AISB AI, Consciousness and Ethics Conference Submission

Mapping Conscious States with AI

300 WORD ABSTRACT

Using advanced AI research methods, this work investigates how signals in the brain can be used to create an anatomically grounded representation of emergent neural patterns in 10 experimentally defined states of consciousness. From a large dataset consisting of multichannel EEG and MEG signals that capture lobe-specific activity, a comprehensive set of static, stochastic and dynamic complexity features was extracted to characterise the underlying trends associated with each state. A hybrid supervised model was trained to distinguish lobe-level activation patterns to 93% accuracy, demonstrating distinct and learnable neural signatures for each state of consciousness.

Local and global subpopulations were established with a hybrid clustering pipeline allowing detection of coherent signal clusters that shared similar complexity features, as well as the identification of outlier patterns that may correspond to transitional or unstable neural configurations. A multi-layer cellular automata was developed to simulate the temporal evolution of neural activity in each state, incorporating biologically inspired mechanisms including latent-space similarity metrics, synaptic weightings and spatial grounding based on anatomical adjacency and lobe-level modulation. Together, these components enabled the model to generate plausible trajectories of state-specific neural propagation, offering a mechanistic framework for understanding how distributed patterns of activity in the brain emerge and evolve over time within different states of consciousness.

This work has significant implications in challenging negative attitudes towards AI, and guiding the direction of research towards improving mental health diagnosis and understanding the difference between biological consciousness and AI representations, while ensuring that such technologies are developed responsibly, with attention of lived personal experiences of individuals whose neural activity they seek to model. The ability to classify states of consciousness, identify subpopulations from neural activity, and simulate neural dynamics has potential in clinical care, communications systems and governance of organisations, including privacy, autonomy and the objective interpretation of subjective experiences.

1000 WORD ABSTRACT

Understanding consciousness is a challenging scientific project spanning neuroscience, artificial intelligence, philosophy of mind and ethics. Recent advances in machine learning and computational neuroscience have enabled new approaches to characterising neural correlates of conscious states, offering both empirical insight and

Charlotte Patmore

conceptual tools about artificial systems. This project investigates how multichannel EEG and MEG recordings can be used to construct anatomically grounded representations of emergent neural patterns across 10 experimentally defined states of consciousness, and explores the ethical implications of such modelling for future AI systems designed to interface with, interpret, and support human cognition.

The dataset, comprised of high-resolution recordings, captures both global and local lobe-specific activity. From these signals, a comprehensive set of state-stochastic and dynamic complexity features was extracted, including PCI, Lempel-Ziv complexity, autoregressive residuals, cross-correlation metrics among others. These features characterise the information processing landscape associated with each state, reflecting differences in volatility, entropy, complexity and oscillatory structure of an averaged signal in a given state. A supervised model was trained on these features, achieving 93% accuracy in predicting the state of consciousness from lobe-level activation patterns. The result demonstrated that each state exhibits a distinct and learnable neural signature, and that machine learning AI can reliably identify these signatures from biologically meaningful structures found in the signals.

To interpret these structures, the model was trained on averaged signals computed per state and per lobe. Latent vectors derived from a bottleneck layer provided a compact embedding of the neural dynamics, enabling the identification of neural structure both within and across states. These embeddings revealed clusters of states with shared dynamical properties, as well as lobe-specific contributions to global patterns of activity. To further investigate patterns, a hybrid clustering pipeline was developed. This approach enabled the detection of coherent subpopulations, defining neural activity that shared similar complexity features, as well as identification of outlier patterns that may correspond to transitional or unstable configurations. The resulting subpopulations map provides a biologically interpretable decomposition of each state into local and global components.

Building on these findings, a multi-layered cellular automata (CA) was developed to simulate the temporal evolution of neural activity within each state. The CA incorporated several biologically inspired mechanisms: latent space similarity metrics, synaptic weighting systems, spatial grounding and lobe-level modulation. Together these components enabled the CA to generate plausible trajectories of state-specific neural propagation. The emergent patterns align with current neuroscience findings, demonstrating that CA based models can serve as a mechanistic framework for visualisation of emergent and distributed temporal patterns of activity in the brain, and how these evolve over time in different states.

Beyond technical contributions, this work raises ethical considerations. As machine learning becomes increasingly capable of interpreting neural data, questions arise about how systems should be designed, deployed and governed. The ability to classify

states of consciousness, identify subpopulations of neural activity, or simulate neural dynamics has potential in clinical care, communication support for individuals with neurological conditions, and the development of AI systems that interact with human cognitive states. However, these possibilities also introduce risks related to privacy, autonomy and the interpretation of subjective experience.

One ethical dimension concerns the use of neural data to support individuals with impaired communication or cognitive function. Dementia, Alzheimer's disease and traumatic brain injury can disrupt memory, language and self-expression. Computational models capable of identifying neural signatures or mapping emergent patterns may contribute to assistive technologies that help individuals communicate or re-gain agency. Yet applications require consideration of consent, data ownership, and the potential for misinformation of neural signatures. Consciousness is not reducible to neural patterns alone, and any system that claims to infer internal states must be designed with epistemic humility and robust safeguards.

A second ethical dimension refers to broader discourse on AI consciousness. While the models developed in this project do not attempt to create or simulate consciousness, they contribute to the growing body of work that uses computational tools to analyse and understand conscious states. This raises conceptual questions about whether artificial systems could exhibit properties analogous to human consciousness, and what ethical status such systems might hold. The project argues that empirical modelling of human consciousness can inform these debated by clarifying the structural and dynamical features associated with conscious processing, while also highlighting the profound differences between biological and artificial systems with respect to the illusion of real consciousness in AI companions.

A third dimension concerns the societal impact of technologies that interpret or manipulate neural activity. As AI systems becomes more integrated into healthcare, education, and daily life, there is a risk that individuals may defer too readily to algorithmic interpretations of their mental states. Ethical design must therefore prioritise transparency, user empowerment, and the prevention of overreach. Models should support human understanding rather than replace it, and should be developed in collaboration with educated clinicians, care workers, as well as children who are exposed to unregulated AI.

This project demonstrates that machine learning and biologically inspired modelling provides powerful tools for analysing neural signatures of consciousness. The combination of feature-based classification, latent-space analysis, clustering and cellular automata offers a multi-scale perspective on how local and global patterns of activity interact across states. The work also underscores the importance of grounding technical advances in ethical reflection. Researchers must ensure that such

technologies are developed responsibly, with attention to lived personal experiences of individuals whose neural activity they seek to model.

The findings represent an early step toward computational frameworks that integrate neuroscientific insight with ethical design principles, extending to how computational representations of consciousness can support individuals with neurological conditions. By combing rigorous empirical modelling with a commitment to ethical responsibility, this research aims to contribute to a future in which AI systems enhance human wellbeing while respecting the complexity and dignity of subjective conscious experience.

This work aims to rectify negative attitudes towards the evolution of AI, especially prominent in the realm of consciousness, and guide the direction of research towards improving mental health diagnosis and understanding the difference between a biological consciousness and an AI representation.