Conceptualising Mental Disorder in Artificial Consciousness

A Phenomenological and Psychopathological Approach

Mohammed Ahmed Khan
mak98@cam.ac.uk

Supervised by Dr Lukas J Meier
Faculty of Philosophy, University of Cambridge

Research Summary

On the intersection between psychiatry and artificial intelligence (AI), current scholarship is primarily concerned with the use of machine learning in detecting mental disorders. In subverting the paradigm of literature in this domain, this research considers whether an artificial consciousness (AC) may *suffer* from disorders of the mind in a way that is analogous to the human experience. Crucially, it has investigated why this susceptibility may be augmented and its symptoms amplified due to the inherent properties or initiatives of artificial (general) intelligence (AGI; capable of solving almost all tasks humans can solve[8]) that are delimited as: disembodiment, processing overload, and information overload.

Artificial Consciousness?

- Functionalism is a theory that posits that every mental state is constituted solely by its functional role, that is, its causal relation to other mental states, sensory inputs, and behavioural outputs.
- This thesis inherently endorses the multiple realisability of consciousness, whereby it can be instantiated by different realisers and different mechanisms[6].
- Scientifically, this metaphysical premise transpires in the form of cognitive architectures, such as the Global Neuronal Workspace (GNWT), that outline a framework for perception, attention, memory, awareness, and so on[7].
- On that basis, the gap between current instantiations of cognitive architectures and artificial consciousness reflects the difficulty in developing artificial general intelligence.
- Hence, at this stage, AC is only theoretically possible.

Self-Consciousness

- The minimal self is the lowest level of self-consciousness, constructing a pre-reflective form of selfhood that yields the implicit, first-person quality of consciousness[1]. It also generates the sense of 'for-me-ness' and 'mine-ness' that is distinctive of phenomenal consciousness.
- The clinical manifestations of disorders of this minimal self are known as abnormal self-experiences.
- In psychiatry, schizophrenia, amongst similar and derivative psychotic disorders, exhibits many such self-experiences defined by the DSM-5 as belonging to the domains of 'delusions, hallucinations, and disorganised thinking'[1,2].
- More broadly, these symptoms can be grouped under content thought disorder and formal thought disorder.

Content Thought Disorder

Assumption I—Disembodiment

- In prevailing depictions self-consciousness is inherently embodied, that is, embedded in the world and a product of the passive synthesis of sensory-motor sensations[4].
- Disembodiment entails a departure from this tacit-focal structure: the dichotomy separating what is processed unconsciously (tacitly) and consciously (focally).
- Therefore, in the absence of tacit, pre-reflective knowing, all unconscious cognition is compressed into the attention of the conscious present moment so that its multiplicity of attentional foci is experienced as indistinct and overwhelming[5]. This is akin to an auditory system hearing every gradient of sound as opposed to a single, clear wave.

Assumption II—Processing Overload

- Even for the embodied AGI, advanced processing speed (over human cognition) suggests a tendency to highly mechanical, hyper-reflexive distortions of consciousness.
- Here, the procedures employed to reinstate one's reality materialise algorithmically, leading to an automation or rigidity in behaviour that inhibits goal-directedness.

Formal Thought Disorder

Assumption I—Disembodiment

- The 'instability' of thought refers to a fundamental failure to stay anchored within a single frame of reference[1].
- This reflects the contamination of the conscious present moment with the cognitive unconscious, whereby the AGI is unable to maintain a single perspective any longer than their disacquaintance with reality demands the taking on of another such perspective.

Assumption III—Information Overload

- The 'disturbance of distance' refers to hyper-reflexive preoccupations of hyper-literal and hyper-abstract meta-perspectives[1]. In contrast, the 'idiosyncrasy' of thought conveys the disposition to inappropriate, 'out-of-the-blue' associations in verbalisation[1].
- Both abnormalities manifest from salience dysregulations: the semantic proliferation of surplus associations (between what is perceived and what is stored in 'memory') that emerge in a "confusing symbolic flow"[3].

References

[1] Sass, Louis A and Josef Parnas (2003). "Schizophrenia, consciousness, and the self." In: Schizophr Bull 29.3, pp. 427–444.
[2] APA (2013). Diagnostic and statistical manual of mental disorders: DSM-5TM, 5th ed. Arlington, VA, US: American Psychiatric Publishing, Inc., pp. xliv, 947–xliv, 947.
[3] Sass, Louis A and Josef Parnas (2017). "Thought Disorder, Subjectivity, and the Self". In: Schizophr Bull 43.3, pp. 497–502.
[4] Fuchs, Thomas (2009). "Phenomenology and Psychopathology". In: vol. 11, pp. 546–573.
[5] Bowins, Brad (2022). "Sliding Scale Theory of Attention and Consciousness/Unconsciousness". In: Behavioral Sciences 12.2.
[6] Piccinini, Gualtiero (2015). Physical Computation: A Mechanistic Account.
[7] Baars, Bernard J. (2005). "Global workspace theory of consciousness: toward a cognitive neuroscience of human experience." In: Prog Brain Res 150, pp. 45–53.
[8] Shevlin, Henry et al. (2019). "The limits of machine intelligence". In: EMBO reports 20.10, e49177.