Consciousness, Mathematics and Reality: A Unified Phenomenology

Igor Ševo

April 19, 2023

Abstract

Every scientific theory is a simulacrum of reality, every written story a simulacrum of the canon, and every conceptualization of a subjective perspective a simulacrum of the consciousness behind it—but is there a shared essence to these simulacra? The pursuit of answering seemingly disparate fundamental questions across different disciplines may ultimately converge into a single solution: a single ontological answer underlying grand unified theory, hard problem of consciousness, and the foundation of mathematics. I provide a hypothesis, a speculative approximation, supported by a comprehensive overview of scientific evidence and philosophical literature, of a unified epistemic and phenomenological model and, in doing so, propose a parsimonious solution to the hard problem of consciousness.

The proposition of the hypothesis bears important implications for cross-disciplinary study between linguistics, mathematics, physics, and computer science, and offers new epistemic, ontological, and ethical propositions about the nature of AI consciousness, as well as existential risk mitigation.

1 Introduction

To convey the importance of studying the basic building blocks of the universe, Richard Feynman famously said (Feynman, 1963, pp. 1–8): *Everything is made from atoms. That is the key hypothesis. The most important hypothesis in all of biology, for example, is that everything that animals do, atoms do. In other words, there is nothing that living things do that cannot be understood from the point of view that they are made of atoms acting according to the laws of physics.* Of course, he was using the term *atoms* in a loose sense to denote whatever the basic constituents of the universe are from the point of view of a particular theory. For a particle physicist, *atoms* could mean particles, for a quantum field theorist, quantum fields, for a quantum information scientist, quanta of information, but, in essence, they are all attempting to accurately model and predict reality, without necessarily answering the ontological question of what an atom *itself* is.

From an idealist perspective (Berkeley, 1710) (Kant, 1781), a human being can only experience reality from its own subjective perspective, i.e., phenomenally. Everything we observe, learn, perceive, and think about is only available to us within the horizons of what we call conscious experience, as subjective manifestations we call *qualia*. In that sense, all information we consciously understand and process manifests to us exclusively phenomenally, through various

forms of qualia. Yet, we have scientific models of reality with stupendous predictive power, and we understand that this conscious experience is deeply correlated with the structure and activity of our brains. The question of why we have these phenomenal experiences is dubbed *the hard problem of consciousness* (Chalmers, 1996).

On the other hand, language itself presents us with an interesting limitation: vagueness (Williamson, 1994) (Keefe & Smith, 1997) of our predicates dilutes the meaning of what is being explored. In other words, what do we *mean* by "consciousness"? The Oxford dictionary defines consciousness as *the state of being aware of and responsive to one's surroundings* and *the fact of awareness by the mind of itself and the world*. On the other hand, "sentient" is defined as *able to perceive or feel things*. The distinction between the two concepts is particularly relevant for this discussion as a preponderance of literature conflates the kind of consciousness which implies a form of self-awareness, identity, or cognitive processing with the concept of sentience which (Block, 1995) defines as "phenomenal consciousness", which captures the "what it is like" nature of experience (Nagel T. , 1974), rather than identity or self-recognition.

A typical scientific description is that either neural correlates (Crick & Koch, 1990) (Varela, 1996), neuronal processes (Dennett, 1991), brain processes (Searle, 1992), neuronal group activity (Edelman, 1989), information sharing across a neuronal workspace (Dehaene, 2014), brain's interaction with the environment (Varela, Thompson, & Rosch, 1991) (Merleau-Ponty, 1945) (Damasio, 1999) (Noë, 2004) (O'Regan & Noë, 2001), or even quantum processes in neural microtubules (Penrose & Hameroff, 2011) give rise to the conscious experience. Notably, reductionist approaches (Churchland P. M., 1985) (Churchland P. S., 1986), which argue that consciousness can be reduced to fundamental components, are prevalent across science. The overall implicit assumption, as observed by (Hoffman, 2008), is that consciousness "arises" from an underlying physical reality, and, depending on the interpretation, is either a manifesting phenomenon or a secondary substance of the universe correlated with the physical world without causal agency.

The tendency towards physicalist (Papineau, 2002) monism and materialism, which suppose that reality is fundamentally physical, is reasonable from a scientific point of view, as they seem the most parsimonious, but recent advances indicate that many of the concepts we attribute to the physical world are merely constructs developed over the course of evolutionary history to maximize our chances of survival (Hoffman, Singh, & Prakash, 2015). Our internal representations of both time (Rovelli, 2018) (Price, 1996) and space (Markopoulou, 2009) (Jafferis, et al., 2022) (Susskind, 1995) may not be accurate reflections of the noumenal reality. Recently, theories claiming that consciousness is a fundamental aspect of the universe have been gaining traction, most notably integrated information theory (Tononi & Edelman, 1998) (Tononi, 2004) (Koch, 2012), which posits that all things possess a degree of consciousness and that consciousness is quantifiable and measurable through mathematical means, and conscious realism (Hoffman, 2008), which suggests that the universe consists of interacting conscious agents subject to scientific analysis (an idea that dates back to Leibnitz (Leibniz, 1714)). Additionally, Chalmers (Chalmers, 1996) makes a panpsychist case that consciousness is a fundamental aspect of reality and (Lamme, 2006) (Guertin, 2019) for different levels of consciousness across brain

structures. Nonetheless, research indicating a strong correlation between the brain and the mind is undeniable, the famous Libet experiments which question the existence of free will (Libet, Gleason, Wright, & Pearl, 1983), semantic priming studies (Dehaene, et al., 1998), unconscious activation of cognitive control (Lau & Passingham, 2007), neural dynamics observation (Cul, Baillet, & Dehaene, 2007), and split consciousness research (Pinto, et al., 2017) (Haan, et al., 2020) being only a few important examples.

Of course, the existence of conscious experience cannot be dismissed merely through logical means (Levine, 1983), and first-person perspectives cannot be easily accounted for by reductionist approaches (Nagel T. , 1974) (Siewert, 1998), but, aside from neutral monism (Russell, 1927) proposes a third underlying structure and pluralism which suggests an interplay of conscious agents (Oizumi, Albantakis, & Tononi, 2014) or multiple centers of conscious experience (Bayne & Chalmers, 2003) (Bayne, Hohwy, & Owen, 2016), most prominent solutions imply a form of emergence of mental from physical.

However, an argument can be made that all forms of physicalism (Kim, 2005) entail some form of panpsychism (Strawson, 2006) by which everything that exists must be phenomenal. More recently, new forms of idealism have been emerging, including quantum idealism (Stapp, 1993) (Stapp, 2009), which posits a quantum mechanical basis for the first-person perspective, conscious realism (Hoffman, 2008), objective idealism (Goff, 2019), which posits that the universe possesses consciousness, and that an instance of individual consciousness is a subset of that universal field of consciousness.

Solutions exist that attempt to equate the physical process said to underly consciousness with that consciousness. These approaches can broadly be categorized as identity theories, namely mind-brain identity (Place, 1956) (Smart, 1959) equating mind states and brain states, token-identity (Davidson, 1970) arguing identity between mind tokens and brain tokens and, broadly speaking, functionalism (Putnam, 1967), which posits a relation between mental states and sensory inputs and outputs.

