Dancing with Synthetic Data: AI Educational Research using an AR Ballet

Genevieve Smith-Nunes[1] and Alex Shaw[2]
[1] University of Roehampton, London, UK
[2] Glastonbridge Software, Edinburgh, Scotland
ges52@cantab.ac.uk

Research In Progress

Abstract

Synthetic data (SD) is becoming an increasingly important tool in artificial intelligence (AI) research, particularly in domains where real-world data can be difficult or costly to obtain. In this research-in-progress paper, we explore the use of SD derived from brainwave and movement data to power an augmented reality (AR) episodic ballet experience. The goal of this WIP is to prompt discussions around the ethical use of body data in computing education through immersive technologies and to explore new technologies for teaching and learning within computing education. By leveraging SD rather than real user data, we aim to create an immersive AR experience that allows exploration of the relationship between physical movement, cognition, and artistic expression, while avoiding potential privacy and consent issues associated with the use of personal biometric data. Through this WIP, we investigate the technical challenges and opportunities in using SD to enable novel educational experiences, as well as the broader implications for the role of synthetic data in AI-powered educational research and applications. Our findings have the potential to inform best practices around the ethical development of data-driven educational technologies that respect individual privacy and autonomy.

Keywords: Synthetic Data, Biometrics, Data Ethics, AI, Computing Education, XR

1.0 Introduction

As artificial intelligence (AI) applications grow across diverse fields, the need for accessible, cost-effective data is more pressing than ever. Synthetic data (SD) has emerged as a valuable tool to address this need, especially in scenarios where obtaining real-world data is challenging, costly, or raises ethical concerns. In this research-in-progress paper, we investigate the use of SD generated from brainwave and movement data to create an augmented reality (AR) episodic ballet experience. The educational research from this project is in the field of creative computing through the following: programming for SD generation and AR development, ethical discourse on data ethics and AI.

This work in progress (WiP) is an exploration of synthetics for computing education and an opportunity to reflect on the ethical implications of using biometric data within immersive learning environments. By substituting synthetic for real biometric data, we aim to create a secure and engaging AR experience and learning resources that invites users to explore connections between programming, movement, cognition, and expression, while safeguarding privacy. This study seeks to spark discussion on the responsible integration of AI approaches in creative computing education, outlining both technical challenges and opportunities of using SD in AI-driven educational applications. Highlight the potential for SD to reshape data ethics in immersive learning technologies

1.1 What Is Synthetic Data

SD in simple terms refers to artificially generated data that replicates the statistical properties and patterns of real-world datasets without exposing sensitive or personally identifiable information (PII) (Rubin 1993). SD potentially offers solution to privacy, data scarcity, and financial challenges of real-world data generation. Recent research emphasises it role in domains beyond education such as Health Care (Goncalves et. Al., 2020), Governmental Census data (Abowd & Hawes, 2020), and Finance (Altman et al, 2024).

1.2. Why Use Synthetic Data?

SD can be a valuable tool in cases where using real-world data raises ethical concerns. For example, in research involving minors or other vulnerable populations, SD enables researchers to conduct analyses without directly involving these groups. Particularly relevant in educational research, where privacy concerns are paramount, and the use of real-world data involving minors requires strict ethical protocols (Adams et al., 2023). We aim to explore and develop a publicly available product without compromising participant privacy or violating ethical guidelines. SD approach(s) allow research to move forward while respecting the rights and well-being of individuals represented in the original data. It should be noted that there are no specific guidelines on the best use-case for SD (Dankar & Ibrahim, 2021), especially in education. In this WiP we use movement and brainwave datasets to generate the synthetic data. We purposefully selected python for SD generation aligning with computing education practices in England (Hadwen-Bennett & Kemp, 2024, p.10).

2. Project Overview / Background

This paper focuses on the processes, data, and creative computing techniques that will be developed into computing education resources aimed at pre-university students and non-technical artists. It forms part of a larger project, a data-driven AR ballet. The AR experience consists of five episodes, built using real and synthetic biometric datasets. The AR storyline, see table 1, follows four astronauts as they make the first ever manned journey into the vastness beyond our solar system. We follow them as they train, blast off, and explore, visualised through a web of personal journeys, biometrics, and digital simulations. It explores how our data-driven society shifts the way we perceive reality, each other, and presents the contention between enriching and dehumanising ourselves through data measurement. Only one episode will rely entirely on synthetic data.

