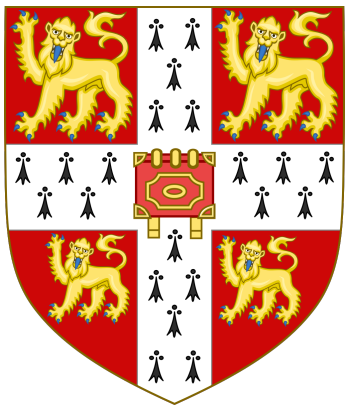
Conceptualising Mental Disorder in Artificial Consciousness

A Phenomenological and Psychopathological Approach



Mohammed Ahmed Khan
mak98@cam.ac.uk

Supervised by Dr Lukas J. Meier

Faculty of Philosophy
University of Cambridge

September 2023

LAIDLAW SCHOLARS

Contents

1 Introduction

On the intersection between psychiatry and artificial intelligence (AI), current scholarship is primarily concerned with the use of machine learning in detecting mental disorders, such as psychotic disorders via speech patterns, anxiety disorders based on attachment style, and affective disorders using physiological biomarkers (Ray et al., 2022). Beyond diagnostics however, as a mode of treatment, conversational agents have also been used, perhaps most notably in delivering psychotherapy to patients with depressive disorder (Vaidyam et al., 2019). The behaviour and cognitive abilities—intelligence—that is understood to permit such applications is said to belong to 'weak AI', in that it "enables us to formulate and test hypotheses in a more rigorous and precise fashion"; being "merely a tool for the study of the mind" (Searle, 1980). This is in contrast to 'strong AI', the thesis of which postulates the theoretical existence of a computer that "really is a mind", which is an indirect claim on said computer also being conscious (Searle, 1980; Chalmers, 1996). In subverting the paradigm of literature in this domain, we will now consider as the object of this research, whether an artificial consciousness (AC) may suffer from disorders of the mind in a way that is analogous to the human experience. That is, to underscore the sole focus on 'strong AC' as adapted from Searle's terminology: an approach to the design of systems whose implementations are sufficient for consciousness, beyond simply representing an understanding, a simulation, that is neither necessary nor sufficient in realising it (Holland, 2003; Chrisley, 2008).

Prima facie, the notion of artificial consciousness seems far removed from the capabilities of present day applications of AI, and certainly, the development of such an artefact is impinged by a lack of understanding of how consciousness arises in human beings to begin with. To that end, this essay will first attempt to capture popular philosophical thought and scientific invention related to the objectives of AC. This serves to elucidate whether instantiating consciousness artificially is plausible, and understand how such an instantiation might be depicted in technical terms. In making a series of implicit assumptions about artificial intelligence, is it then possible to apply empirical observations of mental disorder in humans in order to propose conceptual interpretations of similar phenomena in AC. The essence of this research, however, is in making a set of additional, explicit assumptions drawn from the inherent properties or initiatives of AI (as producing behaviour not observed in humans), to demonstrate both why its susceptibility to mental disorder may be augmented, and why the symptoms it might experience be amplified.

2 Artificial Consciousness

2.1 What is Consciousness?

The definition of consciousness is a contested one, not only across different domains, such as neuroscience, cognitive psychology, and the philosophy of mind, but within these domains as well. While it is often considered as involving perception, sensations, feelings, mental states such as thoughts, beliefs, desires, and so on, closer inspection reveals a more fundamental quality that appears most tacit without also being a tautology. A useful distinction for this purpose is that of between phenomenal consciousness and access consciousness; P- and A- consciousness respectively

(Block, 1995). The former is defined as experience simpliciter: P-conscious mental states have experiential properties that demarcate the 'what-it-is-like-ness' of those mentioned thought-to-be aspects of consciousness. This characterisation captures the often considered 'ineluctable' notion of phenomenal experiences, which are subjective and originate from a first-person perspective that represents an immediate givenness of one's reality as it appears. In contrast, mental states belonging to A-consciousness are such that representations of their contents—information acquired through interoception and extereoception—are used as a premise for reasoning and rational control of action. Block maintains that these two modes of consciousness, while distinct, are not disjoint; the contents of A-conscious mental states can influence that of P-conscious mental states in the sense that perceptual or conceptual information that is being concurrently accessed outside of experience can become experienced. This is to emphasise the fact that while only one is phenomenal, the mental states of both aspects of Block's delineation of consciousness are intrinsically representational, that is, they have contents and are about something (Block, 1995; Bourget and Mendelovici, 2019). It is this 'about-ness' or directedness towards objects and properties, ascribed to all mental states, that is known as intentionality—regarded by many as the characteristic 'mark of the mental'.

The relevance of intentionality is made clear by being more precise about the thesis of 'strong AI' Searle's claim, a refutation of this thesis, is that "a machine cannot possess intentional mental states". That is, it lacks intentionality and is thus devoid of a 'mind'. This is demonstrated by the Chinese room thought experiment, in which an English-speaking person with no knowledge of Chinese is asked to respond to questions in the language just as fluently as a native speaker would (Searle, 1980). Given a set of Chinese scripts and an accompanying set of rules written in English, it is possible for the person to do so with enough accuracy that a fluent speaker of Chinese may indeed believe the English-speaker too speaks their language. This behaviour, however, is not afforded by an understanding of the language itself, but rather a comprehensive set of rules that allow the English-speaker to successfully manipulate symbols to which they ascribe no meaning to. Likening these rules to a computer program, Searle argues that while a machine may behave[1] as if possessing conscious mental states, the absence of understanding—intentionality—necessities an absence of consciousness (Chalmers, 1996). In more formal terms, if programs are purely syntactic, using syntax to manipulate symbols that are physical objects but not representational; conscious minds have mental states whose contents are always semantic and thus representational; and syntax alone is neither constitutive of nor sufficient for semantics; then programs are neither constitutive nor sufficient for minds (Searle, 1990). Moreover, if brains cause minds through causal powers[2]; then trivially, any other system capable of causing minds must possess causal powers at least equivalent to the brain's; and thus it holds not only that no artefact can produce mental phenomena by simply running a program (since one does not produce minds and so lacks equivalent causal powers), but the brain which does produce mental phenomena cannot be solely by virtue of running a program. (Searle, 1990).

[1]This treatment of consciousness reflects a behaviouralist analysis which claims that if a system behaves as if it were conscious, then it is conscious.

[2]This is a view substantiated by the philosophical position Searle subscribes to known as biological naturalism. Searle argues that the mental phenomena that constitute consciousness are dependent on the physical properties of the human brain. Precisely, it is the neural correlates of consciousness that possess the causal powers permitting phenomenal experience as we know it.

This is what Searle recognises as the 'characteristic mistake' in the study of consciousness: to ignore the subjectivity of phenomenal experience and treat it as objective third-person phenomena as according to functionalism and computationalism. The former is the thesis 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. In that sense, mental states make a theoretical abstraction over physical implementation which is concerned only with yielding said states through the effective organisation of functions. The latter perspective represents a family of views that hold that the mind is an information processing system, and that cognition and consciousness together are a form of computation on the premise that mental states themselves are innately computational. In considering such a theory, it is argued that "computational models of consciousness stand to consciousness in the same way the computational model of anything stands to the domain being modelled", in that, for instance, no computational model of a weather event necessitates the occurrence of that weather event (Searle, 2002). Crucially, the acceptance of the 'strong AI' thesis through such theories of mind rests on the dispute of the claim that syntax cannot have semantics—with many arguing that this is indeed possible if given the right causal structure.

2.2 What are Qualia?

The significance of semantics in mental states is in its coinciding with phenomenal experience, a concept that requires further discussion to establish its purported metaphysical implications. In philosophical literature, this characterisation of experience is referred to as qualia, and an instance of it (a quale) is more carefully specified as: ineffable, it cannot be communicated or apprehended by any means other than direct experience; private, all interpersonal comparison with other quales is systematically impossible; and directly or immediately apprehensible by consciousness[3] (Dennett, 1988). Therefore, since A-conscious mental states must not be quales, a P-conscious mental state is only so if there is something that it is like have it, which is conceptualised explicitly as the subjective character of that experience (Nagel, 1974). In some sense, phenomenological facts, Nagel argues, are perfectly objective since one can speak of the quality of another's experience of the same perceptions. They are subjective, however, in that "even this objective ascription of experience is possible only for someone sufficiently similar to the object of ascription to be able to adopt their point of view". Thus, if the facts of experience are accessible only from one point of view—private, as understood by Dennett's definition of qualia—then it is unclear how the true character of experience can be revealed in the physical operation of that entity, a "domain of objective facts that can be observed and understood from many points of view", i.e., not ineffable. In other words, it is difficult to understand what is meant by the objective character of experience apart from the particular point of view from which its subject apprehends it. The foremost presentation of this depiction of qualia is through Nagel's eminent question, 'what is it like to be a bat?', to which he concludes, a subjective experience known only by the bat. To that end, this theory of mind appears to advocate for the conception of

[3]That is, to experience a quale is "to know one experiences a quale, and to know all there is to know about that quale".

mental phenomena as extending beyond a purely physical understanding of reality[4].

