JASs Journal of Anthropological Sciences
Open Access
forum Vol. 103 (2025), pp. 217 - 220
doi. 10.4436/jass10316

Faking it for science: generative artificial intelligence at the crossroads of biological and cultural anthropology

Aicha Ben Taher

School of Biological and Behavioural Sciences, Queen Mary University of London, UK
e-mail: a.bentaher@qmul.ac.uk

The Lie That Learns, Ctrl + C, Ctrl + GenAI = DATA

Generative artificial intelligence (GenAI) models are gradually transforming the way the scientific community engages with data. What truly sets them apart from traditional statistical-based algorithms and other machine learning models, often used in prediction, classification, or decision-making, is GenAI's ability to mimic the statistical properties and patterns of different types of datasets. In some cases, the results are so realistic that they challenge our ability to distinguish between the synthetic and the authentic. Building on the widespread adoption of GenAI across numerous domains, it is now gaining traction as a valuable resource for use in anthropological research. Much like how archaeological relics unveil narratives of the past, GenAI offers insight into the hidden structure and behaviour of complex datasets, helping uncover connections and anomalies that might otherwise remain buried. One exciting implementation of GenAI is the use of Generative Adversarial Networks (GANs), a technology introduced over a decade ago that has since made a significant impact across various fields thanks to its ability to generate high-quality synthetic data (Goodfellow et al. 2014).

Rewriting humanity using data we didn't collect

The most immediate benefit of this technology is particularly evident in medical imaging, as GANs can produce high-resolution medical images from low-quality scans. When trained well, they're also excellent at removing noise and reconstructing corrupted parts of medical images. Oraby et al. (2025) demonstrated the use of GANs to enhance the resolution of MRI scans, thereby amplifying diagnostically relevant features that may be indistinct in standard imaging. These refined images were subsequently used to differentiate between stages of Alzheimer's disease. In a similar application, Dee et al. (2024) addressed limited training data by generating synthetic histopathological thyroid cancer images. Despite a small sample size, their GAN model produced realistic images that augmented a deep learning classifier for subtype identification. This improved the generalisability of external datasets, especially in the minority class tumour subtypes, which are notoriously difficult to classify due to their rarity. Such applications of GANs demonstrate the practical utility of synthetic data in mitigating the risks of misclassification, even with limited real-world annotated data.

Beyond medical image generation, GenAI is transforming human genomics by generating synthetic genetic data that explores diversity, population history, and evolution while safeguarding privacy. In the first study of its kind, Yelmen et al. (2021) used GANs to generate high-quality synthetic human genomes that closely mirrored real genetic datasets. Their model successfully replicated key features critical to understanding human variation, such as allele frequencies, linkage disequilibrium, and population structure. Importantly, the researchers evaluated the privacy risks of synthetic genome generation using log-likelihood ratio attacks, demonstrating that their method posed minimal risk of individual

re- identification. This has significant implications for anthropological genetics, as it potentially expands access to sensitive datasets from underrepresented groups. Expanding on this, Wang et al. (2021) developed the first GAN for population genetics, which allows researchers to infer complex demographic scenarios, such as population bottlenecks, migrations, and admixture events. This approach reduces the need for extensive parameter tuning and allows for more flexible modelling of human evolutionary histories. For anthropologists, this opens new possibilities for reconstructing past population dynamics without relying solely on limited or ethically restricted datasets. As the focus on GAN applications in human genomics continued to grow, Booker et al. (2022) introduced a novel approach in their study *This Population Does Not Exist,* where they trained GANs to generate synthetic population genetic alignments rather than full genomes. They successfully modelled a wide range of evolutionary scenarios, including selection, population subdivision, and demographic change. This method allows researchers to explore hypothetical evolutionary pathways and test theories about human ancestry without needing direct access to sensitive or proprietary genetic data.

Despite these advancements, the effectiveness of GenAI models in generating synthetic data for underrepresented populations remains limited. Recent efforts by Marchesi et al. (2025) show promise in mitigating representation bias and improving fairness in synthetic health data generation, particularly for minority subpopulations. However, further work is needed to ensure that synthetic data reflects the full spectrum of human biological variation.

The monster we made: synthetic data. Human stakes

As we embrace GenAI, we must also confront the profound ethical challenges it brings. First and foremost, we must reconsider how we define data. GenAI offers new modes of data creation and analysis, challenging traditional notions of what constitutes "real" or "authentic" data. If GenAI can simulate biological or cultural patterns using synthetic data, then what does it truly mean for data to be "real" or "authentic"? Therefore, it is essential to establish clear and consistent definitions distinguishing synthetic data from authentic data. Current distinctions vary significantly across disciplines, often leading to ambiguity. In fields such as cultural anthropology, the distinction is relatively straightforward: data collected through immersive fieldwork is typically considered authentic, whereas data produced by models like GANs would be classified as synthetic. However, in other fields, the boundary can be far less discernible, especially when algorithmically generated data is used alongside real-world data. One particularly well-articulated definition of synthetic data comes from the Allan Research Institute and the Royal Society, which describes it as data created using purpose-built mathematical models or algorithms to address specific data science tasks (Jordon et al. 2022). This framing helps clarify the intent and application of synthetic data, contributing to a more unified definition that can be applied consistently across disciplines.

