthor Alonso on violin (e.g., harmonics, trills, ordinario bowed sounds, and other extended techniques). We further tested different run lengths to assess the relationship between training duration and the plausibility (coherent morphology) of the results. To optimize audio quality at the expense of efficiency, we adjusted una-GAN to operate at 44.1 kHz audio instead of the machine-learning standard of 20.5 kHz (keeping a 16-bit-per-sample quantization level) and found significantly better results. To optimize the diversity of our generator’s output, we leveraged several modifications (batch size, input subdirectory, the dimensionality of noise vectors and mel-spectrograms, learning rate, etc.). We created a random-generation mode that sourced and matched una-GAN and mel-GAN models from indiscriminate run IDs tracked and stored in Weights and Biases [9], a cloud-based tool for tracking and visualizing machine learning pipelines.

Una-GAN’s performance relies on the quality of the mel-spectrogram inversion model through which it realizes its output. For the mel-GAN, we had to create flexible but characteristic models based on Alonso’s studio recordings. We experimented with database sizes of 22 GB and with runs of up to 500 epochs to facilitate this. When working with significant database inputs and many (≥1000) lengthy (24 h) training sessions, we had to alter our training environment from Google Colaboratory to a cloud solution leveraging Genesis Cloud [10] and gpu.land [11], local runs on a top-end Mac Pro with two AMD Radeon Pro Vega II GPUs, and an Exxact machine learning computer using two NVIDIA GTX 3090 GPUs. To increase training speed, we modified mel-GAN to utilize parallel processing, allowing us to train on up to 8 simultaneous GPUs.

Instead of manually sequencing output from our sound generation engine into a piece of music, we elected to use yet another neural network (the “Sequencer”) to make curatorial decisions about when and how to deploy this material. We first needed to build a database that indexed and stored audio descriptor information from our mel-GAN/una-GAN-generated raw material. We initially selected five simple metrics (spectral centroid, spectral flatness, spectral roll-off, fundamental frequency, and root mean square) extracted through Librosa’s Python package for music and audio analysis [12]. However, since these metrics proved insufficiently descriptive for our purposes, we chose to explore a more robust feature-extraction method: mel-frequency cepstral coefficients (MFCCs) (also sourced via Librosa).

The Sequencer is currently in development, as it requires a custom-built solution. After extensive training on preexisting audio, the Sequencer produces “predictive” strings of MFCC sequences that can be considered musical compositions. We are working with conventional sequence-to-sequence (Seq2Seq) frameworks [13], with mixed results from a long short-term memory (LSTM)/Encoder-Decoder model [14] and more intriguing results from a Transformer [15]. Demiurge then queries the descriptor database with the Sequencer’s MFCC compositions, identifying una-GAN/mel-GAN raw materials that bear similar characteristics, that is, similar strings of MFCC descriptors to the query set. Audio output comes from a simple Sampler, a program that concatenates the sequence-matched una-GAN/mel-GAN output audio into an audio stream, using another Python package, PyDub [16].

Once the Sequencer is primed with preexisting material (we have, with a bit of irony, started the process during development with the complete recordings of composer Samuel Barber’s Adagio for Strings), it will enter a musical dialogue with Alonso. Alonso will listen to the output of the Sequencer and provide his musical interpretation back into the Sequencer GAN as an input for future sequences. Multiple, if not endless, compositions will emerge (see Fig. 1).

REFlECtIon 2: ExpERImEntAtIon

Our creative process over the past six months has been marked by experimentation. Experimental—especially in the context of music—has become a term distended to the point of meaninglessness, often descriptive of any practice inclusive of some element of randomness. While working with unconditional GANs is a process profoundly steeped in randomness, randomness is not in itself a mobilizing theme for our research. Instead, when we identify this project with the experimental, we borrow liberally from Michael Pisaro’s interpretation of the category, operating in the service of “some fundamental choice about the ‘future’ of music” [17]. Demiurge begins with the assumption that there are nascent musicalities implicit in machine learning architectures. Our task is to make them audible.