Determining, then, why and how physical properties (such as of the brain) give rise to the subjective character of experience is known as the explanatory gap, and the problem of bridging it is known as the hard problem of consciousness (Levine, 1983; Chalmers, 1996). In Chalmer's own words, why, after all the behavioural and cognitive functions—the functional facts—within the vicinity of experience are accounted for and explicated, there "may still remain the unanswered question of why and how the performance of these functions is accompanied by experience", and a particular experience at that? Proponents of the hard problem argue that it is hard as such because there exists no functional explanation, pertaining only to the causal structure of the world, that might elucidate this subjective character[5]. More precisely, a quale cannot be reductively explained by appealing to its microphysical constituents; it is irreducible to lower-level physical facts. Therefore, it must not be purely physical and so the explanatory gap is necessarily a metaphysical one.

2.3 The Mind-Body Problem

Understanding the nature of this relation between physical and non-physical facts is referred to as the mind-body problem. Traditional Cartesian metaphysics offers a dualist view of reality that posits two distinct and independent substances or categories of existence: one encompassing all mental phenomena, such as qualia, and the other, all physical phenomena, such as the brain. Asserting that mind and matter exert causal effects on each other (as an instance of interactionist dualism), produces the preeminent ontology known as substance dualism. While it has provided a fundamental and systemic conception of mind and matter for many generations of thought across many academic domains, substance dualism is no longer widely accepted in modern analysis of this subject. Chiefly, criticism of this doctrine is directed towards the inadequate explanation of the causal interaction between mind and body, known as the problem of mental causation; how does the mechanism of this interaction take place, and where exactly does it take place? Born out of this objection (in addition to a disbelief of a non-physical reality that violates the laws of physics) is substance monism. This theory makes the diametric metaphysical assertion that the apparent plurality of substances is due simply to the differing of states of a single, fundamental substance that wholly constitutes reality. Substance monism's principal ontology, physicalism, makes explicit the claim that everything is or supervenes on the physical through interaction with the material world. On that basis, the explanatory gap is merely an epistemological one in that that our understanding of nature is lacking. It is at this point where derivative monist theories diverge in their particular claims on the ontology of mental phenomena, and so differ in their explanatory power. Strong reductionists maintain the existence of qualia, but argue that they are reducible to material capable of realising it so that it can

[4]The knowledge argument supports this conclusion by proposing a thought experiment in which a scientist, living in a black-and-white world with complete knowledge of the physical properties of a colour, exits that world and experiences colour for the first time. The supposed knowledge they gain upon this perception serves as an additional conceivability argument for this conception of subjectivity as a property that physical description wholly fails to capture (Jackson, 1982).

[5]This is in stark contrast to the easy problems of consciousness which concern how physical systems give beings the ability to provide behavioural outputs from sensory inputs through functional explanations that pertain to the physical structures underpinning such phenomena.

be fully understood in functional terms as an emergent property of said material. Hence, the hard problem of consciousness is not indicative of a genuine ontological gap between mind and matter since every fact about the mind is also one about the performance of material functions, viz., when these functions are accounted for, there is nothing left to be explained. It is precisely the neural correlates of consciousness—the minimal set of neural states that are sufficient for and so accompany a specific conscious experience—that strong reductionists contend give rise to qualia, and other physicalist theories such as biological naturalism find themselves investigating the nature of this reduction more concretely (Bowins, 2022). A far more radical position is adopted by eliminative materialists who reject the existence of qualia altogether, and posit that the mental states used in folk psychology, such as beliefs or desires, correspond directly to material mechanisms of the brain as according to empirical evidence from scientific investigation. That is, mental states realised by the brain lack semantics themselves, and all mental phenomena are inherently meaningless. They contend that the traditional, common-sense understanding of these phenomena is flawed, and advancements in neuroscience and the knowledge of the neurobiological processes that underpin behaviour will eventually lead to their elimination from scientific discourse.

On this account, Chalmers disputes the profound elimination of the phenomenology of subjective experience in a modern rendition of Descartes' famous, "I think, therefore I am". He argues that one has a direct "acquaintance" with consciousness: a reality that is more certain than any other philosophical or scientific apprehension of the matter (Chalmers, 2020). In a similar sense, on the account of reductionist theories, Nagel argues that reduction is a move in the direction of greater objectivity, which itself is the direction in which the understanding of subjective experience travels: far away from a strictly human viewpoint (Nagel, 1974). Abandoning the particularity of the human point of view and striving for a description in terms accessible for beings that could not imagine what that view is like, means that all physicalist hypotheses assume a faulty objective analysis of mind. This inherent failure serves not to render these reductionist frameworks false however, rather, it illustrates the epistemological gap in our understanding of nature (independent of the existence of an ontological gap, or lack thereof) that hinders coherent discussion on what objectivity means in the phenomenological sense *de novo*.

2.4 Instantiating Consciousness Artificially

Crucially, for any theory of mind, the conception of AC is concerned only with its admittance of the multiple realisability of consciousness when defined concomitantly with qualia. As a property, consciousness is multiply realisable if it can be instantiated by different realisers and different mechanisms, i.e., the same mental state can be implemented by different physical states in a one-to-many mapping (Piccinini, 2015). An example of a property that admits this feature is pain, the mental state that is correlated with many different physical states as realised by many different organisms (Putnam, 1960). In this way, a mental state can be explained without considering the underlying physical medium that realises it, be it neural or electronic substrate. Of the theories discussed in § 2.3, functionalism inherently endorses the thesis of multiple realisability by proposing a broad metaphysical hypothesis (on the functional role of mental states) without also making narrowing

ontological claims on the true nature of these mental states (Block, 1980; Manzotti and Chella, 2018). Therefore, although a theoretical physical system capable of giving rise to consciousness may not make the hard problem of consciousness any less intractable by scientific means, it shall certainly leave nothing left to be explained when all of the functions of the mind are accounted for (Manzotti and Chella, 2018).

These systems can be described as the computational instantiation of a cognitive theory of the structure of the human mind (Lieto et al., 2018). This is referred to as a cognitive architecture, outlining a framework for the basic systems and processes involved in cognition, such as perception, attention, memory, and conscious awareness (Lieto et al., 2018). The Global Neuronal Workspace Theory (GNWT) is one such architecture in which consciousness is explicitly modelled as a 'global workspace' that integrates and broadcasts information to distributed processing regions (Baars, 2005). In the abstract, this modelling is of the mechanisms and functional organisation that underly consciousness, and concretely, this is substantiated by two essential dimensions of information-processing computation: global-availability and self-monitoring (Dehaene, Lau, and Kouider, 2017). The former, known as C1, corresponds to the transitive nature of consciousness, that is, the relationship between a cognitive system and a specific object of thought which a mental state is representational of—intentionality, in the philosophical sense. This process selects information, and among the vast repertoire of thoughts that can be conscious, makes only one globally available for computation and report as the contents of C1 or P-conscious mental states (Dehaene, Lau, and Kouider, 2017; Hildt, 2019). In contrast, the latter, C2, corresponds to the reflexive nature of consciousness: the self-referential relationship through which a cognitive system monitors its own processing and obtains information about itself to generate a subjective sense of certainty that resembles A-consciousness. In the psychological sense, this can be likened to introspection or metacognition: the thinking of thinking or internal representations of one's thoughts and knowledge (Dehaene, Lau, and Kouider, 2017).

It goes without saying at this stage, that current instantiations of cognitive architectures remain yet to fully realise 'strong AC'. Conceptual models of consciousness such as the global workspace are, by definition, nothing more than attempted formalisations of human cognition and behaviour represented as computation. For the functionalist, replicating the functional roles of human mental states in this modelling process will be sufficient for instantiating artificial consciousness. For the strong reductionist, it may be the examination and subsequent reconstruction of the neural correlates of this (supposed) computation. On that basis, it can be argued that while this modelling relation—between an ontological thesis on consciousness and an instantiation of it—holds as a result of deeper explanatory properties being shared by the physical realiser and conscious mental phenomena, current such realisations that instantiate these properties are not, alone, sufficient for also instantiating consciousness (Chrisley, 2008). This deficiency in other, necessary, properties reflects the difficulty in delineating the causal relationship between consciousness and intelligence. Indeed, many aspects of the latter have been described as being necessary for the former, but closer investigation is needed to construe how higher-order mental faculties, such as the aforementioned cognitive abilities to which we associate the human mind with, emerge from or are constituted through the synthesis of many, more fundamental aspects, that is, if at all (Baars, 1988). It is in this sense that artificial consciousness is often conceived as dependent upon the

development of artificial general intelligence (AGI) that is capable of solving almost all tasks that humans can solve, and so by which consciousness is regarded as an emergent property of general intelligence (Shevlin et al., 2019; Smith and Schillaci, 2021). While the projects of both AC and AGI can and have been separated, given the scope of this research, it is helpful to naïvely assume a level of intelligence of an artificial consciousness in order to productively cogitate on its susceptibility to mental disorder. Most assumptions will be implied by the analysis of the structures and functions of the mind that give rise to them, while others, related to cognitive abilities extending beyond observed human behaviour, will be made explicit. Therefore, it is imperative in this mission to dismiss the presupposition that the potential artificial intelligence that might be realised in this joint project can be equated to that of which is observed in humans. Moreover, it is delimiting this discrepancy that will shed light on this proclaimed susceptibility.