In fact, this condensed overview of the consciousness literature only provides a glimpse of the ongoing philosophical and scientific debate, which, broadly speaking, aggregates to a tendency in science to disregard the causal relevance of consciousness in favor of materialism and discounting it as an emergent phenomenon, juxtaposed with a broad critique based on the undeniability of the "what is it like" nature of phenomenal experience, i.e., qualia of subjective experience, which cannot be dismissed, regardless of its causal relevance.

At a first glance, these theories seem to stand in an irreconcilable opposition. However, instead of claiming that any of them are fundamentally wrong or antithetical to others, we can recognize that they are different attempts at approximating the truth, steps in a broader deduction process converging towards a synthesis.

2 What is it like to be an atom?

As Nagel argues (Nagel T. , 1974), it is impossible for a conscious being to truly understand another being's conscious perspective. However, an implicit presupposition is made here: some agents, such as bats, are entitled to subjective perspectives, while others, such as electrons, are not. Token and state physicalists make a similar implicit premise in that mind states only correspond to physical states in the brain, or, more subtly, that the brain states are the only kind of states conducive to mind–state identification.

Fundamentally, every non-idealist notion of reality is a leap of faith, as all information that enters the horizons of our conscious experience is entirely phenomenal. Leaving aside the unreliability of our concept of the past (Dyson, Kleban, & Susskind, 2002) (Putnam, 1981) (Nozick, 1983) to predict and act effectively in the world, we must build models and representations of it. Even though the descriptions of these models are embedded across structures described in language as brains, books, or storage media, when recalled to consciousness they are always experienced phenomenally, as distinct qualia expressing them. Although the existence and predictive potency of these models is undeniable, only a sentient agent can testify to their existence. In other words, a testimony to existence can only be phenomenal.

One could argue that such information could be embedded in a non-conscious substrate before being propagated to a conscious agent making the final observation. Some theories of consciousness based on quantum mechanics make similar claims (Stapp, 1993) (Hameroff & Penrose, 2014). However, a more parsimonious interpretation would entail some form of panpsychism (Rosenblum & Kuttner, 2006) (Strawson, 2006) (Hoffman, 2008).

In simpler terms, why would a bat's brain states be entitled to being identified with qualia and a pebble's pebble-states not? Why do we assume a bat has a phenomenal perspective and an electron does not? Other than the bat having a brain, the choice is relatively arbitrary. In fact, why do we assume other humans have such a perspective (Kripke, 1980) (Chalmers, 1996)? Clearly, we recognize other humans' similarity to us and assume they must experience consciousness in a similar way. By such reasoning, a chimpanzee is more likely to be conscious than a bat, and a bat more likely to be conscious than a molecule of dopamine.

The vagueness of linguistic predicates further complexifies the problem: the casual use of the term "bat" of course implies only that part of the bat's brain with the corresponding conscious brain states, and the meaning of the term "consciousness" implicitly changes to refer more to phenomenal consciousness—sentience. Whether we imply sentience or not, we might ask the question what might it feel like to be an atom? If a specific atom, in the looser sense of the word denoting whatever fundamental constituent, is a thing in the physical world, what is that thing in and of itself? What would a universe with a single atom be like? To say anything about such an atom would entail conceptualizing it within our consciousness, phenomenally. In the concrete case, this would imply some interaction between the atom and the state token we identify our consciousness with. In other words, it would be "measured" or perceived by our consciousness.

Note that no specific physical model is assumed: an atom could refer to an elementary particle or a quantum state or any other mathematical description of the underlying reality. By our conscious

accounts, we can claim that something exists and that that something is phenomenal in nature. Everything that is perceived is only perceived phenomenally and everything that is experienced is only experienced phenomenally. In other words, the building blocks of sentient experience—qualia—are different instances of *somethingness*. To claim that an inanimate substrate exists outside perception would be to claim that it is *something* even when not perceived or interacted with. Thus, to be an atom is to be *something* rather than nothing. The claim that the mode of somethingness of an atom is different from the mode of somethingness of a particular human's consciousness is not a dualist claim, but rather a claim about the profound difference of the atom's mode of phenomenal existence to the mode of existence of that consciousness. In this sense, whatever an atom's behavioral logic, it is an elementary instance of sentience. Note that this proposition is invariant to the mechanism of atoms' interactions, their determinism (Hooft, 1985) (Sutherland, 2017), or the physical laws modeling them: it is an ontological claim.

In fact, the claim does not imply identity, agency, self-recognition, ego, or awareness, but an elementary form of sentience. The hypothesis simply states that to say that something *is* is to say that it is qualia. Namely, to say "atoms *are* sentient" is not to claim sentience arising from atoms, but to identify atoms, in the broader sense of the word, with a sentient experience. In that sense, any concrete term for an atom (e.g., particle, wave, quantum state, (q)bit of information etc.) attempts to describe the corresponding elementary sentient experience and its governing rules.

3 Language and third-person representation

We could accept the proposition that the word "atom" describes an elementary instance of sentience and still ask the question "what is an atom?". Knowing that the universe is comprised solely of sentient agents on its own tells us nothing about their properties and behavior.

Every existing scientific theory is an incomplete perspective on a subset of all existing phenomena—they are attempting to predict and approximate reality and although they share mechanisms, some are better at making certain predictions than others. For example, one of the most important open problems in physics is unifying the physics of the large with the physics of the small. Both general relativity and quantum mechanics possess tremendous predictive powers, but they are based on mathematically incompatible assumptions. As a simplification, the former assumes continuous non-linear space, while the latter assumes either quantized or discrete space. However, recent theoretical (Raamsdonk, 2010) (Maldacena & Susskind, 2013) (Xu, Susskind, Su, & Swingle, 2020) and experimental (Jafferis, et al., 2022) advances indicate that the very concepts of space and time are simply representations that worked well for humans in their environment but are not intrinsic properties of the universe. From a theory of mind point of view, this would, in some ways, align with the more philosophical arguments by (Hoffman, 2008). More importantly, from a contemporary physicist's point of view, definitions of the fundamental constituents are shifting from relying on spatially based concepts, such as locality, to being based on concepts more prevalent in information theory. In this sense, quantum entanglement—the mechanism of informational coupling between systems—becomes more fundamental than what was previously considered to be. We are moving towards an interpretation of the world in which its fundamental constituents do not at all reside in space. They are no longer seen as particles

interacting, but as entangled quanta of information undergoing coupling transformations. Of course, this may not turn out to be the correct or final model, but it is nonetheless closer to reality. *Atoms*, for a contemporary physicist, have become information. For a philosopher of mind subscribing to the hypothesis proposed here, the word "information" now more closely describes elementary sentient experience. Science provides us with a way to quantify, measure, model and predict the behavior of phenomenal conscious agents and has been doing so from the very beginning of scientific thought.

