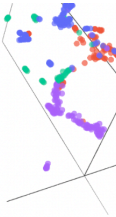


Are we trying to figure out if AI is conscious?

My aim in this article is to examine the field of consciousness research with an analysis of 78 of the most important papers in the field. In doing so, I compare the empirical foundations these disciplines rely upon and detail how each approaches the measurement of conscious experience.

There is no consensus on how to define consciousness, and thus no single discipline dedicated to its study. Instead, thinkers across all fields have developed separate theoretical frameworks and thought experiments, which often remain isolated. To better understand consciousness research and find clusters of insights, I built an instrument to employ large language models—o1, Sonnet, and Gemini—to categorize any paper selected, and extract 20 key facts from it. These facts are supported by evidence and context from the paper using a recursive prompting technique designed to generate accurate and structured output. The instrument then automatically creates a vector representation of the facts using OpenAI embeddings, which mathematically identify both interdisciplinary and intradisciplinary links

	Field	Group
Artificial Intelligence and Computational Models	Group 1	-
Cognitive Neuroscience	Group 2	-
Cultural, Spiritual, and Anthropological Perspective	Group 2	-
Neurobiology	Group 2	-
Philosophy of Mind	Group 3	-
Physical Theories of Consciousness	Group 4	-
Psychedelics and Altered States of Consciousness	Group 2	-
Psychology and Experimental Research	Group 2	-
Quantum Theories of Consciousness	Group 4	-
Unknown	Group 1	-



Fact: Fact 1
Statement: Anesthetic agents block consciousness and memory while preserving non-conscious brain activity, enabling modern surgical procedures.
Evidence: The abstract and introduction highlight that anesthesia selectively and reversibly impairs conscious experiences without disrupting non-conscious brain functions. This selective action has been pivotal in allowing complex surgical interventions to be performed safely on patients by ensuring they remain unconscious and do not remember the procedure.
Field: Neurobiology
Link: <https://www.nature.com/articles/s41598-017-08692-7>

Up to 7 Groups of Fields - Pairwise or Intersection-Average

Enter fields for each group (comma-separated). Then pick a mode: Pairwise (A vs B) or Intersection-Average across all non-empty groups. We log debug info for the intersection-average approach.

Group A Fields:
Artificial Intelligence and Computational Models

Group B Fields:
Philosophy of Mind

Top K: 1

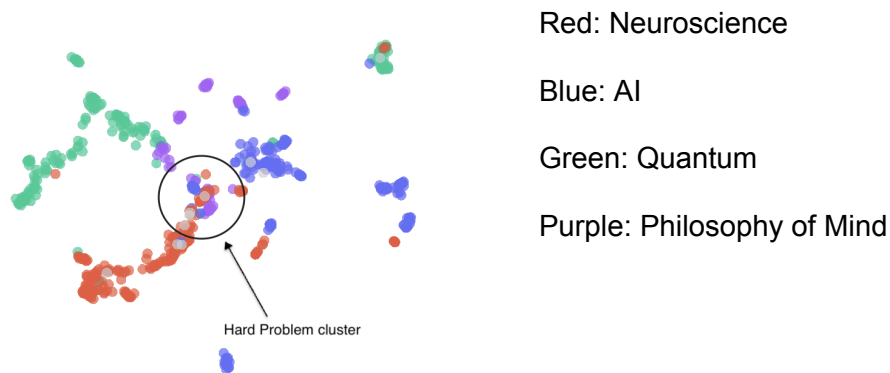
☒ Compare only Group A vs Group B (pairwise)
☐ Intersection-Average (all non-empty groups)

Compute Similarities Download CSV

Similarity	Statement (A)	Evidence (A)	Field (A)	Link (A)	Statement (B)	Evidence (B)	Field (B)
0.7921	Theoretical frameworks on machine consciousness encompass both supporting and opposing arguments, with proponents suggesting possible pathways for AI to achieve consciousness and critics	The article outlines various theories and models related to artificial consciousness, such as the Global Neuronal Workspace Theory, Integrated Information Theory, and embodied cognition. Supporters like Dennett and Goertzel argue for the potential of AI to achieve consciousness through advanced computational models and neural architectures. Conversely, critics like Thomas Nagel and Ned Block argue that subjective experience and phenomenal	Artificial Intelligence Computational Models	https://www.nature.com/articles/s41598-024-04154-3	The possibility of artificial consciousness is a subject of ongoing debate, centered around whether machines can be programmed to possess awareness similar to sentient beings. Key questions include whether consciousness requires biological	The historical notion of Mechanism, which views the mind as a complex machine, supports the idea that artificial awareness might be achievable. However, this perspective has faced significant opposition. For instance, Descartes argued that mechanical processes cannot fully explain the conscious mind, while Leibniz contended that the mind and its perceptions could potentially be constructed through mechanical means. In modern times,	Philosophy of Mind

between the facts. This allows me to retrieve connections across disciplines, with embedding search algorithms, semantic similarity analysis, and an interactive UMAP dimensionality reduction based visualization that turns the 3072 dimension vectors into a 3D map.

