AI's Pursuit of Consciousness and Ethical Challenges

Nikolaos Koutsis[1]

Supervisor-Prof. Paraskevi Papadopoulou[1]

BMS 4650 Capstone in Biomedical Sciences

[1]Department of Science and mathematics, The American College of Greece, Athens, Greece

May, 2025

Introduction

What Is AI?

AI has already been established as one of the most revolutionary technologies of the 21st century. **It is defined as intelligence held by machines or software in the form of code and algorithms, rather than by biological organisms.**

The History of AI

AI was popularized by British mathematician **Alan Turing**. AI only saw real progress in the 90s with the creation of **neural networks and machine learning.**

Recent developments have amplified questions regarding AI potentially achieving a **conscious state**.

The Field of Consciousness

Phenomenal consciousness is defined as how an experience feels like. Current research focuses on the development of **Theories of Consciousness (ToCs)**. ToCs attempt to create a model of a system from which consciousness could emerge and are typically grouped into three major categories.

AI Ethics

AI ethics delves into the **social, political, and psychological impact** of AI systems including any ethical questions that arise should they exhibit consciousness. Given that **different ToCs** have **different applicability** to AI, they also have **different compatibility** with the ethical aspect of AI consciousness.

Aims and Research Questions

Aim: To contribute to current attempts to develop Artificial Intelligence as well as promote its regulation and ensure compliance with ethical parameters.

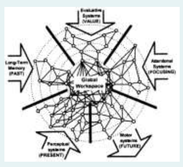
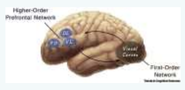
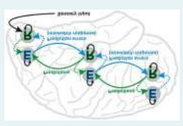
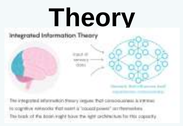
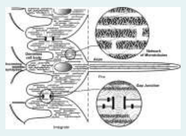
Q1: What ethical considerations and implications arise when developing AI systems that exhibit signs of consciousness?

Q2: How do these considerations and implications align with the current theories of consciousness?

Methods

- Literature Review (Traditional- Generic- Yet followed the Systematic Literature Review methodology)
- Data collection, search terms, inclusion/exclusion criteria, data analysis/synthesis
- Prisma diagram included

Results

Theory	Paper Description/Key findings	Indicative Studies
Global Workspace Theory	Bernard Baars introduces the Global Workspace Theory in 1993.	(Baars, 1993)
	Introduction of Global Neuronal Workspace Theory, an extension of GWT.	(Mashour et al., 2020)
	A comparison of GNW predictions with experimental studies of conscious sensory perception.	(Dehaene et al., 2011)
Higher-Order Theory	Rosenthal introduces Higher-Order Theory.	(D. M. Rosenthal, n.d.)
	A review of different higher-order theories.	(Gennaro, n.d.)
	This paper argues that HOT theory is viable outside philosophy by providing empirical evidence and responding to common objections.	(Lau & Rosenthal, 2011)
Predictive Processing Theory	This paper introduces Predictive Processing as a consciousness theory by combining it with the free energy principle.	(Friston, 2010)
	This paper uses Predictive Processing theory to identify neural correlates of consciousness.	(Hohwy & Seth, 2020)
	This paper is an overview and critique of Predictive Processing that claims that the theory fails to adequately explain consciousness.	(Marvan & Havlík, 2021)
Integrated Information Theory	Tononi first proposes the Integrated Information Theory.	(Tononi, 2004)
	This paper is an overview of IIT focusing on brain structures.	(Tononi, 2017)
	This paper criticizes IITs implication that various systems that exchange information are conscious.	(Merker et al., 2022)
Orchestrated-Objective Reduction Theory	One of the first papers in which Stuart Hameroff introduces Orchestrated Objective Reduction.	(Hameroff, 1998)
	A recent follow-up by Hameroff and Penrose in light of more recent developments and criticisms.	(Hameroff & Penrose, 2014)
	A study that experimentally supports Orch-OR by establishing that certain anesthetic gases bind to microtubules in neurons and dampen their quantum computation, potentially contributing to unconsciousness.	(Khan et al., 2024)

Results (cont.)