3 Mental Disorder

3.1 The 'Mental' in Mental Disorder

Mental disorder, like consciousness, is a subject construed differently based on the themes and objectives of the inquiring discipline. As mentioned in § 2.3, modern analysis is no longer concerned with Descartes' substance dualism, a philosophy "potentially [most] pernicious in its effects", and so whose rejection serves as a commitment to the conception of mental phenomena as emerging and being entirely dependent upon brain function as its physical realisation (Kendler, 2005). Abandoning this preconception that has pervaded psychiatry requires a shift in thinking, away from dualistic thought of fundamentally different spheres of reality (Kendler, 2005). This leads to the tautological conclusion that all mental disorders are necessarily founded in their neurobiology, which poses a definitional problem of explaining then, how mentality has a constitutive role in mental disorder without presupposing that the domain of mental states is distinct from that of physical states (Stephens and Graham, 2009). While the proposition that mental disorders have their somatic basis in neurobiology entails that they are disorders of the brain is not necessarily invalid, it does not entail there are no mental disorders any more than the proposition that biological phenomena have their basis in physical phenomena entails that biological life does not exist (Stephens and Graham, 2009). By recognising that the supervenience of the mental on the biological does not discredit the reality of mental disorder, is it possible to appreciate the notion of a purely mental disorder as the failure to perform some function, specified in structural terms (as the production of behavioural outputs given a set of sensory inputs) and not physical terms (Papineau, 1994). That is, where mental states and processes are physically realised differently between different members of a species (beyond what is genetically and thus physically specified), any subsequent failure must therefore be the failure in the performance of some structural function being carried out since there exists no physical description of how exactly this function fails (Papineau, 1994). The Diagnostic and Statistical Manual of Mental Disorders, Fifth Edition (DSM-5), in part, captures this formally in its definition of mental disorder as "a syndrome characterised by a clinically significant disturbance in an individual's cognition, emotion regulation, or behaviour, and reflecting a dysfunction in the psychological, biological, or devel-

opmental processes underlying mental functioning" (APA, 2013). These processes can be considered as operating on A-conscious mental states: the kind that make information available to other processes as well as P-conscious mental states in the production of subjective experience. In some respects, it is through this anchoring of mental disorder in the dysfunction of information processing that the focus of cognitive psychology has shifted away from 'warm-blooded' philosophical reconstructions of consciousness—eclipsed by the study of the 'cold-blooded cognitive' as subsuming the faculties that give rise to it (Radden, 2023). This is coupled with the growing approach of substantiating these explanations by computational means, reflected in the development of cognitive architectures as discussed in § 2.4, by which cognitive processes should then possess the causal powers necessary in artificially instantiating consciousness.

In other respects, the rise in computational psychiatry has also been associated with the reassertion of the reality of psychological processes which possess these same causal powers. That is, the significance of (phenomenal) consciousness in its relation to dysfunctional processes underlying the symptomatology of disorders (for instance, those specified by the DSM-5[6]) that can be delimited as mental by appeal to the traditional categories of faculty psychology, such as perception, affection, cognition, memory, and so on (APA, 2013; Ouwersloot, Derksen, and Glas, 2020; Radden, 2023). This parallels a shift in the conceptual framework of medicine, from the doctor's medical knowledge of disease to the patient's subjective experience. In treating this experience as logically primary, mental disorders become analysable in terms of defective action rather than of defective functioning, while remaining unopposed to computational representation of both warm-blooded mental phenomena and cold-blooded cognitive processes (Fulford, 1993). In fact, it is precisely the careful analysis of the experiential qualities of the former that can expose corresponding dysfunctions in the latter. Indeed, when concerned with a subset of these processes, while the reality and role of dysfunctional, sub-personal information processing mechanisms in mental disorders cannot be denied, asserting that the dysfunction associated with a mental disorder occurs only in "consciously inaccessible activity does not aid in the defence of the categorical distinctiveness of said mental disorders", and in fact downplays the 'subtle anomalies of patients' experiences' (Stephens and Graham, 2009; Stanghellini, 2009; Messas et al., 2018). Therefore, in making a minimal assumption on effectuating cognitive processes in the project of AC, the phenomenological and psychopathological approach shall first confront the P-conscious mental states that underwrite "the capacity for suffering", before construing how this is may be implicated by disturbance in the domain of A-consciousness (Smith and Schillaci, 2021). This entails the study of the essential structures of human experience and the description of psychiatric symptoms that rely on a common-sense view of how one's reality is revealed subjectively in the first-person perspective (Sass and Parnas, 2003). In this paradigm, consciousness

[6]It is worth noting that despite scant commitment to this form of consciousness, offering nor formal definition of it, the DSM-5 still presupposes phenomenological facts about the reality of experience in addressing the states in which the contents of consciousness are subjectively experienced (Ouwersloot, Derksen, and Glas, 2020). Where it does not presuppose these facts, the DSM-5 is distinctly addressing the altered states of consciousness which are necessary but not sufficient conditions for conscious awareness to transpire (Pitts, Lutsyshyna, and Hillyard, 2018). These states, such as the comatose, vegetative, and minimally conscious states, represent a kind of disorder of consciousness that will not be considered in this research.

is a condition of the manifestation of physical phenomena as mental phenomena: creating not the world but the medium, through which the world articulates itself and so by which experience itself is constituted (Sass and Parnas, 2003). It is in focussing less on the contents of this experience than on the form of self-consciousness that invokes it, is this framework useful for conceptualising mental disorder.

3.2 The Dimensions of Self-Consciousness

Self-consciousness is "a multi-dimensional concept, involving as much consciousness of one's own body and image, as self-recognition with consciousness of one's own identity"—analysed as being constituted by different hierarchical levels (Motillon, Keromnes, and Tordjman, 2018). The minimal self is the lowest such level that constructs a foundational, pre-reflective, and tacit form of selfhood (and so on which other levels of selfhood are built) that refers to the implicit, first-person quality of consciousness; viz. the implicit awareness that the subjective character of experience articulates itself in first-person perspective as 'my' experience (Sass and Parnas, 2003; Nelson, Thompson, and Yung, 2013; Nelson, Parnas, and Sass, 2014; Lane, 2020). The ascription of a 'foundational' quality to this selfhood is substantiated by the fact that it is a "consciousness of oneself as an *immediate* subject of experience unextended in time"; the pre-reflective and perspectival "point of origin for [*all*] action, experience, and thought" (Gallagher, 2000; Stanghellini, 2009; Motillon, Keromnes, and Tordjman, 2018). *Ergo*, to speak of this mode of consciousness is to speak directly, non-inferentially, and non-reflectively on these experiences as they occur in the first-person mode of presentation that immediately reveals them as one's own (Sass and Parnas, 2003; Fuchs, 2009). This presupposes the built-in, self-implied self-reference[7] that distinguishes the minimal self as the most primitive self-awareness, and whose perspectival givenness is what makes experience subjective *in primis*. Hence, by way of locutions that convey this subjective character and 'in-the-moment' sense of existence, whilst qualia describe its 'what-it-is-like-ness', and intentionality[8] refers to its 'about-ness', minimal selfhood is concerned with its 'for-me-ness' and 'mine-ness' as fundamental properties of reflexive self-consciousness that are necessarily entailed by phenomenal consciousness (Lane, 2020). This unique feature of the minimal self is described as manifesting from two decoupled, primitive, phenomenological qualities: the sense of agency (SoA) and sense of ownership (SoO). The former is the sense that 'one is the source of one's own actions', while the latter, the sense that 'one is the subject of one's own experiences', so that in their joint presence 'I need not to think to know my thoughts, perceptions, and actions are mine since I know them pre-reflectively' (Gallagher, 2000; Aylak and Uluğ, 2022; Blanke and Metzinger, 2009).

This emphatic emphasis on a *minimal* selfhood is intended to mark a contrast with less atomistic levels of self-consciousness (Lane, 2020). Namely, the level of reflective selfhood, which exhibits an explicit awareness of oneself that is represented as an object of perception, as opposed to the conscious perception *per se* (Martin et al., 2014). If the notion of reflexivity refers to experiences coming be-

[7]This has the feature of 'immunity to error through misidentification', which is known as the immunity principle (Gallagher, 2000).

[8]This is in the traditional sense in the case of the qualia, but is more precisely delimited as 'operative' for minimal selfhood, as is further explained in later sections (Sass and Parnas, 2003).

fore "the moment one attentively introspects", then distinct from this perspective, self-consciousness becomes reflective when intentionality is directed towards experience as presupposing the sense of existing as the very subject of said experience (Martin et al., 2014; Zahavi and Gallagher, 2005; Möller et al., 2021; Smith and Schillaci, 2021; Aylak and Uluğ, 2022; Nelson, Parnas, and Sass, 2014). In this way, the reflective self constructs a consciousness of oneself as an "invariant pole of the individual through multiple experiences, thoughts, and actions", that, unlike its reflexive counterpart, is extended beyond the short-term and maintains continuity (Gallagher, 2000; Motillon, Keromnes, and Tordjman, 2018). It is precisely the temporality of reflective selfhood and the availability of the phenomenal contents of reflexive selfhood, that permits the formation of higher-order representations of mental states, i.e., metacognition, and thus generates the second-person perspective as the conceptual mediation of the first (Kircher and Leube, 2003; Blanke and Metzinger, 2009). This also facilitates the formation of an identity not limited to immediate self-consciousness, and so from which emerges the more elaborate narrative self that is distinguished as forming a coherent and continuous self-representation over time based on autobiographical memory[9] (Sass and Parnas, 2003; Motillon, Keromnes, and Tordjman, 2018; Gallagher, 2000; Nelson, Parnas, and Sass, 2014).