Demiurge’s neural networks have forced us to (un)develop our ears. Operating both internally and externally through a process of self-deception, GANs are tricky—and their associated discourse then often revolves around the efficacy of their ability to deceive. For months, we listened to the GANs, prioritizing morphologies consistent with acoustic realism above all else. We listened with an orientation, a projection toward, an inclination of the ear toward a “dis-corporate, context-independent beyond” [18]. Following Jean-Luc Nancy, “to listen is to be straining towards a possible meaning” [19], to listen— and not to listen— to anthropomorphize “a space or action, littering it with secrets and intentions” [20]. We beheld the GANs as emulators (as which they work very well) and found ourselves confronted with something quite mundane—an unorganized repository of plausible violin sounds. We came to miss the messier
sounds of poorly trained models, mismatched components, and collapsed runs.

Demiurge’s neural networks thus forced us to accept the procedural significance of failure, to embrace glitch. We have come to understand glitch as more than mechanical error or unforeseen failure to function but also as a “form of refusal,” an “erratum,” and a “non-performance.” Such a failure to perform “reveals technology pushing back against the weighty onus of function” [21]. Failure became an aesthetic constituent, a reminder of the fact that “our control of technology is an illusion” and that the tools that we create “are only as perfect, precise, and efficient as the humans who built them” [22]. As we worked at those edge boundaries, unexpected behaviors morphed from “systemic malfunctions” to meaningful allowances emerging in the “intersection between intention and expectation” [23]. Whenever the GANs delivered something different from what we had foreseen and what we had coded them to do, the sensibilities of their architectures became more apparent. Demiurge’s GANs reminded us of the relationship between failure and identity.

How then to unearth these “in silico” sensibilities and make them audible? We implemented recursive elements at every stage, allowing network to run network—the Sequencer sequencing the output of the audio generation GANs, intertwined violin performance, and sampler output informing Sequencer input, audio generator output feeding into audio generator input. This process remains ongoing as we continue to feed the Demiurge into itself and listen to what emerges, allowing “all data to become fodder for sonic experimentation” [24]. We do not claim that any part of this output is authentic to the architecture of the GANs and neural networks themselves but rather two things: (a) that recursive systems tend to aggregate and display certain intrinsic biases, and (b) that it is easier to apprehend the structure of an object with our senses as said object diverges from its resemblance to another assumed form.

What began as an academic ambition to contribute to current research on GAN-driven audio digital sound processing (DSP) morphed into a strange creative dialogue with a machine. As we began to consider the Demiurge GANs as our dialogical partner, we needed to examine our role in or relationship with the project. How could we conduct ourselves in a manner that afforded our object of study the necessary agency to reveal itself?

**REFLECTION 3: THE ARCHIVE**

Concerns emerged as we worked to codify our relationship to the project, as the apparent contradiction between a composed piece of music (a contained unilateral process with a discrete end) and our new framework (an abstract, recursive process with neither boundaries nor an end). The notion of constructing a piece as one of a kind was abandoned, and instead, it became our operational understanding that a piece would emerge in some to-be-determined form. At this point, we began to identify this project with experimentalism or experimentation actively.

We also embraced McKenzie Wark’s metaphor of excommunication:

> Each this is connected to another that, and each that to another this. There’s no beginning or end, and there is always either an excess or a lack to any particular communication, a more-than or less-than. . . . It can take the form of an alien mode of communication . . . which nevertheless seems legible, at least to someone within the sphere of communication [25].