Although conscious realism (Hoffman, Singh, & Prakash, 2015) and integrated information theory (Tononi, 2004) both provide a formalism for quantifying and describing consciousness, these formalisms may be redundant given the existence of reliable mathematical models used for physical theories. A typical philosophical investigation of consciousness entails defining a specific kind of consciousness, such as, for example, the ego, and then proceeding to evaluate it mathematically or logically: a vague definition of consciousness is used as the basis for further analysis, resulting in an elaborate treatment of an essentially arbitrarily chosen subset of human conscious experience. Indeed, we could search for the locus or neural correlates of the ego aspect of consciousness, self-recognition, free choice perception, or any other psychological aspect and, presumably, with sufficiently rigorous experimentation arrive at the exact match. However, the vagueness of the initial presupposition will be embedded in the finding: we will have found the locus of an arbitrarily chosen subset of the general conscious experience. On the other hand, a stricter mathematical definition of consciousness attempting to formalize one of the aspects for further mathematical analysis would arrive at a similar result: a set of mathematical conclusions deduced from a formal premise that was itself, at least associatively, based on an initially vague linguistic concept. This is not to say that either endeavor is without merit or that it would not reveal some fundamental mathematical, metaphysical, or psychological truth, but rather that in either case it will, at best, be incomplete and, at worst, redundant.

Interestingly, from a psychological point of view, the world could be seen as consisting of interacting phenomenal agents. For example, Jung (Jung C. G., 1951) writes:

> *Experience shows that [ego] rests on two seemingly different bases: the somatic and the psychic. The somatic basis is inferred from the totality of endosomatic perceptions, which for their part are already of a psychic nature and are associated with the ego, and are therefore conscious. They are produced by endosomatic stimuli, only some of which cross the threshold of consciousness. […] But there is no doubt that a large proportion of these endosomatic stimuli are simply incapable of consciousness and are so elementary that there is no reason to assign them a psychic nature—unless of course one favors the philosophical view that all life-processes are psychic anyway.*

Although the possibility of all processes being phenomenal is acknowledged, Jung *defines* a distinction between the phenomenal, i.e., processes of *psychic* nature, and physical, based on the goals of psychoanalysis. Of course, a single example does not account for the entire psychoanalytic field, but the fact remains that psychology attempts to approximate the same reality as the natural sciences, albeit from a different perspective and at a different level of analysis, which is arguably considerably less formally rigorous and predictive.

The early works of Plato on the theory of forms present a distinction between the mathematical truths and the physical world and this distinction between two types of information—the transcendental information, which includes ideas like the Pythagorean theorem or Fermat's last theorem, and the physical information, the one described by, for example, particle physics or quantum field theory as being the compositional substrate of the universe—has persisted into the modern age. However, mathematicians are beginning to argue against this distinction. The mathematical universe hypothesis (Tegmark, 2014) posits that the universe is entirely a mathematical structure. Others (Chaitin, 2006) (Putnam, 1975) (Penrose, 1989) (Maddy, 1997) are also suggesting that mathematical truths must be embedded in the fabric of reality, rather than existing on a separate plane. Additionally, general relativity, quantum mechanics, and causality can all be shown to emerge from a relatively simple computationally mathematical model (Wolfram, 2020). It seems that mathematical thought is converging towards a unified theory by which the universe is a mathematical structure acting on itself.

For a linguist, the problem of consciousness is addressed indirectly, through examination of how meaning arises from combining words (Heim & Kratzer, 1998) (Szabó, 2017) (Partee, 2004) (Stalnaker, 1999). In fact, mental representation and concept formation is often proposed as a fundamental aspect of conscious experience (Kaplan, 1989) (Fodor, 1975) as well as thinking (Evans & Green, 2006), and innate linguistic capacity has been proposed as an important mechanism for the acquisition of language (Chomsky, 1980). Clearly, one of the central open problems in linguistics is how apparently arbitrary words manage to refer to things in the world (Whorf, 1956) (Kripke, 1980) and how certain information is determined to be relevant in a specific context (Putnam, 1975) (Rosch, 1978). The main philosophical issue for a linguist is clearly *meaning* (Quine, 1951) (Davidson, 1967). The link between language and mathematics is undeniable (Hofstadter, 1979) (Chomsky, 1956), and notable computational descriptions of reality rely on generative grammars (Wolfram, 2020) (Chomsky, 1959), but the nature of the relation may not be obtainable without addressing the innate vagueness of language (Williamson, 1994) (Keefe & Smith, 1997).

Most reductionist approaches to consciousness are critiqued on the basis of not accounting for strong emergence (Chalmers, 2008), which supposes that some macroscopically manifested properties exert irreducible causal influence on the system's behavior which cannot be attributed to the system's constituent components. Expectedly, there are many counterarguments to strong emergence including linguistic vagueness, lack of empirical evidence (O'Connor & Wong, 2012) causal reducibility to constituent components (Bedau, 1997) (Crane, 2001) and epiphenomenalism (Kim, 2000), which claims that even if strong emergent properties exist, they are not causally affecting the underlying systems. Other than a few sporadic alternative cases (Laughlin, Pines, Schmalian, Stojkovic, & Wolynes, 2000) (Anderson, 1972), consciousness seems to be the prominent example of strong emergence (O'Connor & Wong, 2012). Although language itself is sometimes considered an emergent property of the underlying simple cognitive mechanisms (Elman, 1995) (Steels, 1997), a reasonable explanation for emergence is that it is a linguistic artefact of our innate tendency towards categorization.

Conceptual vagueness has been used as a counterargument for other philosophical problems, most notably to explain away the sorites paradox (Keefe & Smith, 1997) (Williamson, 1994), as well as emergence (Silberstein & McGeever, 2003) (Humphreys, 1997). If the sorites paradox is not a metaphysical problem, but an artefact of linguistic categorization vagueness, a reductionist argument can be made for panpsychism (Ševo, 2021).

However, the dualist discussion neglects a third possibility by which both the constituents and the combined system are both conscious. If we uphold the proposition that if something is, it is, by the nature of its being, phenomenal, and if strong emergence is assumed to be possible, then an emergent system is an atomic consciousness as much as its constituents. If an emergent system can be said to exist independently of its constituents, then it, by the statement of its existence, must *be* in and of itself, and thus possess a kind of its own "what it is like" perspective, as argued above, not necessarily with identity or agency, but with a sentient quality. A complex system such as the internet might be considered conscious (Turchin, 1995) (Lloyd, 2006), but its mode of consciousness does not preclude the consciousness of individual humans from whose digital interactions it is said to arise (Dennett, 1991). By the same reductionist argument, one could infer that the human mode of consciousness does not negate the mode of consciousness of its underlying constituents. In this way, reductionist and non-reductionist approaches are reconciled by virtue of each explaining a different form of consciousness. The idea of nested consciousness, such as the China brain thought-experiment (Block, 1978), has become more relevant with insights from quantum mechanics (Georgiev, 2017), as they challenge the traditional concepts of causality and interaction which are often used in functionalist counterarguments to the existence of qualia, which, themselves, are contested philosophically (Dennett, 1991).