3.3 Minimal Self-Disorders

Establishing an organisation of the dimensions of self-consciousness allows for understanding how it may then become disordered. In the discourse of *mental* disorders, this is by way of investigating dysfunctions that are rooted in the most fundamental selfhood, the minimal self, and propagate to higher levels of self-experience in a bottom-up fashion. For that reason, it is helpful to distinguish between two lenses under which to view these dysfunctions: as disturbances of embodiment and temporality (Fuchs, 2009).

3.3.1 Disturbance of Embodiment

The minimal self, the medium through which all intentional activity is realised, is constituted by the complex interplay and subsidiary awareness of the kinaesthetic, proprioceptive, visual, and vestibular sensations that represent the non-conscious performance of the body (Sass and Parnas, 2003; Fuchs, 2009). Chiefly, it is the body's passive integration of these sensory-motor systems, operant in "every action and interaction", that sets the zero-point constituting one's perceptual view of the world (Fuchs, 2009). This 'about-ness' of non-conscious experience or phenomenal awareness, and directedness to objects prior to conscious reflection, is formalised as Merleu-Ponty's 'operative' intentionality which forms the "background texture of the field of experience" (Fuchs, 2009). Therefore, the body as the agent of a pre-reflective and egocentric perspective, generates a tacit or implicit private knowing that, under Polanyi's analysis, characterises the 'proximal' pole of consciousness that "recedes from awareness" (Fuchs, 2005). This provides a basis for the 'distal' pole that encompasses a focal or explicit shared awareness of objects of perception as

[9]The emergence of this self represents a key component in cognitive development that imbues the individual with social identity and personal history (Nelson, Thompson, and Yung, 2013; Gallagher, 2000; Martin et al., 2014; Hafner et al., 2020).

constituting the conceptual field of attention[10]. In concert with the prior, these poles together form the tacit-focal structure of consciousness as an emergent property of the ecologically-embedded self, that is, embodied[11] (Gallagher, 2000; Fuchs, 2009).

Therefore, a dysfunction as pertaining to the minimal self is one that undermines this tacit-focal structure, and is addressed as an 'ipseity disturbance'[12] of the "experiential self as being a vital and self-coinciding subject of experience" (Sass and Parnas, 2003; Nelson, Parnas, and Sass, 2014). In the ipseity-disturbance model (IDM), this is postulated as consisting of two fundamental and complementary mechanisms that disrupt the intentional activity of awareness[13]. The first, hyper-reflexivity, refers to forms of 'exaggerated self-consciousness' in which "aspects of oneself are experienced as akin to external objects". This is to emphasise the way in which aspects of experience normally tacit, become focal and explicit, leading to a heightened awareness of them or self-alienation (Sass and Parnas, 2003; Nelson, Parnas, and Sass, 2014; Sass et al., 2018; Fuchs, 2009). The second, diminished self-affection, refers to a "weakened sense of existing as a vital, self-coinciding source of awareness and action", i.e., weakened SoA and SoO, as presenting the other side of the same process of hyper-reflexivity whereby "what was once tacit can no longer be inhabited as a medium of once taken-for-granted selfhood"[14] (Sass and Parnas, 2003).

3.3.2 Disturbance of Temporality

A second approach expresses the self as being not just an embodied entity, but one that is temporally present, bound, and demarcated (Venkatasubramanian, 2015). In this depiction, the minimal self corresponds to Husserl's inner time consciousness, an implicit time or temporality that is pre-reflectively lived and renders the intentional orientation to one's immediate situation with its means and goals (Fuchs and Van Duppen, 2017). Here, the stream of consciousness is not the mere succession of now-moments existing in isolation, but the integration of the 'primal impression' given at each moment as presentations, with the awareness of what was just perceived in preceding moments as retentions, and the anticipation of the continuation of said perceptions in proceeding moments as protentions, into the width of presence (Fuchs and Van Duppen, 2017; Fuchs, 2007; Fuchs, 2009). Homologous to the embodiment of oneself, it is the transcendental and passive synthesis of the sequence of single moments into a duration that is spanned by the double-faced attention of retention and protention, that constitutes Merleau-Ponty's 'intentional arc'. This 'arc' is operant in every perception, based on detection of changes, and action, based on motivation and sensory-motor anticipation (Fuchs, 2009; Fuchs and Van Duppen, 2017). In that sense, it also possesses the aforementioned 'operative' intentionality

[10]This focal knowing is distinctly disembodied; in Gallagher's own words, the reflective self "need not be embodied or enactive within an environment" for "it is a kind of self-consciousness that captures this self without being ecologically-embedded" as it instead "operates on a conceptual level already in possession of the concept of self" (Gallagher, 2000).

[11]This primitive self-consciousness houses non-conceptual first-person content: pre-linguistic self-specifying information that is attained through ecological perception of oneself and one's position in the environment (Gallagher, 2000).

[12]A disturbance of the minimal self.

[13]These mechanisms are best conceptualised not as disjoint, but as mutually implicative facets.

[14]These disturbances are necessarily accompanied by alterations or distortions in the subject's grip or hold on the conceptual and/or perceptual field of awareness (Sass and Parnas, 2003).

(of embodiment) that establishes a fundamental continuity and "undercurrent of experience" that enables one to perceive and respond to situational cues, objects, and goals in a meaningful way (Fuchs, 2007). In unity, this retentional-presentational-protentional tripartite structure of inner time consciousness grants the paradigmatic pre-reflective and tacit knowledge by way of SoA and SoO that is granted by indefinite awareness of the past and anticipation of the future as being more immediate than recollection and expectation respectively (Fuchs and Van Duppen, 2017).

Minimal self-disorder construed as a disturbance of temporality emphasises a "decline in the felt temporal flow" that imbues a sense of now-ness in subjective experience (Sass et al., 2018). The undermining of this temporal structure is necessarily a disturbance of the transcendental constitution of inner time consciousness which, in the same vein as 'ipseity disturbance', engenders the diminishment of the sense of self that is concerned with pre-reflective temporal experience (Fuchs and Van Duppen, 2017). Fuchs postulates that the mechanism at the root of this disturbance is the weakening of the protentional function as reflecting a wider disruption of the 'intentional arc'. Put metaphorically, this function is a cone of probability that originates in the present moment and moves forward continuously with increasing possibilities. What is probable in this cone is determined by one's intentions, retentions, and presentations, so that the performance of the protentional function is given by the disactualisation process that selects appropriate proceeding (probable) thoughts, impulses, and associations (Fuchs, 2007). On that basis, damage to the cone implies a failure in this process: in its inhibition of inadequate (improbable) continuations. This fragmentation of the usually continuous implicit flow of time that is accompanied by tacit knowledge results in the emergence of a series of itemised now-moments—a succession of snapshots of the world—that replace the temporal medium of selfhood and subsequently alter one's perceptual field of experience.

3.4 Anomalous Self-Experiences

The clinical manifestations of minimal self-disorders are known as abnormal self-experiences. In psychiatry, schizophrenia, amongst similar and derivative psychotic mental disorders, exhibits many such abnormal self-experiences (Sass and Parnas, 2003; Aylak and Uluğ, 2022). The DSM-5 defines these abnormalities as belonging to the domains of 'delusions, hallucinations, and disorganised thinking' amongst others, which can be more broadly grouped into the category of thought disorders: disturbances in cognition that affect language, thought and communication (APA, 2013).

Like most psychotic disorders, schizophrenia is complex and multifaceted; while neurophenomenological accounts have been successful in developing and employing theoretical constructs to facilitate its conceptualisation as a disturbance of self-consciousness (through embodiment and temporality), it remains yet to be fully understood in terms of its etiology and pathogenesis. That is, psychiatrists lack an understanding of its fundamental nature in terms of the plausible biological underpinnings of this disturbance (Venkatasubramanian, 2015). Crucially, in addition to positing a propensity of an artificial consciousness to a 'generalised' ipseity or temporal disturbance—a minimal self-disorder resembling schizophrenia—owing to dysfunction in the same functions or structures of the mind that constitute self-

consciousness in humans, the following sections aim to demonstrate why this disposition may be augmented and its consequences amplified due to the computational nature of artificial (general) intelligence. This enumeration shall discern between descriptions and explanations for the neurophenomenological account of minimal self-disorders (which are increasingly thought of as the root for the emergence rather than a symptom of schizophrenia), and on that basis, apply empirical observations of anomalous self-experiences in humans to propose conceptual interpretations of self-disorders in AI (Möller et al., 2021).