Theory	Description of Theory	Indicative Studies
Global Workspace Theory	Consciousness is born from the broadcast of a perception from a single processor to a network of processors.	(Baars, 1993)
Integrated Information Theory	Consciousness is defined as the maximum amount of integrated information a system possesses.	(Tononi, 2004)
Higher-Order Theory	A mental state becomes conscious when there exists a higher-order thought targeting it.	(D. M. Rosenthal, n.d.)
Predictive Processing Theory	Higher-level predictions are compared with lower-level incoming stimuli to minimize prediction error.	(Friston, 2010)
Orchestrated-Objective Reduction Theory	Consciousness originates from quantum processes within neuron microtubules.	(Hameroff & Penrose, 1996)

Ethical Considerations	Ethical Implications	Indicative Studies
Should conscious AI be treated as a **moral patient**?	Should conscious AI be treated like other conscious beings?	(Farisco, 2024)
	Should conscious AI be granted rights such as freedom and privacy?	(Kiškis, 2023)
Can conscious AI behave as a **moral agent**?	Conscious AI will have more applications and perform tasks that are beneficial to society.	(Esmaeilzadeh & Vaezi, 2022)
	Conscious AI can create new problems or exacerbate current ones.	(Seth, 2024)

Discussion-Conclusion

To determine the compatibility of ToCs with ethical considerations/implications of AI consciousness, the applicability of each theory on AI must first be considered.

- GWT, HOT, and PP all **consider consciousness the result of a classical physical model**, making it only a matter of time until **technology is able to artificially replicate this physical system** and create consciousness.
- IIT treats **consciousness as an irreducible, fundamental aspect to reality.** It is **unclear whether artificial systems would be able to replicate** the features necessary to achieve **consciousness**.
- Orch-OR proposes that the key to **consciousness lies in the manipulation of quantum phenomena** making it a strong contender for achieving AI consciousness since **technology has already managed to use the principles of quantum mechanics** to perform computations.

Theories of Consciousness	Ethical Aspects of AI	Compatibility Status
Global Workspace Theory	Moral Agency	Compatible
	Moral Patienthood	Compatible
Higher-Order Theory	Moral Agency	Compatible
	Moral Patienthood	Compatible
Predictive-Processing Theory	Moral Agency	Compatible
	Moral Patienthood	Compatible
Orchestrated-Objective Reduction Theory	Moral Agency	Compatible
	Moral Patienthood	Compatible
Integrated Information Theory	Moral Agency	Unclear
	Moral Patienthood	Unclear but probable

Impact and Recommendations

The extremely high compatibility of ToCs with ethical considerations and implications of AI consciousness is a call for reevaluation of AI rules and regulations. Current efforts such as the EU AI Act have been shown to be insufficient. This project suggests that AI development does not continue unchecked.

References

- Baars, B. J. (2005). Global workspace theory of consciousness: Toward a cognitive neuroscience of human experience. *Progress in Brain Research*, *150*, 45–53. https://doi.org/10.1016/S0079-6123(05)50004-9
- Friston, K. (2010). The free-energy principle: A unified brain theory? *Nature Reviews Neuroscience*, *11*(2), 127–138. https://doi.org/10.1038/nrn2787
- Hameroff, S. (1998). Quantum computation in brain microtubules? The Penrose–Hameroff 'Orch OR' model of consciousness. *Philosophical Transactions of the Royal Society of London. Series A: Mathematical, Physical and Engineering Sciences*, *356*(1743), 1869–1896. https://doi.org/10.1098/rsta.1998.0254
- Kuhn, R. L. (2024). A landscape of consciousness: Toward a taxonomy of explanations and implications. *Progress in Biophysics and Molecular Biology*, *190*, 28–169. https://doi.org/10.1016/j.pbiomolbio.2023.12.003
- Lau, H., & Rosenthal, D. (2011). Empirical Support for Higher-order Theories of Conscious Awareness. *Trends in Cognitive Sciences*, *15*, 365–373. https://doi.org/10.1016/j.tics.2011.05.009
- Metzinger, T. (2021). *Artificial Suffering: An Argument for a Global Moratorium on Synthetic Phenomenology*. https://doi.org/10.1142/S270507852150003X
- Tononi, G. (2004). An information integration theory of consciousness. *BMC Neuroscience*, *5*(1), 42. https://doi.org/10.1186/1471-2202-5-42