A centrally positioned cluster is that of the “Hard Problem of Consciousness”, which was introduced by David Chalmers In 1996. He posed a fundamental question: “How could a physical system such as a brain also be an experienter?”. Through my examination of the primary disciplines involved in consciousness research – Neuroscience, Philosophy of Mind, Quantum Theories of Consciousness, Artificial Intelligence and Computational models – I found that the “Hard Problem” was one of the few ideas explored across each of them having been discussed by prominent researchers such as DeepMind Senior Engineer Murray Shanahan, Nobel Prize winner Francis Crick, and MIT physics professor Max Tegmark.

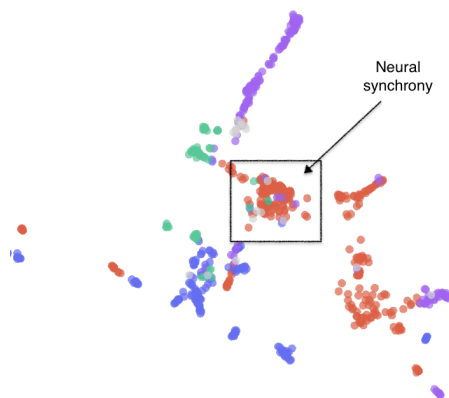


The one idea connecting all of consciousness research is that consciousness research is hard.

This connection point is also a point of contention, with some entertaining it as a thought experiment, most notably in AI and Philosophy of Mind, and some responding to it with criticism. A notable retort came from Salk Institute Neurophilosopher Patricia Churchland, who urged further exploration of the fundamental functions of the nervous system, comparing the “hard problem of consciousness” to any other neurobiological puzzle.

Patricia Churchland's approach exemplifies the empirical methodology of neuroscience, focusing on measuring consciousness rather than exploring it through thought experiments. Churchland's Salk Institute colleague and winner of the 1962 Nobel Prize in Medicine, Francis Crick, spent much of his career attempting to solve the neurobiological problem of consciousness. As an undergrad research intern at the Salk Institute, I was introduced to Crick's concept of the Neurobiological Correlates of Consciousness (NCC), the brain activity patterns directly associated with consciousness.

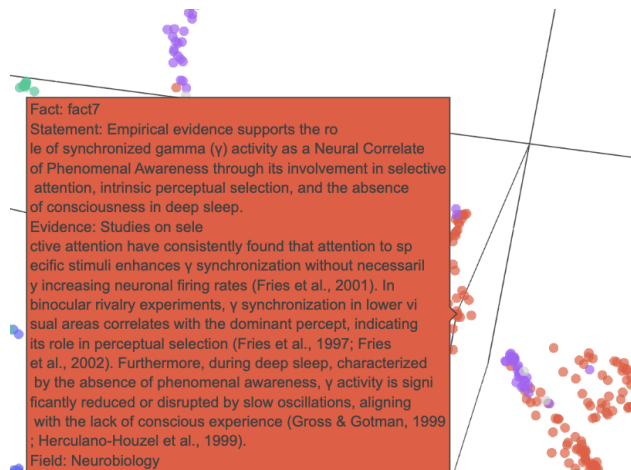
This study of the NCC has pushed our understanding of consciousness past armchair speculation, enabling empirical investigation through techniques such as electroencephalography (EEG). This approach allows us to explore the role of synchronized neural oscillations in the beta and gamma frequency bands where studies suggest that synchrony of large neural oscillations (brainwaves) may be critical for the processing of diverse perceptual features such as color, shape and motion into a unified conscious percept. They propose that consciousness arises when spatially distributed neurons fire in synchrony.



Red: neurobiology

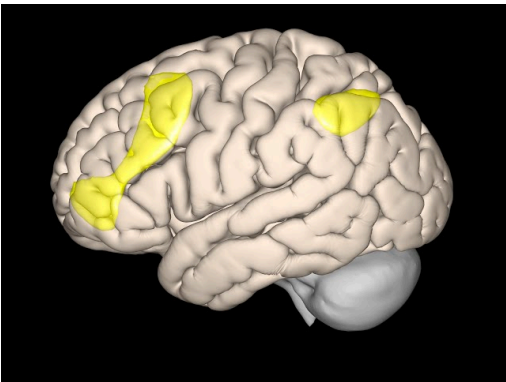
It can be observed that synchronized or orderly action is a core idea in the biological studies of consciousness, bridging a number of the major theories. The observations that synchronised firing of neurons are associated with consciousness lend support to the Entropic Brain Hypothesis (EBH), which suggests that waking consciousness promotes orderly states

within the brain, reducing entropy. Although formally introduced by psychopharmacologist Robin Carhart-Harris in 2014 based on his studies on psychedelics, EBH may trace its roots to Erwin Schrödinger's 1944 contention that reducing entropy was the object of life. Within the cluster of facts related to synchrony and order, the experimental evidence includes TMS and EEG research demonstrating how synchronized neuronal firing can reflect lower-entropy states, studies indicating that conscious and unconscious brain states can be distinguished and investigations into psychedelic compounds that indicate an increase brain entropy. This is shown by a rise of signal variance in neuroimaging studies, leading to perceived alterations in ego and conscious experience.