2.1 AR Episodes

Episode	Story	Pedagogical
1 Training for Space	Focuses on the astronauts' physical and cognitive preparation using only synthetic biometric data.	Illustrate the foundational relationship between biometric analysis and performance optimisation.
2. The Launch	Simulates the physical and emotional intensity of leaving Earth, combining real and synthetic data for immersive effects.	Highlight SD's role in simulating extreme scenarios.
3. Space Station Alpha	A pause in the journey: data testing, communications, and preparation for interstellar travel	Practical computing education of data transmission, latency, and biometric data. Facilitate ethical SD discussion applications.
4. **Interstellar Travel**:	Vast distance from earth, consent monitoring.	Recursion, iteration. Data less predictable, outliers
5. **Personal Journeys:**	Astronauts feeling so far removed from 'home', detached and less able to see the importance of their data at this distance from earth. Comms glitches	Connect data to storytelling and emotional expression. Bridge real-world and SD augmentation with unknown future horizons.
6. **Signal Failure**:	Concludes with signal failure – unknown ending, loss of data, catastrophe, for astronaut agency of personal data.	Designed to be unknowable and potentially uncomfortable for discussion. Communication at this point is reduce the raw 1's and 0's

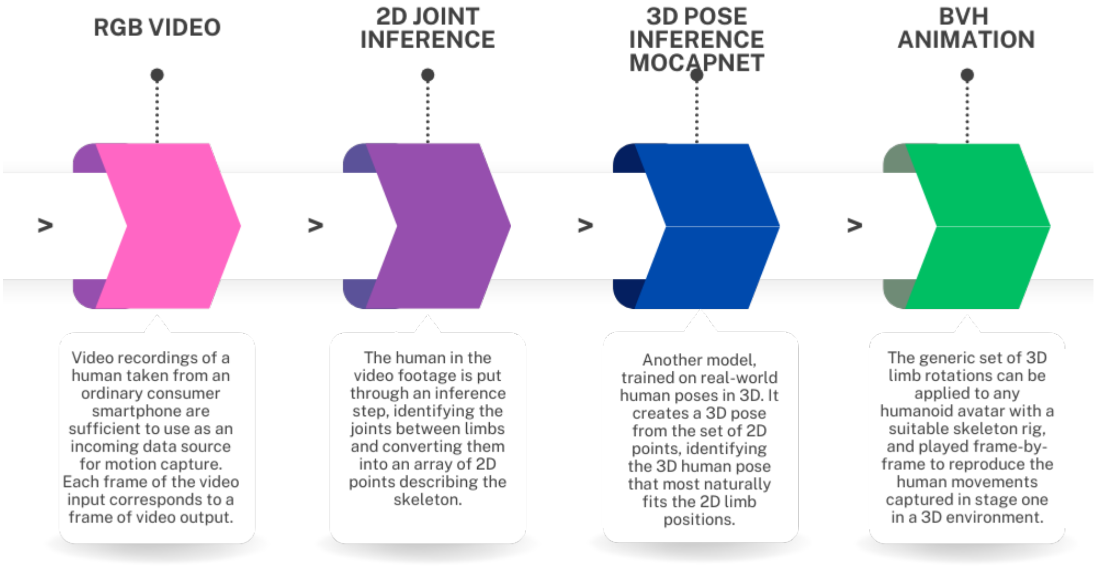
Table 1. AR Episode Overview and Pedagogical Intent.

3. Method

For the AR ballet, dancers' biometric datasets were (i) motion-captured using an extended pipeline, fig 1. Motion data, using a markerless motion capture system and MocapNET (Qammaz & Argyros, 2019, 2023) to create 3D avatars in BVH (BioVision Hierarchy) format. (ii) CSV format of Electroencephalography (EEG) data for augmenting the graphic and sound effects in the AR ballet, not for neuroscience purposes. Note that these techniques are solely to produce mathematical representations of our dancers' movement and are not related to the current trend of AI and ethics that repurposes existing creative work.

3.1 Movement Data: Real > Synth

Our process is based around MocapNET (Qammaz & Argyros, 2019, 2023) , a research motion capture project relying on a single RGB video stream to generate a 3D pose estimation of a human dancer. MocapNET uses two inference stages, the first one identifies 2D joint positions from an image, and the second estimates the 3D pose of the human from 2D joint positions. This final inference result can be exported as BVH and applied to rigged 3D models.



Video recordings of a human taken from an ordinary consumer smartphone are sufficient to use as an incoming data source for motion capture. Each frame of the video input corresponds to a frame of video output.

The human in the video footage is put through an inference step, identifying the joints between limbs and converting them into an array of 2D points describing the skeleton.

Another model, trained on real-world human poses in 3D. It creates a 3D pose from the set of 2D points, identifying the 3D human pose that most naturally fits the 2D limb positions.

The generic set of 3D limb rotations can be applied to any humanoid avatar with a suitable skeleton rig, and played frame-by-frame to reproduce the human movements captured in stage one in a 3D environment.

Figure 1. Motion Capture Pipeline

RBG video - 2D joint data – 3D pose – BVH 3D skeleton

3.2 EEG: Real > Synth

EEG (brainwave) data was recorded using Emotiv's Insight 5-channel headset and exported as unprocessed data to a CSV format. The data is anonymised and synthesised to create new SD sets. This process is simpler than synthesising movement data due to data structure.

3.3 Teaching Synthesising Data with Ethics

Google Colab with SynthDataVault (Patki et al., 2016) served as an interactive tool for SD generation. Pedagogical methods emphasized interdisciplinary learning by integrating Python-based programming, ethical discussions, and real-world applications in XR. For example, datasets such as dancer biometric capture stories facilitated ethical debates, while programming exercises focused on SD generation principles.