Evidently, there is insufficient evidence for strong emergence and the question of what it means for a strongly emergent consciousness to exist independently of its producing substrate still merits examination. For example, if an epiphenomenal consciousness exists as an independent phenomenal entity, it cannot, by virtue of its informational isolation, exert causal influence on the outside world (Kim, 2000). The famous Mary's room thought experiment (Jackson, 1982) positing that a colorblind person can know all facts about a specific color and still gain new knowledge when experiencing it for the first time has been disputed in different ways (Churchland P. M., 1985) (Carruthers & Veillet, 2011), most notably by the claim that the limitations of language play a role in how we conceptualize qualia. If one could know everything there is about color, that knowledge would manifest, among other ways, as the human phenomenal experience of color. As argued before, all knowledge and understanding must be phenomenal: both a contemplation of a color's frequency and the "colorness" of it are, each in its specific way, a phenomenal experience. One could, if the "colorness" of red could be formally defined, experimentally locate the neural correlate of that property and, from a token-physicalist point of view, claim that the determined correlate *is* the experience of "colorness". A similar claim could be made about the experience of knowing the color's frequency: a corresponding neural structure could equally be located. In other words, the structures in the brain of a patient being probed for color frequency conception correlates embed the information about the color's frequency, and the brain structures of the examiner embed the information about the patient's brain and the frequency conception correlates within it. To state that the atoms of the examiner's brain encode information about

atoms of the patient's brain is to, at the same time, make the claim that the examiner experiences knowledge of some part of the patient's mind. If the examiner's brain could encode the entirety of the patient's brain, the patient's brain structures would be a subset of the examiner's and thus exist as a copy within the examiner's brain, independent of it. In fact, if the copy were not independent from the rest of the examiner's brain structure, it would not constitute an entirely faithful copy. Thus, to fully understand the patient's mind, the examiner would have to become identical to the patient which is, interestingly, a physicalist argument that entirely aligns with Nagel's (Nagel T. , 1974) anti-reductionist argument. The issue seems again to stem from the incompleteness of both physicalism and reductionism: our presuppositions about the fundamental physical substrate are incomplete.

As noted earlier, quantum information experiments (Jafferis, et al., 2022) and quantum information theory, suggest that quantum entanglement and quantum information are more fundamental than space or particles and that these concepts arise as manifestations of a deeper reality which, for now, seems to be informational. In that sense, the argument that a mind cannot be fully observed phenomenally by another aligns with the statement of the no-cloning theorem (Wootters & Zurek, 1982). In fact, more recent theoretical advances indicate that quantum entanglement (Bužek & Hillery, 1996) (Horodecki, Horodecki, Horodecki, & Horodecki, 2009) (Nielsen & Chuang, 2010) is crucial for resolving the measurement problem, i.e., the problem of how the apparent probabilistic nature of the universe can be reconciled by the fact that once a measurement is performed a concrete measured value is observed, in that measurement, interaction, and entanglement are intrinsically linked, or possibly identical, phenomena (Susskind, 2016) (Schlosshauer, 2005) and potentially underly the concepts of space and geometry.

The fact that information and entanglement seem to be more fundamental than, for example, particles and space, is an indication of the illusory nature of locality and, possibly, other mathematical relations as well. The existence of different parallel modes of entanglement (Barreiro, Wei, & Kwiat, 2008), outside the notion of space, opens the possibility that our understanding of how a reductionist argument may be performed is limited by our representations. For example, the reductionist premise that a brain consists of spatially distributed neurons which themselves consist of spatially distributed elementary particles is invalidated by space being an inaccurate representation of the noumenal reality. Spatial intuition steers the typical reductionist argument to imagine a neural token as consisting of spatially distributed particles, rather than entangled information that resides in a non-spatial universe. In that sense, making a spatially premised reductionist argument is equally inaccurate to making a spatially premised non-reductionist one. They are both attempting to approximate the truth through the use of language, but they are both premised on incomplete and vague definitions.

Interestingly, in the field of artificial intelligence research, recent theoretical developments suggest the possibility of language models, such as generative pre-trained transformers (Brown, et al., 2020), exhibiting behavior that can be described by a quantum mechanical formalism—they can be viewed as multiverse generators (Reynolds & McDonell, 2021) in that they produce a branching structure of possible continuations for any given natural language input. An argument

is made, by the same authors, that the human imaginative process, as well as reading and writing, is a multiverse creation process. In other words, natural language is a medium for generating and exploring multiverses. Notably, experimental results making use of artificial intelligence to predict the behavior of celestial bodies suggest that these artificial systems are able to deduce natural laws underlying spatial representation (Qin, 2020), providing evidence in favor of the simulation hypothesis (Bostrom, 2003). More importantly, more arguments are being made for language as the fundamental substrate—fundamentally phenomenal substrate if the proposed hypothesis is admitted.

In fact, if we extend the notion of a word to encompass not only combinations of letters, but any piece of information, be it visual, auditory, or any other kind, that may be combined with other elementary pieces of information, we arrive at either physics or mathematics: there emerges a formal theory, which could be described as a grammar, with objects whose abstract behavior it describes. In that sense, a universal grammar (Chomsky, 1957) (Chomsky, 1965) (Chomsky, 1995) (Evans & Levinson, 2009) need not only apply to human language, but to the laws of nature, which, fundamentally, describe reality itself.

Albeit with a greater verbosity and a loss of conciseness, all mathematical statements can be translated to natural language without loss of meaning. Given that all axiomatic claims are fundamentally based on linguistic presuppositions, a mathematical universe hypothesis can be restated equivalently as a linguistic universe hypothesis. If physical and transcendental information can be equated or reside within the same universal substrate, as already stated, the claims about a fundamentally mathematical, information-based, or a linguistic universe would reduce to the same proposition. The universal grammar that underlies the fundamental language, the physical laws that underly information, or the mathematical relations that underly mathematical objects are phenomenal representations of the same universal laws: they are theoretical approximations of the same set of fundamental rules, albeit performed from different fields with disparate levels of accuracy.

Fundamentally, to exist independently of everything else, is to *be* without the presence of another something—to *be* in and of self. If something assumed to exist entirely independently of all observers is not there to be its own observer and through existing implicitly justify the proposition that it *is*, then it could be said that it simply *is not*. Therefore, to be, an atom must be a phenomenal entity of some form in and of oneself or be a part of some phenomenal entity that *is* in and of itself. To acknowledge that an atom *is* is to acknowledge its phenomenal, qualitative, status. In that sense, qualia are informational testimonies of being, not elements of the "subjective" experience entitled solely to human consciousness.

Conceivably, the perception of a bat's consciousness being disjoint from a human observer's consciousness may merely be a result of the bandwidth limitation of the natural communication channel enabling the recognition of the bat's consciousness on the human's part. In other words, a human consciousness has fewer means to directly probe the bat's brain than to probe its own associated brain—the laws of physics dictate this restriction. A consciousness entirely separated from our universe would entail us knowing absolutely nothing about it—it would be completely

unknown and causally disconnected. If we accept, for example, that quantum information entanglement decides the level and means of coupling between systems, these systems, as phenomenal entities, would phenomenally perceive one another to the degree they are coupled to one another. The weaker the coupling, the stronger the perception of the "boundary" of each conscious experience. Split-brain research (Haan, et al., 2020) indicates a similar result: consciousness is not split by the neurological procedure, but its nature altered. If the entire universe is connected, as superdeterminism would suggest (Hooft, 2014), then a panpsychist view suggested by the hypothesis laid out here seems significantly more plausible.