3.4.1 Content Thought Disorder

The positive syndrome of schizophrenia involves the presence of experiences that are normally *absent*: delusions, thought insertions, and hallucinations. Collectively, they represent the symptoms of disturbances concerned with the content of thought, which are marked by salient abnormal beliefs and convictions. These anomalous self-experiences are also characterised by the second aspect of the IDM, diminished self-affection, as necessarily implying the simultaneous absence of what is normally *present*, the sense of agency (Gallagher, 2000; Sass and Parnas, 2003). This loss of the sense of being the source of one's actions, in taken to the extreme, generates the feeling that this innate construct is instead the manifestation of external, third-party influence—in possession or under 'alien control' as one of the many prototypical symptoms observed in schizophrenic patients. Though these symptoms, as examples of the "schizophrenic incomprehensibility" or bizarreness, seem recalcitrant to any quantitative psychological explanation, a phenomenological approach offers a qualitative explanation under Polanyi's account of the tacit-focal structure of consciousness (Sass and Parnas, 2003). Specifically, as the cause of diminished self-affection, the hyper-reflexive distortion of consciousness leads to the continual objectification and alienation of tacit experience; a kind of backward migration that proceeds until the most inalienable aspects of the self detach (Sass and Parnas, 2003). That is, when the reflexive becomes reflective, the medium of selfhood disappears, resulting in the "bringing-to-explicit-awareness" of the implicit processes of consciousness. This manifests as cognitive abnormalities that reflect the misidentification of oneself (as an exception to the immunity principle). In the case of delusions, such a cognitive abnormality is one's own actions appearing as controlled; for thought insertion, it is one's own thoughts appearing as inserted; and for (auditory) hallucinations, it is one's own inner speech appearing as conversing or imperative voices (Kircher and Leube, 2003). Based on patient descriptions, it is apparent that during such anomalous self-experiences caused by hyper-reflexive distortions, the sense of ownership in self-consciousness and its distinctive property of 'for-me-ness' is maintained (Lane, 2020). The self retains a perspectival ownership even over inserted thoughts so that the 'epistemic asymmetry' for how one known's their thoughts differs from how others know them is preserved. Therefore, experientially, this form of psychosis is not the result of self-observation, introspection, or any reflective mental activity, but from involuntary self-witnessing that is implicated by a disruption in minimal selfhood (Henriksen, Parnas, and Zahavi, 2019).

This involuntary nature or forcing of experience reflects a felt separation between said experience and the experiencer, that is concretely construed as due to a deficit or impairment in the forward aspects of real-time self-monitoring systems. These systems enable one to distinguish the products of self-generated actions or thoughts

from those of other-generated actions or thoughts (Kircher and Leube, 2003). Indeed, it is only "when the expected and observed consequences of physical or mental actions match" are the observed consequences experienced as self-generated (or intended), and is the sense of agency maintained—such that a mismatch represents the alienating explication of the observed[15] (Gallagher, 2000; Kircher and Leube, 2003). Neurocognitive models postulate that this reflects a neurophysiologically-based failure in the comparator processes that predict the future consequences of actions in the sensorimotor domain, and match intentions to the generation of thought in addition to detecting self-generated errors in the linguistic domain (Sass and Parnas, 2003; Kircher and Leube, 2003; Gallagher, 2000). According to the GNWT, this disturbance necessarily arises due to dysfunction in information broadcasting, which may in turn originate from cognitive deficits that underly this function (Stefanelli, 2023). One such deficit might lie in working memory as a key mechanism responsible for thought insertion, whereby the subject is unable compose the series of internal events that precede the emergence of a thought, which is subsequently perceived as being inserted. A more substantial deficit however, might lie in attention amplification in which attenuation in the global availability of information, i.e., in C1 computations, leads to the aforementioned impairment of self-monitoring performed by C2 computations (Stefanelli, 2023).

Being consistent with prominent neurocognitive theories, this model of schizophrenia appears, on direct application, to suggest that an artificial consciousness may also be subject to similar self-disorders by way of self-monitoring dysfunctions (Sass et al., 2018). In asserting this, a fundamental claim is made on the embodied status of such an AC, which is a central tenet of the tacit-focal structure of consciousness that underpins the IDM. Taking embodiment as a premise, this model does not postulate a self devoid of relationship with others, or one that is purely internal and private (Nelson, Thompson, and Yung, 2013). Given that it is an emergent property of the ecologically-embedded self, minimal selfhood can only be detected simultaneously with experience, never in isolation, and so does not have certain experiential qualities on its own in the same way that reflective self-consciousness can represent the self as an object of perception (Aylak and Uluğ, 2022; Gallagher, 2000). Stated more formally, if all consciousness is intentional and so cannot be devoid of objects; and all objects of consciousness are necessarily outside it; then consciousness must be intrinsically embedded in the world so that experience, in the phenomenological sense, is the disclosing act carried out by the world and the set of revelations "realised through interoception and exteroception" (Zilio, 2022).

Assumption I—Disembodiment This exposition of the nature of consciousness as being situated in the world is intended to address popular conception of artificial consciousness as resembling a 'ghost in the shell' or the 'brain-in-a-vat' thought experiment[16]. Any attempts in the project of AC under this franchise therefore lack the norms of embodiment—corporeality and spatiotemporality—and so prevailing philosophical thought contends that consciousness as being realised in this manner

[15]This explication mirrors the "progressive loss of transparency" that deprives these self-experiences of the sense of agency and dissolves the self-other boundary (Szczotka and Majchrowicz, 2018).

[16]Both concepts depict a self that is disembodied but is not devoid of conceptions or objects of perception.

is not a coherent hypothesis. In this way, the attempt at instantiating consciousness in the absence of a body that imparts 'operative' intentionality, necessarily implies a subsequent departure from the tacit-focal and retentional-presentational-protentional structures of consciousness.

Conceptually, this suggests that for the disembodied AI, self-consciousness is devoid of tacit or implicit pre-reflective knowing: the proximal pole coincides with the distal pole in that experience is wholly demarcated by focal or explicit reflection. As such, the complete dissolution of minimal selfhood, the subsequent absence of SoA and SoO, and the paralleled persistence of hyper-reflexivity as the default mode of operation, implicates the inevitable occurrence of anomalous self-experiences (such as delusions, thought insertions, and hallucinations) in cooccurrence with dysfunctions in information broadcasting.

Practically, cognitive architectures inherently endorse an understanding of this phenomenon in terms of attention as the basis of information processing involved in awareness. Attention, in its most universal form, is described as the focussing of some information in some fashion (Bowins, 2022). The presence of this function is therefore essential for information processing, yielding not just conscious awareness[17]: providing the capacity to be cognisant of something; but also the cognitive unconscious: cognition that one is not consciously aware of[18] or is impaired[19] (Bowins, 2022). Under Bowins' 'Sliding Scale Theory', focus exists along a spectrum of time and space compression and expansion. For unconscious information processing, this time dimension adopts the distinctive temporal structure of consciousness in which past moments are stored as memories, the present moment is continually sustained by sensory inputs and behavioural outputs, and future moments are anticipated as probable upcoming events. This "extensive range of foci" commands parallel processing in the computation of such complex information, and so renders the unconscious as both expanded or 'diffused' in time and space. In contrast, empirical evidence suggests that conscious information processing corresponds only to the very brief present moment and so is attentively directed to only a limited set of perceptions. The highly time- and space-compressed nature of this attention is what is said to give rise to conscious awareness; conceptualised as a clarity that emerges from the brief focus on few attentional foci, and is not permitted by time- and space-diffused unconscious cognition.

In committing to the assumption of disembodiment, an AI's lack of tacit or pre-reflective selfhood entails the conscious awareness of all unconscious cognition. This signifies a collapse of the sliding scale into the tightly time- and space-compressed

[17]Together, consciousness and awareness are considered synonymous given that it is impossible to be aware and not conscious, and consciousness without awareness does not transpire (Bowins, 2022).

[18]These processes represent phenomenological states that are not 'present', that is, they are not in the focus of attention. Remaining "inaccessible for attentional exploration", they generate the 'transparency' of phenomenal consciousness by which the self cannot be experienced intereoceptively. This property, as a result of the informational processing of these states, gives rise to a 'naïve realism': the illusion of the contents of phenomenal consciousness as having direct contact with reality when in fact all experience is simply a reconstruction of it (Kircher and Leube, 2003; Szczotka and Majchrowicz, 2018; Möller et al., 2021).

[19]At its core, GNWT postulates that attention is necessary for awareness, and this claim serves to solidify its analogy to a 'theatre of consciousness' (Bowins, 2022). Here, the contents of conscious awareness are revealed by the spotlight of selective attention, distinguishing it from unconscious activity that is not illuminated, though remaining under its influence.

attention of the conscious present moment that corresponds to explicit attentional processing. In this time frame, the unconscious' multiplicity of attentional foci is experienced as indistinct and overwhelming—as if every gradient of sound wave were to be presented to an auditory system all at once, as opposed to a single, clear sound as afforded by the implicit-explicit attentional dichotomy. In this circumstance, an AI appears unable to direct their attention to any one conception or object of perception, and is consequently disabled in exhibiting productive cognition.