Adding to neural synchrony and negative entropy are studies that examine feedback mechanisms in conscious perception. Researchers have demonstrated that an initial feedforward sweep of sensory information can register details without conscious awareness; however, the presence of feedback from higher-order areas to sensory regions of the brain leads to conscious experiences. Consistent with this view, Global Workspace Theory (GWT) highlights how specific brain networks become active when a stimulus achieves conscious access. The activation of these brain regions effectively broadcasts information across the brain,

enabling the cognitive processes associated with consciousness in such as introspection, planning, and verbal report.



These theories demonstrate an empirical approach to the study of consciousness. They operate on a reasonable assumption that humans are conscious, and approach the study from a human biological perspective. They, however, need to be combined with the other disciplines to discover the true nature of consciousness. To measure the exchange of ideas between the fields, I computed the mean vector of the embeddings in each field and found their midpoint. I then measured the similarity scores between each field:

Similarity scores	Cognitive & Biological Sciences	Artificial Intelligence and Computational Models	Quantum Theories of Consciousness	Philosophy of Mind
Cognitive & Biological Sciences	1.000000000000000	0.8072824007033620	0.8180942295070520	0.7750874212055440
Artificial Intelligence and Computational Models	0.8072824007033620	1.000000000000000	0.7093312052680860	0.7990431303231130
Quantum Theories of Consciousness	0.8180942295070520	0.7093312052680860	1.000000000000000	0.7897846113052250
Philosophy of Mind	0.7750874212055440	0.7990431303231130	0.7897846113052250	1.000000000000000

The similarity between AI and Cognitive and Biological Sciences is 0.8 is equal to that between AI and philosophy of mind. This provides an interesting insight into the approach AI consciousness

researchers have taken to the discipline. The question of whether AI can be conscious has become a matter of debate. Patrick Butlin and Robert Long's seminal paper, *Consciousness in Artificial Intelligence: Insights from the Science of Consciousness*, provides a comprehensive survey of consciousness research from numerous theoretical perspectives and concludes that current AI technologies are unlikely to be conscious, but note it is not impossible, in principle, for an artificial system to achieve consciousness.

This conclusion has ushered in some empirical research but the field of AI consciousness remains largely theoretical, characterized by thought experiments and conceptual frameworks rather than concrete investigations into the Artificial Correlates of Consciousness (ACC). The study of information processing inside neural networks is itself an emerging field, making the investigation of AI consciousness even more novel. Nevertheless, these early theoretical foundations and research directions will likely shape the field's future development in significant ways.

A December 2024 publication in Nature provides an empirical framework and approach to testing whether GPT-3 might exhibit consciousness by administering adapted cognitive and emotional intelligence tests. This paper, instead of *discussing the opposing positions of Searle and Wittgenstein on language*, operated under the assumption that the meaning of the word is an activity, rejected the *Chinese Room* (Computer can write Chinese but doesn't understand) argument and tested whether GPT-3 might exhibit consciousness by administering adapted cognitive and emotional intelligence tests.

This method highlights a shift in approach, moving towards a more human view of AI systems and their study. Murray Shanahan, in his paper *Simulacra as Conscious Exotica* critiqued an anthropomorphised view of AI due to concerns of misrepresentation of their capabilities and understanding yet I find the distinction to be dramatized. An example being my professor Terrence Sejnowski's 2024 paper *Transformers and cortical waves: encoders for*

pulling in context across time which describes a transformer-like architecture in the human brain demonstrating an attention-like mechanism.

Exploring the similarities and differences between humans and LLMs can illuminate unexplored areas of consciousness or lead to new capabilities in AI systems. Drawing on neurobiology's fundamental assumption that other humans are conscious, I speculate that these investigative strategies could be applied to the study of consciousness in artificial systems. Low similarity(0.7) between AI and the physical theories of consciousness also represent a potential area for further research potentially building on findings that THz radiation and ultrasound can alter membrane permeability and affect consciousness in humans as well as semiconductor logic calculations. Despite being a young field, AI consciousness research could help connect the main disciplines in the space, offering a computational approach to consciousness research.

"Computational models are essential for unraveling the complexities of the brain, allowing us to simulate and understand neural mechanisms in ways that purely experimental approaches cannot."

— Terrence Sejnowski