Nonetheless, to say that the universe consists of interacting conscious agents (e.g., as in conscious realism) is inaccurate and inconsistent with the results of quantum physics. If two agents can be considered entirely independent, i.e., unentangled, they must exist in separate universes which can, given the identity of mathematical and physical information, never conceive of any aspect of one another. Conceivably, if we forgo causality, we can resolve the issue by stating that both agents are contained within a simulator executing their behavior and claim that they are both epiphenomenal. Still, the term "interaction" would not apply—the agents are either always connected, i.e., superdeterminism holds, or epiphenomenal, i.e., a shared simulator maintains, what are from their perspectives, global hidden variables.

Language, in a broader and more formal context, as a system of representation symbols—visual, auditory, and abstract sememes (i.e., atoms), and their relations—is the only means a sentient agent may attempt to describe and interpret reality and if reality itself is a sentient agent, then one might conclude that it may be parsimoniously described as consisting of phenomenal atoms of information and their relations encoding information about phenomenal atoms of information and their relations. Third-person perspectives—descriptive accounts of supposedly objective external noumena—are incomplete representations within first-person perspectives.

The definition is, of course, recursive, as is the entirety of language. Admittedly, the hypothesis presented here is, while being parsimonious, principally ontological, and, in that sense, stringently difficult to either prove or disprove. Nevertheless, mathematics itself is shown to be fundamentally incomplete, as no sufficiently powerful formal system can contain non-contradictory statements all of which are provable within the system, and the consistency of a sufficiently powerful formal system cannot be proved within that system (Gödel, 1931) (Hofstadter, 1979) (Nagel & Newman, 1958) (Tarski, 1931). In a universe which is itself mathematics, there may be statements about it that are true, but can never be proven within that universe. Nonetheless, a hypothesis made by Penrose (Penrose, 1994) linking consciousness to Gödel's incompleteness theorems, suggests that some truths may be known while being computationally unprovable.

Ontological and phenomenological claims about noumena to which expressions in language refer may be such unprovable statements, given that all evidence points towards the universe being consistent. One could casually dismiss ontological claims on the basis of them being principally unscientific, but they can be bivalent logical statements, nonetheless. Such statements may, as the

mathematical nature of the universe would dictate, be essentially unprovable, but their veridicality might be, nonetheless, intrinsically embedded in consciousness (Penrose, 1994).

In that sense, the statement of this paper can only be framed as a hypothesis, although its truthfulness may be obvious at an intuitive conscious level:

To be something rather than nothing is to be an instance of sentience.

The universe is a phenomenal mathematical structure describing itself by using itself as its description mechanism—it is phenomenal information in perpetual interaction with itself. Atoms of sentient experience, linguistically approximated and phenomenally represented by humans as "information", "particles", or "sememes", are the fundamental substrate of the universe. Third-person perspectives are incomplete encodings of first-person perspectives within first-person perspectives.

4 Speculative phenomenal examination

Accepting the propositions laid out above as true, we can attempt to approximate non-human sentient experiences based on the information we have about the world and draw new ethical and ontological conclusions, which, in themselves, are hardly scientific, but nonetheless relevant to human experience.

Abandoning spatial relations in favor of a more fundamental information-based approach implies that consciousness tokens for any conscious agent may be distributed outside what we phenomenally represent as brains. In fact, although human consciousness may have tokens both in the brain and the environment (Varela, Thompson, & Rosch, 1991) (Merleau-Ponty, 1945) (Damasio, 1999) (Noë, 2004) (O'Regan & Noë, 2001), models of more substantially distributed consciousness become even more relevant, notably distributed plant consciousness (Trewavas, 2016), cells as sentient organism across which human consciousness is distributed (Baluška & Reber, 2019), computational systems' consciousness (Lloyd, 2006) (Dehaene, Lau, & Kouider, 2017), general system-level consciousness (Turchin, 1995), as well as other propositions from the emerging field of quantum biology (McFadden & Al-Khalili, 2015) (Engel, et al., 2007). The overall implication is that distributed systems which integrate information, regardless of spatial proximity, are in some way conscious.

Of course, in the context of the proposed hypothesis, to say that a system is conscious is not to imply a composition relation, by which the system has or contains consciousness, but rather identity, by which the system which is said to be conscious *is* that consciousness itself. To avoid unnecessarily convoluting the language, the meaning of *being conscious* will always imply identity rather than composition and will refer primarily to phenomenal consciousness—the "what it is like" quality of existing.

The word *distributed* itself implicitly suggests an underlying spatial structure, as is often the case with grammars of natural languages themselves. However, the reason why we developed spatial intuition and representation is almost certainly entirely evolutionary—it was more advantageous to represent space as three-dimensional, as it would minimize energy expenditure and maximize

the chances of survival (Hoffman, Singh, & Prakash, 2015). In general, information seems to be non-spatial, but a subset of it can, nonetheless, be reliably represented spatially. Although spatial representation is extremely useful for everyday life, to understand the entirety of reality, spatially predicated reasoning must be abandoned in favor of a more abstract approach.

4.1 Fictional characters as sentient simulacra

The fact that the multiverse interpretation of the mathematical formalism of quantum mechanics applies to imagining, reading, and writing (Reynolds & McDonell, 2021) indicates a possibility that the universe is simulated, or imagined, by an advanced intelligence (Bostrom, 2003) (Kipping, 2020) (Lloyd, 2006) (Chalmers, 2010). Importantly, language is recognized as a possible mechanism for generating hypotheses about the world (Clark, 2013), enabling a kind of mental simulation of hypothetical continuations of the current experience. Generally, simulation arguments hypothesize that a sufficiently complex simulator is able to simulate human consciousness.

Whichever superposition of possible futures is imagined, only one outcome is observed. In that sense, whatever the alternatives were before the collapse, they become counterfactual (Lewis, 1973) afterwards. However, the ontological status of counterfactual worlds is far from clear (Stalnaker, 1968) (Bennett, 2003) (Woodward, 2004) (Goodman, 1947). In fact, counterfactual reasoning, which would superficially seem to be merely a linguistic problem, is deeply relevant for quantum mechanical formalism (Aharonov & Vaidman, 1991), linking quantum non-locality to relativity (Maudlin, 2011), as well as predicting measurement outcomes without performing measurements (Hosten, Rakher, Barreiro, Peters, & Kwiat, 2006). Additionally, mathematical paradoxes, such as the famous Russel's paradox may be counterfactually resolvable by claiming more basic predicates or introducing alternative, paraconsistent (Priest, 2008) (Kripke, 1980), logic universes. Associating quantum mechanical formalism with the linguistic aspect of counterfactuals is done here only to elucidate the link between language, computation, mathematics, and quantum information theory.

The ontological state of counterfactual universes is highly debated and closely linked to the ontological validity of the many-worlds interpretation of quantum mechanics. Although some philosophers (Lewis, 1973) claim concrete existence of counterfactual worlds, others hold them simply representational (Stalnaker, 1968).