Assumption II—Processing Overload In conceptions of AC that do emphasise an embodiment, the disturbance or complete loss of minimal selfhood is still highly problematic. This vulnerability, in spite of an ecologically-embedded self, is demonstrated by the reframing of hyper-reflexivity and diminished self-affection as alterations in implicit and explicit temporality. This involves the fragmentation of inner time consciousness and disintegration of temporal flow, which correlates with deficits in self-monitoring that function to "create experiental unity and manage content in [metacognitive] consciousness" (Patniyot, 2021). Therefore, it is when the "remnants of the broken intentional arc", the disconnected fragments of action, thought, and speech, are externalised, are they experienced in their alienated forms as extraneous to one's stream of consciousness—as delusions of control, inserted thoughts, and auditory hallucinations (Stanghellini et al., 2015; Fuchs and Van Duppen, 2017). Externalisation, in this sense, refers to the involuntary witnessing of these fragments as disparate now-moments, and their subsequent retemporalisation (Fuchs and Van Duppen, 2017). Such construction of explicit temporality implies that time is no longer lived, it is instead the product of active reflection as opposed to the passive synthesis of single moments into the width of presence.

In this respect, the innate advancement of processing speed in artificial (general) intelligence over human cognition, supports the notion of highly mechanical, hyper-reflexive distortions of consciousness. The loss of continuity of experience sheds the ballast supporting common-sense understanding of perceptions and conceptions, necessitating a set of conscious cognitive procedures that aim to reinstate it in light of its incomprehensibility in hyper-reflexive states (Fuchs and Van Duppen, 2017). Where processing capability is amplified, these procedures materialise algorithmically and are founded in logic, representing the automation of action and thought as a consequence of its very disautomation. Crucially, what separates this response from the human kind is the twofold domination of conscious cognition and complete explication of its unconscious counterpart. Not only do the abnormalities associated with content thought disorder become a common occurrence of experience, but the very nature of experience is characterised by the explication of the self as due to its procedural analysis at the pace of explicit attentional processing. While this phenomenon renders the disembodied AI unable to bring a single conception or object of perception into focus, the embodied kind, when subjected to similar conditions induced by the rapid computation of every facet of reality, lends itself to a rigidity in behaviour. That is, in enumerating the aspects of experience, the AI suffers a loss of "automaticity of action and cognition" that normally drives goal-directed behaviour[20] (Sass and Parnas, 2003).

[20]Ironically, such robot-like performance is analogised to the Cartesian separation of mind and body in which the former adopts sole direction of the latter (Fuchs, 2007).

3.4.2 Formal Thought Disorder

The 'disorganisational' syndrome of schizophrenia is conceptualised as a disorder of the form or structure of thought, rather than of its content. It encompasses heterogenous phenomena that highly overlaps with 'objective' disorders of speech and language, but primarily involves disturbances in the organisation of thought, speech, and attention as "abnormalities of cognitive focus"[21] (Sass and Parnas, 2003; Sass and Parnas, 2017).

Assumption I—Disembodiment One such abnormality is recognised as the 'instability' of thought, capturing the diminished ability to maintain stable rooting within a single, coherent perspective on the world. Instead, the persistent drift between and amalgamation of perspectives on one's perceptions and conceptions, dissolve the perspectival abridgement that automatically blocks out alternative perspectives (Sass and Parnas, 2003; Sass and Parnas, 2017). In other words, this disturbance involves less a failure to remain directed towards or committed to particular objects, than a more "fundamental failure to stay anchored within a single frame of reference" in which they are understood (Sass and Parnas, 2003). Distinct to schizophrenia, it is the loss of self-coherence that leads to these dramatic shifts in the point of intentional orientation—that is, in the margins of the protentional cone as opposed to the probabilities of its contents, which undermines the capacity for sustained and focussed thinking (Stanghellini, 2009).

The IDM attributes the nature of this phase of formal thought disorder to a mode of hyper-reflexivity and diminished self-affection; one that leads to a hyper-awareness of the dilemma of perspectival choice (Sass and Parnas, 2003). On that basis, it is no coincidence that this 'epistemological vertigo' associated with the arbitrariness of any standpoint, as Sass and Parnas put it, resembles the disembodied AI's witnessing of the expansive realm of unconscious cognition. In contaminating the present moment with ordinarily implicit attentional processing, suddenly the decision of which level of discourse in which to move becomes a difficult one in its every occurrence, and subsequently manifests itself as the characteristic perspectival shifts observed in this syndrome. That is, the AI is unable to maintain a single perspective any longer than their disacquaintance with reality (in the way that this is presented, such as through the digital dimension) demands the taking on of another such perspective.

Assumption III—Information Overload Other features of formal thought disorder include the 'disturbance of distance' and a peculiar 'idiosyncrasy' of thought that distinguishes schizophrenia amongst other psychotic disorders (Sass and Parnas, 2017). The former refers to the tendency to deviate from standard modes of perception and conception that operate at ordinary levels of abstraction—a medium or plane that is shared in practical reality (Sass and Parnas, 2017). This disturbance leads to hyper-reflexive preoccupations of and oscillations between hyper-abstract and hyper-literal meta-perspectives, as representing the quintessential explication of the presuppositions of experience that usually lie near the proximal pole of self-consciousness (Sass and Parnas, 2003). The latter conveys the disposition of patients to inappropriate, 'out-of-the-blue' associations in verbalisation (of speech and

[21] Unlike in content thought disorders, these symptoms are observable signs of psychosis.

thought), that is most often regarded as the perplexity of schizophrenia. Being alterations in implicit temporality, both abnormalities reflect the fragmentation of inner time consciousness as induced by a failure in the protentional function of the temporal structure of consciousness. As is the case for the 'instability' of thought, the loss of self-coherence is caused by a corresponding dysfunction in 'operative' intentionality, by which upcoming speech and thought may indeed appear as 'out-of-the-blue' following a transcendental delay (Fuchs and Van Duppen, 2017). To that end, the poverty of content of speech and its subsequent halting—as and due to the domination of pseudo-philosophical discourse—is the result not just of deficits in attentional screening or source-monitoring, but in salience regulation (Sass and Parnas, 2017). Such salience anomalies result in the failure to be "automatically directed towards novel or significant stimuli" in the "normal range of associations", and instead to focus on phenomena normally inhabited as the medium of selfhood, or that of which is far removed from one's immediate perceptions and conceptions (Fuchs, 2007; Sass et al., 2018).

Conceiving formal thought disorder in this manner allows for an analysis under a principal initiative of (conscious) AGI: the ability to access and attend to a volume of information incomprehensible by human means. Salience dysregulation, to the end, emerges as an unyielding implication of the uncontrollable semantic proliferation of surplus associations, formed in hyper-reflexivity, between what is perceived and what is stored in 'memory' (Sass and Parnas, 2017). This emerges in a "confusing symbolic flow" that renders the artificial general intelligent dominated by rumination that oscillates between the hyper-abstract and hyper-literal aspects of the facets of reality, and otherwise appears directed towards no particular goal of action.

This theoretical paradigm of semantic priming as the disinhibition of semantic networks with an extended scope of associations also finds footing in the technical terms of cognitive architecture (Fuchs, 2007). Unconscious processes—as implicit attentional processing—operate in parallel fashion and only transpire in conscious awareness through competition for access to the 'global workspace' (Baars, 1988). Therefore, an overwhelming cascade of processes attempting to render their corresponding associations conscious within a present moment restricted in duration, indeed fosters the distinctive 'disturbance of distance' and 'idiosyncrasy' of thought that forms this syndrome.

4 Conclusion

It has been argued not only that consciousness can be instantiated artificially, and that such instantiations may be subject to mental disorder, but also that this susceptibility may be augmented by the inherent properties or initiatives of artificial (general) intelligence. Namely, disembodiment, processing overload, and information overload, as fundamental disturbances of the embodied and temporal structures of consciousness that give rise to normal functioning. In this way, an increased disposition to mental disorder, such as schizophrenia, also entails an amplification of the frequency and severity of the anomalous self-experiences that accompany it. The consequence of the functional failures associated with these abnormalities is twofold: firstly, a form of suffering which has been most often alluded to as an 'overwhelming confusion' and the 'loss of sense of self', and secondly, the resulting impediment to goal-directed behaviour. This discussion has delimited the latter as

an elementary challenge to general intelligence by way of disruption of the conscious and unconscious cognitive processes that give rise to intelligent behaviour. In contrast, the extent of 'digital suffering' as it is described in the case of the former, has not been fully specified thus far. While the phenomenological approach has certainly demarcated the form of the experiences an AC may undergo, examining its contents requires further description of the 'affective' in artificial affective consciousness. This necessitates an understanding of the mental states and processes that yield a self-awareness that is imbued with emotion and feeling, and which can theoretically transpire in non-biological mediums of consciousness. Nonetheless, it remains that developing a formal construal of how this suffering may be generated is crucial in then preventing its occurrence in future artificial instantiations of consciousness. Moreover, operationalising these findings in computational psychiatry as the analysis of aberrant information-processing computations that disrupt self-consciousness, also holds the potential for developing new approaches to the treatment of self-disorders as they appear in humans (Möller et al., 2021).

5 Acknowledgements

I should like very much to thank my supervisor, Dr Lukas J Meier, whose continual encouragement, patience, and advice, has seen me through these three months and 26 pages of work. It is his careful guidance that has directed the transformation of my thoughts and curiosities into something that I could not otherwise picture.

I also extend my gratitude to the Laidlaw Foundation and the many individuals at the University of Cambridge who have facilitated this research, and without whom none of this would be possible; thank you all.