4. Pedagogical Innovation

Pedagogical innovation aims to integrate synthetic data generation, programming, and extended reality (XR) development through creative computing education. Educational objectives (i) developing students' AI competencies, (ii) generation and use of SD (iii) in XR applications, and (iv) foster critical thinking about ethical AI approaches. Building understanding of essential concepts like data privacy including differential privacy (Wood et al, 2018), algorithmic bias, and ethical considerations in synthetic data generation. Additionally, illustrates the contention between data synthesis and the intellectual rights of the person who created the training data, and how we can treat artists fairly in the AI age.

The project aligns with established computing education frameworks, which emphasise the importance of AI competencies, ethical discourse, and practical programming skills (Hadwen-Bennett & Kemp, 2024). It incorporates XR to support diverse learning outcomes and enhancing educational experiences through emerging technologies.

5. Ethics and Limitations

The three main areas of concern: bias, information loss, and ethical implications.

SD may reinforce and potential enhance existing bias form the original dataset. Loss of subtle nuances could manifest misrepresentation. The need for transparent practices for obtaining and processing biometric data are vital. Data subjects must fully understand and explicitly consent to the use of their data, even when presented in synthetic form, to uphold ethical standards and maintain trust. Ethical implications for artists, developers, and participants are significant concerns. Data ownership and intellectual property (IP) of contributions *must* be recognised and protected.

6. Next Steps

Through developing the process and learning resources for SD design for creative computing purposes we aim to test the functionality of the process of both (i) creating synthetic dataset and (ii) building XR experience with those datasets. Alongside ethical discussion of body data ethics.

Research and Design Iterations:

- Test and validate process for creating SD designed for creative computing education. (April - July with pre-service secondary computing teachers and four CS undergraduates)
- Evaluate the effectiveness of using these SD in building XR (Extended Reality) experiences. (June - Sept)
- Examine the ethical implications of collecting, generating, and using body-related data (July onwards)

We are currently generating all the synthetic biometric datasets for use in the AR ballet. It is very difficult to create synthetic ballet movement data. We believe that synthetic data will not be as effective as data generated by real dancers. However, this research will help us understand the limitations of synthetic movement data and develop teaching materials to explain the process and ethics of data synthesis.

References

Abowd, J. M., & Hawes, M. B. (2023). Confidentiality Protection in the 2020 US Census of Population and Housing. *Annual Review of Statistics and Its Application*, *10*(1), 119–144. https://doi.org/10.1146/annurev-statistics-010422-034226

Adams, C., Pente, P., Lemermeyer, G., & Rockwell, G. (2023). *Ethical principles for artificial intelligence in K-12 education.* Computers and Education: Artificial Intelligence, 4, 100131. https://doi.org/10.1016/j.caeai.2023.100131

Altman, E., Blanuša, J., Egressy, B., Anghel, A., & Atasu, K. (n.d.). *Realistic Synthetic Financial Transactions for Anti-Money Laundering Models*.

Dankar, F. K., & Ibrahim, M. (2021). Fake It Till You Make It: Guidelines for Effective Synthetic Data Generation. Applied Sciences, 11(5), 2158. https://doi.org/10.3390/app11052158

Goncalves, A., Ray, P., Soper, B., Stevens, J., Coyle, L., & Sales, A. P. (2020). Generation and evaluation of synthetic patient data. *BMC Medical Research Methodology*, *20*(1), 108. https://doi.org/10.1186/s12874-020-00977-1

Hadwen-Bennett, A., & Kemp, P. (2024). Programming in Secondary Education in England: Technical Report. King's College London. https://www.kcl.ac.uk/ecs/assets/programming-in-secondary-education-in-england-full-technical-report.pdf

Patki, N., Wedge, R., & Veeramachaneni, K. (2016). The Synthetic Data Vault. 2016 IEEE International Conference on Data Science and Advanced Analytics (DSAA), 399–410. https://doi.org/10.1109/DSAA.2016.49

Qammaz, A., & Argyros, A. (2019). MocapNET: Ensemble of SNN Encoders for 3D Human Pose Estimation in RGB Images. https://users.ics.forth.gr/~argyros/mypapers/2019_09_BMVC_mocapnet.pdf

Qammaz, A., & Argyros, A. (2023). A Unified Approach for Occlusion Tolerant 3D Facial Pose Capture and Gaze Estimation using MocapNETs. In IEEE/CVF International Conference on Computer Vision Workshops (AMFG 2023 - ICCVW 2023), Paris, France, October 2023 (pp. 3178-3188). IEEE. https://users.ics.forth.gr/~argyros/mypapers/2023_10_AMFG_Qammaz.pdf

Rubin, D. (1993). Discussion Statistical disclosure limitation. Journal of Official Statistics, 9(2), 461–468. https://www.scb.se/contentassets/ca21efb41fee47d293bbee5bf7be7fb3/discussion-statistical-disclosure-limitation2.pdf

Wood, A., Altman, M., Bembenek, A., Bun, M., Gaboardi, M., Honaker, J., Nissim, K., O'Brien, D., Steinke, T., & Vadhan, S. (2018). Differential Privacy: A Primer for a Non-Technical Audience. SSRN Electronic Journal. https://doi.org/10.2139/ssrn.3338027