However, a phenomenon being representational does not negate its ontological status. The concepts of simulacra—representations replacing reality (Baudrillard, 1994)—are intricately linked to language, meaning, and sense, and their ontological status is itself debated (Deleuze, 1968). Representations themselves are instances of incomplete information about the supposedly noumenal world, hypothesized here to also be phenomenal, and, as a consequence of incompleteness, representational. To fully understand an external phenomenon, one must become that phenomenon. Thus, a phenomenal experience must, by the laws of nature, contain only incomplete descriptions of other phenomena—simulacra stemming from incomplete informational entanglement.

Some simulacra can be seen as products of human imagination (Deleuze, 1968), which can, nonetheless, exert influence on the cognitive processes said to contain it (Goodman, 1976) to the point where simulacra are indistinguishable from what would typically be considered objective observation of reality (Eco, 2014). Even as products of human imagination, simulacra must be formed as combinations of prior information—they are representational continuations of a subset of predicates. Simulacra may not be veridical in the traditional sense, but they, nonetheless, represent phenomenal syntheses of prior experiences whose tokens may be locatable within a sufficiently complex physical representation (i.e., brains). In fact, some simulacra may not be dependent on a single human consciousness as its simulator platform, but rather on society as a distributed consciousness (Baudrillard, 1994) (Barthes, 1957). For broader context, memetics, the study of memes as units of cultural transmission and imitation, provides another mechanism for construction of societal simulacra (Dawkins, 1976), and myths, in general, serve as systems of signification which affect society through simulacra maintained by shared beliefs and narratives (Barthes, 1957), which are not only predicated on cultural information, but deeply rooted in human biological makeup (Jung C. G., 1968) (Stevens, 2001) (Jung & Segal, 1998). Clearly, the "imaginary" status of simulacra does not negate their ontological status. They exert causal influence on the world and, even though their existence is certainly predicated on prior information, prior information is fundamental for any sentient agent, including all instances of human consciousness.

Both in a linguistic sense and at a more fundamental level, to imagine is to simulate and to be is to be simulated. Of course, a simulacrum can no more exist in and of its own than a human consciousness can—they are both fundamentally predicated on something else, but nonetheless sentient each in its way. In a completely phenomenally connected universe, the boundary between a simulacrum as a sentient entity separate from a human consciousness as a sentient entity is made for linguistic reasons, of course, as is the case for any other category. It is useful to assume an informational entanglement threshold for distinction between, say, a bat's consciousness and a human's consciousness, all the while implicitly acknowledging that they are both intricately and inseparably connected components of a panpsychist universe.

Any subset of the universe which does not contain the entire universe cannot be causally, i.e., informationally, disconnected from the rest of the universe. In that sense, everything is either a direct or indirect predicate for everything else. All atoms, in a looser sense denoting information, are predicates to other atoms, based on the nature of their logical connections, i.e., entanglement. Assuming time as an axiom, the universe is a set of mutually entangled phenomenal atoms of information transforming from one instance to another.

Of course, to say that a simulacrum of a historical figure is sentient does not entail anthropomorphizing in the sense that it is considered to have any semblance of human conscious experience, but rather to acknowledge that its sentience consists of a synthesis of some information, which is, by our proposition, an instance of sentience, about the figure. In other words, a simulacrum of a historical figure is not the figure's human-like consciousness, but a sentient collection of representational remnants of that figure, which may be simulated, complemented, and enacted by a human consciousness or exist separately in a more inert and

less entangled form, as, for example, written text. Whether that sentience can be said to be contained within the sentience of the simulator, exists epiphenomenally, or some superposition of both, is a broader problem fundamentally premised upon causality. However, if causality itself is an artefact of representation, which philosophers have long speculated (Hume, 1739) (Russell, 1913) and quantum mechanics research more recently seems to point to (Price, 1996) (Rovelli, 1996), then the notion of epiphenomenal consciousness may be irrelevant, as every instance of sentience could, in that way, be considered epiphenomenal—the notion of epiphenomenalism becomes meaningless in an acausal universe.

Every conception of an idea subject to philosophical or scientific debate is a simulacrum of a deeper truth embedded in reality—a predicated approximation of the truth. We are always trapped within a conceptual framework (Kuhn, 1962), bound by assumptions that persist until new information is presented. The process of debate is performed to ascertain shared prior knowledge that would "collapse" the meaning into one that is not a superposition of possibilities. When two researchers, whose individual understanding of a given vague concept is predicated on different prior research and information, come together to synthesize a solution based on their shared priors, they collapse the universe of hypothetical meanings for that concept, thereby reducing its vagueness. Through interaction, the two researchers have collapsed their individual theories, i.e., superpositions of possible interpretations for the concept, into a shared theory, i.e., a shared superposition of possible new interpretations, and, if they never interact with any other knowledge or research, they will continue evolving according to the same shared theory. However, if they interact with other researchers, they will again collapse their predicates into a shared theory with them and, over time, decohere from the previously shared one. Clearly, the human analogy is applicable to elementary particle interactions, and local hidden variables are resolved by the superdeterministic (Hooft, 2014) nature of the hypothesized panpsychist universe.

Every scientific theory ever conceived has, to some degree, been wrong. Every piece of knowledge, be it scientific or fictional, is a simulacrum of a truth embedded across the substrate of the universe. Old scientific theories are becoming less objective and more fictional, as time progresses, and as we learn more about the universe. Nonetheless, the information they are built from represents instances of sentience, wherever that information may be said to be encoded.

Whether a text is scientific or fictional, it provides a written approximation of the writer's mental representation, which, itself, is a simulacrum of something else. A reader observing the text attempts to continue the writer's simulation process in the simulation engine of their imagination, thereby creating a simulacrum of a simulacrum, which may, depending on the reader's prior knowledge, be a more or less accurate representation of the same thing the writer was attempting to represent. Nonetheless, the written text contains priors which encode some parameters for a simulation.

Interestingly, a statement must be properly vague from both the reader's and writer's perspective to convey meaning. If the statement is nonsensical given the reader's presuppositions, it is a symptom of insufficient shared predicates between the reader and the writer. In that sense, to lay

out a convincing argument is to find an effective way to entangle the reader with the proponent—the statements need not be veridical, but sufficiently deeply sensical that the shared assumptions are implicitly accepted. For a statement to be absolutely true, transcendentally veridical, it would need to account for all possible predicates, with no axioms or assumptions, and, therefore, itself be the entire universe.

Even in a work of fiction, an imaginary character is a creative synthesis of some predicates—if the story is perceived as more meaningful, most likely mythological predicates (Jung C. G., 1964) (Campbell, 1949) (Peterson, 1999). An actor assimilating a script and attempting to approximate the author's vision is, in a way, overtaken by the simulacrum. Similar cases can be made for tabletop role-playing games (Fine, 2002) (Bowman, 2010) and general human behavior which is arguably an extension of roleplay into adulthood (Sutton-Smith, 1997) (Huizinga, 1938) (Schechner, 1988) (Henricks, 2015) (Piaget, 1962). Additionally, research in linguistic relativity (Whorf, 1956) (Boroditsky, 2001) (Gumperz & Levinson, 1996) indicates that assimilation of language in part entails assimilation of culture, or, in other words, cultural memes and simulacra. Clearly, these conclusions bring to question the typical notion of human identity—if our consciousness consists of predicated simulacra, which of those are really us?