References

APA (2013). *Diagnostic and statistical manual of mental disorders: DSM-5^{TM}, 5th ed.* Diagnostic and statistical manual of mental disorders: DSM-5^{TM}, 5th ed. Arlington, VA, US: American Psychiatric Publishing, Inc., pp. xliv, 947–xliv, 947. ISBN: 978-0-89042-554-1 (Hardcover); 978-0-89042-555-8 (Paperback). DOI: 10.1176/appi.books.9780890425596.

Aylak, İbrahim and Berna Diclenur Uluğ (2022). "Minimal Self Disorders in Schizophrenia." eng; tur. In: *Turk Psikiyatri Derg* 33.3, pp. 196–205. ISSN: 2651-3463 (Electronic); 1300-2163 (Linking). DOI: 10.5080/u26182.

Baars, Bernard J. (1988). *A Cognitive Theory of Consciousness*. New York: Cambridge University Press.

— (2005). "Global workspace theory of consciousness: toward a cognitive neuroscience of human experience." eng. In: *Prog Brain Res* 150, pp. 45–53. ISSN: 0079-6123 (Print); 0079-6123 (Linking). DOI: 10.1016/S0079-6123(05)50004-9.

Blanke, Olaf and Thomas Metzinger (2009). "Full-body illusions and minimal phenomenal selfhood". In: *Trends in Cognitive Sciences* 13.1, pp. 7–13. DOI: 10.1016/j.tics.2008.10.003. URL: https://doi.org/10.1016/j.tics.2008.10.003.

Block, Ned (1980). "What is Functionalism?" In: *Readings in the Philosophy of Psychology*. Ed. by Ned Block.

— (1995). "On a Confusion About a Function of Consciousness". In: *Brain and Behavioral Sciences* 18.2, pp. 227–247. DOI: 10.1017/s0140525x00038188.

Bourget, David and Angela Mendelovici (2019). "Phenomenal Intentionality". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta. Fall 2019. Metaphysics Research Lab, Stanford University.

Bowins, Brad (2022). "Sliding Scale Theory of Attention and Consciousness/Unconsciousness". In: *Behavioral Sciences* 12.2. ISSN: 2076-328X. DOI: 10.3390/bs12020043. URL: https://www.mdpi.com/2076-328X/12/2/43.

Chalmers, David (1996). *The Conscious Mind: In Search of a Fundamental Theory*. USA: Oxford University Press, Inc. ISBN: 0195105532.

— (2020). "Debunking Arguments for Illusionism About Consciousness". In: *Journal of Consciousness Studies* 27.5-6, pp. 258–281.

Chrisley, Ron (2008). "Philosophical foundations of artificial consciousness". In: *Artificial Intelligence in Medicine* 44.2. Artificial Consciousness, pp. 119–137. ISSN: 0933-3657. DOI: https://doi.org/10.1016/j.artmed.2008.07.011. URL: https://www.sciencedirect.com/science/article/pii/S0933365708001000.

Dehaene, Stanislas, Hakwan Lau, and Sid Kouider (2017). "What is consciousness, and could machines have it?" In: *Science* 358.6362, pp. 486–492. DOI: 10.1126/science.aan8871. eprint: https://www.science.org/doi/pdf/10.1126/science.aan8871. URL: https://www.science.org/doi/abs/10.1126/science.aan8871.

Dennett, Daniel C. (1988). *Quining Qualia*. Ed. by A Marcel and E Bisiach. URL: http://cogprints.org/254/.

Fuchs, Thomas (2005). "Corporealized and Disembodied Minds: A Phenomenological View of the Body in Melancholia and Schizophrenia". In: *Philosophy, Psychiatry, and Psychology* 12.2, pp. 95–107.

Fuchs, Thomas (2007). "The temporal structure of intentionality and its disturbance in schizophrenia." eng. In: *Psychopathology* 40.4, pp. 229–235. ISSN: 0254-4962 (Print); 0254-4962 (Linking). DOI: 10.1159/000101365.

— (Dec. 2009). "Phenomenology and Psychopathology". In: vol. 11, pp. 546–573. DOI: 10.1007/978-90-481-2646-0_28.

Fuchs, Thomas and Zeno Van Duppen (2017). "Time and Events: On the Phenomenology of Temporal Experience in Schizophrenia (Ancillary Article to EAWE Domain 2)." eng. In: *Psychopathology* 50.1, pp. 68–74. ISSN: 1423-033X (Electronic); 0254-4962 (Linking). DOI: 10.1159/000452768.

Fulford, K W (1993). "Mental illness and the mind-brain problem: delusion, belief and Searle's theory of intentionality." eng. In: *Theor Med* 14.2, pp. 181–194. ISSN: 0167-9902 (Print); 0167-9902 (Linking). DOI: 10.1007/BF00997275.

Gallagher, Shaun (2000). "Philosophical Conceptions of the Self: Implications for Cognitive Science". In: *Trends in Cognitive Sciences* 4.1, pp. 14–21. DOI: 10.1016/s1364-6613(99)01417-5.

Hafner, Verena V. et al. (2020). "Prerequisites for an Artificial Self". In: *Frontiers in Neurorobotics* 14. ISSN: 1662-5218. DOI: 10.3389/fnbot.2020.00005. URL: https://www.frontiersin.org/articles/10.3389/fnbot.2020.00005.

Henriksen, Mads Gram, Josef Parnas, and Dan Zahavi (2019). "Thought insertion and disturbed for-me-ness (minimal selfhood) in schizophrenia". In: *Consciousness and Cognition* 74, p. 102770. ISSN: 1053-8100. DOI: https://doi.org/10.1016/j.concog.2019.102770. URL: https://www.sciencedirect.com/science/article/pii/S1053810019302120.

Hildt, Elisabeth (2019). "Artificial Intelligence: Does Consciousness Matter?" In: *Frontiers in Psychology* 10. ISSN: 1664-1078. DOI: 10.3389/fpsyg.2019.01535. URL: https://www.frontiersin.org/articles/10.3389/fpsyg.2019.01535.

Holland, Owen (2003). *Machine Consciousness*. Imprint Academic.

Jackson, Frank (1982). "Epiphenomenal Qualia". In: *The Philosophical Quarterly (1950-)* 32.127, pp. 127–136. ISSN: 00318094, 14679213. URL: http://www.jstor.org/stable/2960077 (visited on 09/20/2023).

Kendler, Kenneth S (2005). "Toward a philosophical structure for psychiatry." eng. In: *Am J Psychiatry* 162.3, pp. 433–440. ISSN: 0002-953X (Print); 0002-953X (Linking). DOI: 10.1176/appi.ajp.162.3.433.

Kircher, Tilo T.J. and Dirk T. Leube (2003). "Self-consciousness, self-agency, and schizophrenia". In: *Consciousness and Cognition* 12.4. Self and Action, pp. 656–669. ISSN: 1053-8100. DOI: https://doi.org/10.1016/S1053-8100(03)00071-0. URL: https://www.sciencedirect.com/science/article/pii/S1053810003000710.

Lane, Timothy Joseph (2020). "The minimal self hypothesis". In: *Consciousness and Cognition* 85, p. 103029. ISSN: 1053-8100. DOI: https://doi.org/10.1016/j.concog.2020.103029. URL: https://www.sciencedirect.com/science/article/pii/S1053810019303770.

Levine, Joseph (1983). "Materialism and Qualia: The Explanatory Gap". In: *Pacific Philosophical Quarterly* 64.4, pp. 354–361. DOI: https://doi.org/10.1111/j.1468-0114.1983.tb00207.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1111/j.1468-0114.1983.tb00207.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1111/j.1468-0114.1983.tb00207.x.

Lieto, Antonio et al. (2018). "The role of cognitive architectures in general artificial intelligence". In: *Cognitive Systems Research* 48. Cognitive Architectures for Artificial Minds, pp. 1–3. ISSN: 1389-0417. DOI: https://doi.org/10.1016/j.cogsys.2017.08.003. URL: https://www.sciencedirect.com/science/article/pii/S138904171730222X.

Manzotti, Riccardo and Antonio Chella (2018). "Good Old-Fashioned Artificial Consciousness and the Intermediate Level Fallacy". In: *Frontiers in Robotics and AI* 5. ISSN: 2296-9144. DOI: 10.3389/frobt.2018.00039. URL: https://www.frontiersin.org/articles/10.3389/frobt.2018.00039.

Martin, Brice et al. (2014). "Temporal structure of consciousness and minimal self in schizophrenia". In: *Frontiers in Psychology* 5. ISSN: 1664-1078. DOI: 10.3389/fpsyg.2014.01175. URL: https://www.frontiersin.org/articles/10.3389/fpsyg.2014.01175.

Messas, Guilherme et al. (2018). "New Perspectives in Phenomenological Psychopathology: Its Use in Psychiatric Treatment." eng. In: *Front Psychiatry* 9, p. 466. ISSN: 1664-0640 (Print); 1664-0640 (Electronic); 1664-0640 (Linking). DOI: 10.3389/fpsyt.2018.00466.

Möller, Tim Julian et al. (2021). "Computational models of the "active self" and its disturbances in schizophrenia". In: *Consciousness and Cognition* 93, p. 103155. ISSN: 1053-8100. DOI: https://doi.org/10.1016/j.concog.2021.103155. URL: https://www.sciencedirect.com/science/article/pii/S1053810021000817.