Another key ethical concern in the use of GenAI involves questions of consent and ownership. As GenAI becomes increasingly capable of replicating patterns and addressing issues such as data scarcity or underrepresentation, it forces us to confront a fundamental dilemma: who owns the output? Even when the intention behind generating synthetic data is harmless (i.e. enhancing privacy or improving inclusivity), the act of copying or simulating original data raises questions about representation and consent. For instance, if I create a synthetic replica of a cultural artefact, who should rightfully claim ownership of that replica? Does it belong to the model's developer, or the community from which the original artefact emerged? Since GenAI models do not produce truly original work but instead replicate statistical patterns from their training data, generating creations reminiscent of Frankenstein's monster: a creation stitched together from fragments yet void of originality, should we really attribute authorship

to a GenAI model? Shouldn't we revert ownership to the original data sources and their communities? Or should we account for the developer's own biases and experiences that shape how the model is trained?

These questions are especially urgent in the context of human studies and anthropology, where data often originates from communities with deep historical and cultural significance. In these fields, the unauthorised use or replication of cultural data, particularly from marginalised groups, can perpetuate long-standing patterns of appropriation, exploitation, and epistemic injustice. Currently, GenAI models are frequently being trained on vast datasets without the explicit knowledge or permission of the individuals or communities represented within them. While the models themselves may be agnostic during training, their outputs are not. Once trained, they reflect and reproduce the data they have absorbed, often reinforcing dominant narratives or biases without considering minorities who have historically been misrepresented. This issue strikes at the heart of our discipline as we aim to prioritise ethical representation, cultural sensitivity and decolonial methodologies. Ultimately, without transparent consent mechanisms and clear frameworks for ownership, the use of synthetic data in anthropology and other human-centred research risks reinforcing the very inequalities it seeks to address.

Riding the wave to reclaim human insight in a synthetic age

In this data-driven era, information becomes meaningful only when harnessed effectively. Anthropologists can no longer afford to remain passive observers as tools like GenAI continue to evolve. Instead, they must ride the wave and take an active role in shaping the development of culturally informed systems that avoid algorithmic bias and embrace human diversity. Crucially, they have a responsibility to intervene in policy-making to ensure that GenAI and other emerging technologies do not further marginalise underrepresented groups. These tools must be guided to uphold their rights, protect their knowledge systems, and respect their cultural sovereignty, as their voices and agency are too often sacrificed in the name of advancement and innovation. Finally, GenAI is not here to replace the role of Anthropologists. GenAI is a tool, just like the tools that have come before and the many countless tools that will follow. Its purpose in research should never be to replace empirical evidence with synthetic invention, but rather it should be used to complement and inform experimentation. However, this does not mean it should be dismissed or its presence refuted by clinging rigidly to traditional methodologies. Resisting such change risks missing an opportunity to enrich the discipline's methodologies and exclude anthropology from critical conversations where cultural insight is more essential than ever.

References

Booker WW, Ray DD, Schrider DR (2023) This population does not exist: Learning the distribution of evolutionary histories with generative adversarial networks. Genetics 224: iyad063. https://doi.org/10.1093/genetics/iyad063

Dee W, Alaaeldin Ibrahim R, Marouli E (2024) Histopathological domain adaptation with generative adversarial networks: Bridging the domain gap between thyroid cancer histopathology datasets. PLoS One 19: e0310417. https://doi.org/10.1371/journal.pone.0310417

Goodfellow IJ, Pouget-Abadie J, Mirza M, et al (2014) Generative Adversarial Networks. arXiv:1406.2661v1. https://doi.org/10.48550/arxiv.1406.2661

Jordon J, Szpruch L, Houssiau F, et al (2022) Synthetic data - what, why and how? arXiv:2205.03257. https://doi.org/10.48550/arxiv.2205.03257

Marchesi R, Micheletti N, Kuo NI-Hsien, et al (2025) Generative AI mitigates representation bias and improves model fairness through synthetic health data. PLoS Comput Biol 21:e1013080. https://doi.org/10.1371/journal.pcbi.1013080

Oraby S, Emran A, El-Saghir B, et al (2025) Hybrid of DSR-GAN and CNN for Alzheimer disease detection based on MRI images. Sci Rep 15:12727. https://doi.org/10.1038/s41598-025-94677-9

Wang Z, Wang J, Kourakos M, et al (2021) Automatic inference of demographic parameters using generative adversarial networks. Mol Ecol Resour 21:2689-2705. https://doi.org/10.1111/1755-0998.13386

Yelmen B, Decelle A, Ongaro L, et al (2021) Creating artificial human genomes using generative neural networks. PLOS Genet 17:e1009303. https://doi.org/10.1371/journal.pgen.1009303