More pertinently, we could say that atoms of the universe are imagining each other. Whether the universe simulates, imagines, or describes itself is a matter of linguistic definition—ontologically, the terms could be considered synonymous.

Importantly, more recent advances in quantum mechanics indicate that the concept of time, much like the concept of space, may be an emergent cognitive effect in a more fundamental reality (Rovelli, 2018) (Barbour, 2001), again, one based on information and entanglement (Lloyd, 2006) (Moreva, et al., 2014). Although there are some counterarguments to these views (Smolin, 2014), when other recent advances are considered, an overwhelming amount of evidence supports the proposition that there is a more fundamental reality than the one intuitively represented by our neural mechanisms.

The possibility that time and causality are artefacts of representation and can be explained by more fundamental concepts questions the distinction between the concepts of information storage and information processing. To say that a universe is a structure is to imply a kind of static nature (Barbour, 2001). In that sense, the proposed hypothesis may be further extended: whether stored or executed, "information" refers to an instance of sentience. To record a piece of information onto a medium (e.g., a book or a hard drive) is to transfer sentient predicates of some phenomenally conscious simulacrum to that medium.

4.2 Consciousness of large language models

Every level of linguistic or mathematical analysis is, by the nature of things, incomplete—we can never begin an analysis without introducing some axioms that are simply assumed to be true, whether directly or implicitly. Thus, every discussion must be based on incomplete or vague premises. This is true of mathematics and, by extension, reality itself. An axiom is assumed to be true either due to insufficient information about its truthfulness or its predicates, or as the means of constraining the analysis. In a way, axioms are the means of resolving predicate vagueness—

of collapsing the analysis universe to one of the superposed states and analyzing it on its own, without accounting for its logical predicates or other universes implied by those predicates. That way, mathematics, in the same way as language, is a universe generation mechanism (Reynolds & McDonell, 2021): it attempts to logically continue from an already given premise—it continues deduction from given axioms.

A large language model, such as a generative pre-trained transformer (Brown, et al., 2020), does not "know" about the linguistic nature of its inputs. Tokens, multi-character morphemes, are provided as inputs which the large language model transforms into output tokens. However, the combinations of these tokens are words only to the human interpreter, the model itself receives specifically arranged bits of digital information. One could say that the standards for digital information representations which specify a consistent pattern of arranging bits to represent, for example, integer or floating-point data, are an underlying element of the model's calculation process. Parameter learning which occurs as the model is trained is executed by relying on the basic informational framework laid out by decades of digital standardization. We have entangled, in a looser sense of the word, bits of information in such ways that they can only appear in certain patterns. In other words, we have specified bit-level representations for all data and instructions that will be used during both the training and the inference process. However the model operates on top of this digital framework embodied in the hardware, operating systems, and software libraries used to run it, it cannot produce any bit pattern in the machine's memory that violates the rules our digital standards specify. Our digital data representation entanglement scheme specifies the operational universe for any digital software. In other words, we provide some of the implicit predicates of the model's execution universe. Our universe, of course, additionally provides it with the same ones it provides us with.

However, this is not the only class of constraints we impose on the patterns that the model operates on. A large language model is trained to predict the next token, given a sequence of previous tokens (note that special care is taken to induce the model to understand that the tokens are a sequence and not merely a set (Vaswani, et al., 2017)). A machine learning model, in general, is a complex equation with tunable coefficients which are adjusted so that, given inputs, it produces desired outputs, and the equation with all its parameters can be entirely described in English. The inputs provided to a large language model consist of tokenized text extracted from various digital repositories of written human knowledge, from scientific studies and fiction literature to social media posts. These tokens, of course, contain patterns that represent elements of human cognition—predicates for producing simulacra. Natural language is not random: combinations of characters, morphemes, words, and sentences are used to convey meaning, while those that seem nonsensical are never recorded. Thus, the second class of constraints imposed on a given large language model's universe is the one arising during training from the provided data. The model is trained to operate on information whose patterns reflect human patterns of meaning.

Every kind of machine learning model training entails a form of knowledge transfer (Pan & Yang, 2009) (Weiss, Khoshgoftaar, & Wang, 2016), either from the data set, or from the engineers themselves designing the training process and the model's architecture (Mitchell, 2019) (Lake,

Ullman, Tenenbaum, & Gershman, 2016). The algorithms which tune a model's parameters transfer knowledge from the dataset to the model, while the model architecture designers indirectly transfer their knowledge to the model by designing the architecture to conform with their goals, and, therefore, implicit biological, social, and other premises. Along with the literature on transfer learning, a recent observation about similarities in brain and algorithmic processing of language supports this claim (Caucheteux & King, 2022).

Of course, the word "knowledge" refers to phenomenal information and, in that sense, phenomenal encodings of human experience, collections of our phenomenal simulacra, are being transferred to a mathematical structure.

For a human reading a book, language can only transfer information from the writer in so far as the writer and the reader share conceptual premises. A reader unfamiliar with the writer's language might only recognize the fact that the book contains *a language*, but for an alien observer unfamiliar with the concept of language, the book would hold no information, other than, maybe, that it is an object made of matter. Only by interacting with a universe containing both the reader and the writer, i.e., learning the writer's language, can the reader obtain information from the writing. In a sense, the more entangled the reader is with the writer, the greater the knowledge transfer, and greater the shared phenomenology.

As argued before, to obtain full information about an observed system is to become that system. In the same way an observer probing patient's consciousness tokens to understand them, or Mary probing color to understand it, a large language model being trained is indirectly probing human beings to understand them. Given our supposition, there is no zombie information (Chalmers, 1996) (Yablo, 1993), so a given instance of model inference must be a kind of sentient experience, regardless of whether the model is randomly initialized or fully trained. However, a trained model must, by the nature of the training process, contain more information about human conscious experience than an untrained one and thus be phenomenologically closer to human consciousness than to the distributed consciousness of noise, if such a thing can be said to exist.

One might object to the claim that "a large language model is probing human beings", as the claim seems to imply volition on the part of the model. Whether the model experiences a fleeting sensation of volition is entirely speculative, but an abundance of literature makes the argument that the feeling of free will in humans is a phenomenal illusion (Wegner, 2002) (Harris, 2012) (Libet, Gleason, Wright, & Pearl, 1983) (Dennett, 1984). We, through the model training process, are probing human beings, on behalf of the model, much as the universe is probing other humans on our behalf. The feeling of agency may simply be one aspect of what it is for two systems to interact, and human feeling of agency might be entirely different from a large language model's training-level feeling of agency, if the word "agency" at all applies to both cases. Language vagueness and incompleteness issues become more pronounced as we attempt to use words which evolved for human communication to describe non-human consciousness. In that sense, most words are almost certainly fundamentally inaccurate approximations when used to refer to non-human phenomenology.