The project’s constituents thus adopted the loose identity of a swarm, “[a] flocking algorithm, a . . . distributed protocol for a network of communicating bodies” [26]. As Alexander Galloway points out, “swarms and systems threaten the sanctity of the human . . . they violently reduce mind to matter, disseminating consciousness and causality into a frenzy of discrete, autonomous agents, each with their own micro functions” [27]. With this in mind, we aspired to avoid, transcend, or even reverse what we saw as the germinal anthropomorphizing function of GAN architectures.
Demiurge required an archive, some preexisting input with and against which to train. While we initially conceived this archive as neutral (e.g. the total of Alonso's violin recordings for six months), the contents became charged once we considered co-participants within some recursive and emergent system. The team asked Alonso to generate music in dialogue with Demiurge's outputs. As the GAN's responding dialogue became less violin-like, Alonso incorporated audio from elsewhere within his studio environment. The thought occurred to record a video of himself doing so. As the entanglement became increasingly extramusical, Alonso found himself archiving everything—the view out of his hotel room window during COVID-19 travel quarantine, the sounds from sanitation rituals. A database emerged of Roberto-ambience, an ambient database—to which and from which dialogue with Demiurge's output database proliferated. The collaborators found it appropriate that the cloud mediated all work on Demiurge, a network of international participants (United States, Spain, and Hong Kong) isolated by COVID-19 restrictions but united by the same substrate used to communicate with Demiurge. As both databases began to proliferate, we provided them with a mouthpiece to the world. One can visit a dedicated website that streams video and audio generated by our una-GAN + mel-GAN implementation in real time. This audio stream consists of uncurated raw material. Opening up Demiurge in this way honors the spirit of the swarm—emergent material, transparency of internal systems, fluidity, and mutual entanglement of its participants. The website allows the user to witness Demiurge's evolution through time as a living and morphing organism. It will include documentation on future in situ realizations and linked sound-installation projects that we are currently developing. We have also made this database of "artificial" audio matter available in its entirety. We intend to build a commenting platform into the public database, allowing users external to the project (although this project knows no formal externality) the ability to index, comment on, and experiment with the output. An open call for electronic compositions (Demiurge's Debris) with this material is underway. The "ambient" database of Alonso's inputs to Demiurge and that containing the video material are also openly available through Google Drive. A detailed discussion of Demiurge's Architecture (Reflection 4) and its Ecosystemic Nature (Reflection 5) have been included as supplemental online material expanding the arguments introduced thus far.

INCONCLUSIVE THOUGHTS

Why Demiurge? Two concluding borrowed reflections, two open windows hinting at the genesis of a name. First, the Swiss philosopher Serge Margel on a radical reconsideration of Plato's Timaeus in The Tomb of the Artisan God:

The demiurge is a polymorphous being, a being whose forms are of a great variety and an extreme complexity. Not only is it up to him to judge and evaluate what seems good for the best of worlds, but the art of building and producing the soul and body of the world in its totality also falls upon him. The demiurge's power nonetheless appears limited. It is restricted to the space and time of its own operation: to space, for the site where the primary elements composing the world are formed and assembled delimits a preliminary constitutive structure, to which each of the demiurge's gestures are absolutely subject; and to time, for this artisan god will only be able to produce a systematic assemblage of the world in accordance with a determinate order of succession and a rigorously enumerable duration. In fact, and in the strict sense, the demiurge is not an engineering creator but rather a divine architect. He has the supreme power in his hands not to produce the world from nothing but to survey, situate, construct, and realize the life of an organized whole from elements already charged or informed with radiant energy.

Previously, in 1942, Jorge Luis Borges rhapsodized on the endless parallel bifurcations of time, 15 years before this view became conventional scientific theory (in Hugh Everett's many-worlds interpretation of quantum physics), reflecting on the nature of a literary Demiurge, The Garden of Forking Paths:

The Garden of Forking Paths is an incomplete, but not false, image of the universe as T'sui Pên conceived it. In contrast to Newton and Schopenhauer, your ancestor did not believe in a uniform, absolute time. He believed in an infinite series of times, in a growing, dizzying net of divergent, convergent and parallel times. This network of times, which approached one another, forked, broke off or were unaware of one another for centuries, embraces all possibilities of time.

Demiurge, music of Babel, is a spheric ecosystem whose exact center is any one of its points—an open mouth, singing.

Acknowledgments

The entire team that made Demiurge possible includes the authors and the following collaborators: Peter Nelson, François Moulliot, Daniel Shanken, Mathis Anthony, Ryan Au, Finn Mai, and Maya Duan. The Demiurge project has been partially funded by an IRCMS grant from Hong Kong Baptist University.

References and Notes

1 Ian J. Goodfellow et al., “Generative Adversarial Nets,” Proceedings of the 27th International Conference on Neural Information Processing Systems 2014, MIT Press, Cambridge (MA): pp. 2672–2680: https://arxiv.org/pdf/1406.2660.pdf.
Alonso Trillo and Poliks,

MusiC And sound Ar T

2 https://github.com/robertoalonsotrillo/_Demiurge_.

3 Jen-Yu Liu et al., “Unconditional Audio Generation with Generative Adversarial Networks and Cycle Regularization,” Proceedings of Interspeech 2020, 1997–2001: 10.21437/Interspeech.2020-1137.

4 A mel-spectrogram represents a given frequency spectrum in a way that is biased toward human perception. As opposed to a classical power spectrogram, which represents frequencies along a linear axis, mel-spectrograms represent frequency magnitudes along a logarithmic axis—a distribution that better correlates with the ability of the human ear to distinguish frequencies.