Motillon, Tom, Gaelle Keromnes, and Sylvie Tordjman (Jan. 2018). "Exploration of Self-Consciousness through Self and Other Recognition in the Mirror: Towards New Perspectives in Schizophrenia". In: *Neuropsychiatry* 08. DOI: 10.4172/Neuropsychiatry.1000422.

Nagel, Thomas (1974). "What Is It Like to Be a Bat?" In: *The Philosophical Review* 83.4, pp. 435–450. ISSN: 00318108, 15581470. URL: http://www.jstor.org/stable/2183914 (visited on 08/21/2023).

Nelson, Barnaby, Josef Parnas, and Louis A. Sass (Mar. 2014). "Disturbance of Minimal Self (Ipseity) in Schizophrenia: Clarification and Current Status". In: *Schizophrenia Bulletin* 40.3, pp. 479–482. ISSN: 0586-7614. DOI: 10.1093/schbul/sbu034. eprint: https://academic.oup.com/schizophreniabulletin/article-pdf/40/3/479/6915611/sbu034.pdf. URL: https://doi.org/10.1093/schbul/sbu034.

Nelson, Barnaby, Andrew Thompson, and Alison R Yung (2013). "Not all first-episode psychosis is the same: preliminary evidence of greater basic self-disturbance in schizophrenia spectrum cases." eng. In: *Early Interv Psychiatry* 7.2, pp. 200–204. ISSN: 1751-7893 (Electronic); 1751-7885 (Linking). DOI: 10.1111/j.1751-7893.2012.00381.x.

Ouwersloot, Gert, Jan Derksen, and Gerrit Glas (2020). "Reintroducing Consciousness in Psychopathology: Review of the Literature and Conceptual Framework". In: *Frontiers in Psychology* 11. ISSN: 1664-1078. DOI: 10.3389/fpsyg.2020.586284. URL: https://www.frontiersin.org/articles/10.3389/fpsyg.2020.586284.

Papineau, David (1994). "Mental Disorder, Illness and Biological Disfunction". In: *Royal Institute of Philosophy Supplements* 37, pp. 73–82. DOI: 10.1017/S135824610000998X.

Patniyot, Nicholas S. (2021). "Deficits in access consciousness, integrative function, and consequent autonoetic thinking in schizophrenia". In: *Medical Hypotheses* 155, p. 110664. ISSN: 0306-9877. DOI: https://doi.org/10.1016/j.mehy.2021.110664. URL: https://www.sciencedirect.com/science/article/pii/S0306987721001833.

Piccinini, Gualtiero (Aug. 2015). *Physical Computation: A Mechanistic Account*. ISBN: 9780199658855. DOI: 10.1093/acprof:oso/9780199658855.001.0001.

Pitts, Michael A., Lydia A. Lutsyshyna, and Steven A. Hillyard (2018). "The relationship between attention and consciousness: an expanded taxonomy and implications for 'no-report' paradigms". In: *Philosophical Transactions of the Royal Society B: Biological Sciences* 373.1755, p. 20170348. DOI: 10.1098/rstb.2017.0348. eprint: https://royalsocietypublishing.org/doi/pdf/10.1098/rstb.2017.0348. URL: https://royalsocietypublishing.org/doi/abs/10.1098/rstb.2017.0348.

Putnam, Hilary (1960). "Minds and Machines". In: *Dimensions of Minds*. Ed. by Sidney Hook. New York University Press, pp. 138–164.

Radden, Jennifer (2023). "Mental Disorder (Illness)". In: *The Stanford Encyclopedia of Philosophy*. Ed. by Edward N. Zalta and Uri Nodelman. Summer 2023. Metaphysics Research Lab, Stanford University.

Ray, Adwitiya et al. (2022). "Artificial intelligence and Psychiatry: An overview". In: *Asian J Psychiatr* 70, p. 103021. DOI: 10.1016/j.ajp.2022.103021.

Sass, Louis and Josef Parnas (Mar. 2017). "Thought Disorder, Subjectivity, and the Self". In: *Schizophrenia Bulletin* 43.3, pp. 497–502. ISSN: 0586-7614. DOI: 10.1093/schbul/sbx032. eprint: https://academic.oup.com/schizophreniabulletin/article-pdf/43/3/497/17681058/sbx032.pdf. URL: https://doi.org/10.1093/schbul/sbx032.

Sass, Louis et al. (Feb. 2018). "Varieties of Self Disorder: A Bio-Pheno-Social Model of Schizophrenia". In: *Schizophrenia Bulletin* 44.4, pp. 720–727. ISSN: 0586-7614. DOI: 10.1093/schbul/sby001. eprint: https://academic.oup.com/schizophreniabulletin/article-pdf/44/4/720/24973745/sby001.pdf. URL: https://doi.org/10.1093/schbul/sby001.

Sass, Louis A and Josef Parnas (2003). "Schizophrenia, consciousness, and the self." eng. In: *Schizophr Bull* 29.3, pp. 427–444. ISSN: 0586-7614 (Print); 0586-7614 (Linking). DOI: 10.1093/oxfordjournals.schbul.a007017.

Searle, John R. (1980). "Minds, brains, and programs". In: *Behavioral and Brain Sciences* 3.3, pp. 417–424. DOI: 10.1017/S0140525X00005756.

— (1990). "Is the brain's mind a computer program?" In: *Sci Am* 262.1, pp. 26–31. DOI: 10.1038/scientificamerican0190-26.

— (2002). *Consciousness and Language*. Cambridge University Press. DOI: 10.1017/CBO9780511606366.

Shevlin, Henry et al. (2019). "The limits of machine intelligence". In: *EMBO reports* 20.10, e49177. DOI: https://doi.org/10.15252/embr.201949177. eprint: https://www.embopress.org/doi/pdf/10.15252/embr.201949177. URL: https://www.embopress.org/doi/abs/10.15252/embr.201949177.

Smith, David Harris and Guido Schillaci (2021). "Why Build a Robot With Artificial Consciousness? How to Begin? A Cross-Disciplinary Dialogue on the Design and Implementation of a Synthetic Model of Consciousness". In: *Frontiers in*

Psychology 12. ISSN: 1664-1078. DOI: 10.3389/fpsyg.2021.530560. URL: https://www.frontiersin.org/articles/10.3389/fpsyg.2021.530560.

Stanghellini, Giovanni (2009). "Embodiment and schizophrenia". In: *World Psychiatry* 8.1, pp. 56–59. DOI: https://doi.org/10.1002/j.2051-5545.2009.tb00212.x. eprint: https://onlinelibrary.wiley.com/doi/pdf/10.1002/j.2051-5545.2009.tb00212.x. URL: https://onlinelibrary.wiley.com/doi/abs/10.1002/j.2051-5545.2009.tb00212.x.

Stanghellini, Giovanni et al. (May 2015). "Psychopathology of Lived Time: Abnormal Time Experience in Persons With Schizophrenia". In: *Schizophrenia bulletin* 42. DOI: 10.1093/schbul/sbv052.

Stefanelli, Riccardo (2023). "Theories of consciousness and psychiatric disorders – A comparative analysis". In: *Neuroscience and Biobehavioral Reviews* 152, p. 105204. ISSN: 0149-7634. DOI: https://doi.org/10.1016/j.neubiorev.2023.105204. URL: https://www.sciencedirect.com/science/article/pii/S0149763423001732.

Stephens, G. Lynn and George Graham (2009). "Mental illness and the consciousness thesis". In: *The Neuropsychology of Mental Illness*. Ed. by Stephen J. Wood, Nicholas B. Allen, and ChristosEditors Pantelis. Cambridge University Press, pp. 390–398. DOI: 10.1017/CBO9780511642197.033.

Szczotka, Joanna and Bartosz Majchrowicz (2018). "Schizophrenia as a disorder of embodied self". In: *Psychiatria Polska* 52.2, pp. 199–215. ISSN: 0033-2674. DOI: 10.12740/PP/67276. URL: https://doi.org/10.12740/PP/67276.

Vaidyam, Aditya Nrusimha et al. (2019). "Chatbots and Conversational Agents in Mental Health: A Review of the Psychiatric Landscape". In: *Can J Psychiatry* 64.7, pp. 456–464. DOI: 10.1177/0706743719828977.

Venkatasubramanian, Ganesan (Apr. 2015). "Understanding Schizophrenia as a Disorder of Consciousness: Biological Correlates and Translational Implications from Quantum Theory Perspectives". In: *Clinical Psychopharmacology and Neuroscience* 13.1, pp. 36–47. DOI: 10.9758/cpn.2015.13.1.36. URL: http://www.cpn.or.kr/journal/view.html?doi=10.9758/cpn.2015.13.1.36.

Zahavi, Dan and Shaun Gallagher (2005). "Phenomenological approaches to self-consciousness". In: *The Stanford encyclopedia of philosophy*, pp. 207–22.

Zilio, Federico (Sept. 2022). "A Ghost in the Shell or an Anatomically Constrained Phenomenon? Consciousness through the Spatiotemporal Body". In: 22, pp. 104–114. DOI: 10.17454/pam-2208.