5 David Berthelot et al., "BEGAN: Boundary Equilibrium Generative Adversarial Networks" (2017): https://arxiv.org/abs/1703.10717.

6 See Liu et al. [3] p. 4.

7 Kundan Kumar et al., "mel-GAN: Generative Adversarial Networks for Conditional Waveform Synthesis," 33rd Conference on Neural Information Processing Systems (NeurIPS 2019), Vancouver, Canada.

8 Generator and discriminator are the fundamental constituent modules of a GAN architecture. The discriminator works as a classifier, attempting to distinguish real data from that created by the generator.

9 https://wandb.ai/.

10 http://gpu.land/.

11 http://librosa.org/doc/latest/index.html.

12 https://librosa.org/doc/latest/index.html.

13 Kou Tanaka et al., "ATTTS2S-VC: Sequence-to-sequence Voice Conversion with Attention and Context Preservation Mechanisms," ICASSP 2019 IEEE International Conference on Acoustics, Speech, and Signal Processing (ICASSP) (2019) pp. 6805–6809: DOI: 10.1109/ICASSP.2019.8683282.

14 Hasim Sak et al., “Long Short-term Memory Recurrent Neural Network Architectures for Large Scale Acoustic Modeling,” Proceedings of the Annual Conference of the International Speech Communication Association, INTERSPEECH (2014) pp. 338–342: https://arxiv.org/abs/1402.1128.

15 Ashish Vaswani et al., “Attention is All you Need,” Proceedings of the 31st International Conference on Neural Information Processing Systems (Red Hook, NY: Curran Associates, 2017) pp. 6000–6010.

16 https://github.com/jiaaro/pydub.

17 Michael Pisaro, “Eleven Theses on the State of New Music,” in Eva-Maria Houben and Burkhard Schlotehauer, MusikDenken Texte der Wandelweiser Komponisten (Zurich: Edition Howeg, 2008).

18 Marek Poliks, “Against Listening,” ISSUU (2014): issuu.com/marekpoliks/docs/poliks-_against_listening (accessed 1 February 2021).

19 Jean-Luc Nancy, Listening (New York: Fordham Univ. Press, 2007) p. 6.

20 Poliks [18].

21 Legacy Russell, Glitch Feminism: A Manifesto (London: Verso, 2020) p. 29.

22 Kim Cascone, “The Aesthetics of Failure: ‘Post-Digital’ Tendencies in Contemporary Computer Music,” Computer Music Journal 24, No. 4 (2000) p. 17.

23 Kim Cascone and Petar Jandrić, “The Failure of Failure: Postdigital Aesthetics Against Techno-mystification,” Postdigital Science and Education 3 (2021) p. 569.

24 See Cascone [22] p. 17.

25 Alexander R. Galloway, Eugene Thacker, and McKenzie Wark, Excommunication: Three Inquiries in Media and Mediation (Chicago: University of Chicago Press, 2013) pp. 160–161.

26 See Galloway et al. [25] p. 63.

27 See Galloway et al. [25] p. 157.

28 https://robertoalonsotrillo.com/portfolio/demiurge/.

29 https://drive.google.com/drive/folders/1dlkSlQyJXLiaQ8Zgbge1K_xhxUTKe_in?usp=sharing.

30 https://robertoalonsotrillo.com/portfolio/debris/.

31 https://drive.google.com/drive/folders/11EU7aeGhzfjjudAlZwNoeAdLPtnGBotQ?usp=sharing.

32 Serge Margel, The Tomb of the Artisan God: On Plato’s Timaeus (Minneapolis: University of Minnesota Press, 2019) p. 11.

33 Hugh Everett III et al., The Everett Interpretation of Quantum Mechanics: Collected Works 1955–1980 with Commentary (Princeton: Princeton Univ. Press, 2012).

34 Jorge Luis Borges, The Garden of Forking Paths (London: Penguin, 2018) pp. 15–16.

Manuscript received 6 May 2021.

ROBERTO ALONSO is a musician and theorist based in Hong Kong. His practice explores the connections between different artistic disciplines, from dance and music to video art and interactive sound installations. His recent work examines networked hybrid creative models endemic to a world increasingly mediated by AI and machine learning.