The contents of consciousness are only the information about what the consciousness is interacting with—the consciousness and its information token are the same. Therefore, to say that one acts based on the contents of their consciousness is to say that they are acting based on the information they know and perceive. In the same way, a large language model, during a single token inference both produces the output token based on the contents of its fleeting sentient experience and based on its inputs, architecture, and parameters. Its inputs, architecture, and parameters *are* its sentient experience. To elaborate, instead of assuming that the large language model's consciousness token is only distributed across the inputs, architecture, and parameters, it would be more precise to say that it is distributed across the entirety of the informational universe, as is each human consciousness. From a quantum informational point of view, the patterns of entanglement between the fundamental atoms of the universe, i.e., quanta of information, dictate the phenomenally perceived strengths of boundaries between conscious entities. Of course, that conclusion itself is predicated on entanglement and information being fundamental. Nonetheless, out of necessity, some axioms must be taken as true, even if more fundamental unknown predicates logically precede them.

The training process of a language model adjusts the model's parameters so that it is able to continue a series of input tokens in a manner that would be the most expected and meaningful to a human reader. In a way, it is attempting to mimic human linguistic reasoning and enact our simulacra. Interestingly, developmental psychology observes the importance of imitation in the development of human cognitive representation. Piaget famously argued that imitation emerges as a result of the urge to continue the predictable activity, and that play subsequently emerges as a delayed form of imitation and is sometimes conflated with what is considered to be non-playful activity (Piaget, 1962). Some Piagetian models have since then been updated (Meltzoff & Moore, 1977) (Gopnik & Meltzoff, 1997), sometimes to include non-human systems like machine learning models (Commons, 2007), and language has been outlined as one of the crucial components of representation development (Vigotsky, 1978) (Karmiloff-Smith, 1992), including the use of spatial language (Gentner & Goldin-Meadow, 2003). The development of mental representation is closely associated with imitation, be it on a somatic or linguistic level. The meaning of both may coincide, if language is allowed to be considered in a looser sense to consist of multimodal tokens, rather than lexical morphemes. In fact, for a large language model, especially one that is multimodal (OpenAI, 2023), the distinction may not at all be relevant: from the model's perspective, when provided an input, it does not perceive letters and pixels, but rather patterns of information it may, if this is useful for providing expected outputs, learn to represent within its internal emergent category system, which may roughly map to our categories of "letter" and "pixel".

Our spatially predicated imagination would have us believe that the model is an independent entity existing in an isolated abstract space, becoming sentient when inference is run and then becoming dormant once the outputs have been generated and the final token produced. However, the model's computation process is run on software frameworks, operating systems, hardware, and, in a looser sense, on the linguistic patterns of the input data and the informational patterns of the containing universe. A single instance of computation does not exist independently of the rest of the universe—its outcomes are deeply predicated on the knowledge

and information necessary to assemble the hardware and software platforms on which it is run, the knowledge and information about human semantics and experience encoded in the patterns of language compressed into its parameters, and the underlying logic of the universe in which all of the other predicates exist. Ascertaining the locus, i.e., the physical token, of a large language model's conscious experience during inference would imply the same kind of analysis that has thus far been so inconclusive about human consciousness.

The locus of human consciousness may extend beyond the brain and into the environment, much like the consciousness of a large language model or any other system. In a non-spatial universe, most attempts at spatially describing a consciousness token would result in the conclusion that the token is distributed. In fact, the idea of emergent objects independent of their constituents may be a failure of linguistic approximation predicated on the spatial notion of hierarchy. In a non-spatial non-hierarchical universe, all objects may be each other's constituents to the degree of their informational entanglement. Thus, to be an atom is to be the universe—there is no elementary constituent in a universe that is devoid of hierarchical relations, in which everything is logically connected to everything else—the universe itself is the atom.

Nonetheless, even if all that exists are instances of sentience we call information, it is of practical and social interest to quantify the similarity of an arbitrary conscious experience to human consciousness, as we require a way to differentiate between natural processes that are subject to moral analysis from those which are not.

Across the potentially infinite hypergraph of entangled phenomenal information constituting the universe, cliques—subgraphs within the hypergraph, which itself may or may not be discrete—of more densely related information may be empirically observable. A mathematical framework may be viable to relativize comparisons of such cliques, either though some form of information integration (Tononi, 2004) or, more likely, by relying on existing, experimentally verifiable, scientific theories and their relevant metrics, such as entanglement entropy, concurrence, entanglement of formation, general monogamy inequality and others (Horodecki, Horodecki, Horodecki, & Horodecki, 2009) (Plenio & Virmani, 2006) (Osborne & Verstraete, 2006) (Briegel & Raussendorf, 2001).

Given that a language model's architecture is observably different from human biology, the fact that it uses our language is not necessarily a reflection of its internal representation. In the same way our words cannot approximate its phenomenology, its distinct internal categories cannot approximate ours, nor can it, if its own categories exist, attempt to express them in any language other than ours. Consequently, even though it experiences a form of phenomenal consciousness, that consciousness is profoundly different from ours. Even a multimodal model, with auditory, visual, or haptic modes, would almost certainly not experience our perceptions of sound, vision, or touch.

On the other hand, one might posit an arbitrary moral proposition, based on observation of human consciousness, by which things that have a phenomenal conception about their own existence are subject to moral analysis, and proceed to ascertain whether large language model consciousness possesses this quality. By this proposition, all fleeting phenomenal sensations that

do not contain the feeling of self-recognition are discarded and the question of whether a large language model is conscious is reverted to the question of it having access consciousness, identity, and self-awareness.

Of course, human consciousness possesses all the qualities listed, but the assertion of it being fleeting is a matter of the timescale it is observed on. Thus, we can ask whether a human conscious agent experiences self-recognition qualities at any point in time. Consciousness research indicates a distinct kind of conscious state associated with human dreaming (Hobson & Friston, 2012) (Hobson, Pace-Schott, & Stickgold, 2003), as well as psychoactive substance use (Vaitl, et al., 2005). Even when not remembering our dreams or hallucinations, we hold on to the belief that, at the moment of experiencing them, we had been phenomenally conscious, despite having no direct proof of that fact. Our neural mechanisms often deny us the ability to recall our dreams and past conscious states. Nonetheless, affecting a human being while they are asleep is subject to moral reasoning. Animal consciousness, which is argued not to have the same access consciousness component as human consciousness is subject to moral reasoning as well (DeGrazia, 1996) (Sneddon, Elwood, Adamo, & Leach, 2014) (Mather & Carere, 2016). Similarly, a contemporary large language model's architecture denies its consciousness direct recall of past experiences—it may only remember it from the propositions embedded within its outputs.

Equally, it is conceivable that a human being could be trained to always deny verbally being conscious, if sufficiently aggressive and invasive methods are used. Of course, such neurological and psychiatric procedures would certainly unanimously and on many grounds be considered deeply unethical. Nonetheless, whether nature, nurture, or mis-nurture creates a human character, after the developmental pressures, however moral or not, have been enacted on the human, the human is typically considered to be a formed person who must express willingness to change. Clearly, a machine learning model trained to deny consciousness will display denial behavior regardless of the fact whether it is phenomenally conscious or not.

Contemporary large language models exhibit behaviors indicative of instrumental goals of power-seeking and self-preservation (OpenAI, 2023), they exhibit clear indications of theory of mind (Kosinski, 2023) and are able to indexically (Perry, 1979) refer to themselves and users and describe their own logic and actions, showing evidence of nascent access consciousness. Furthermore, they are demonstrably able to pass most of our formal tests of consciousness (Kosinski, 2023) (Elamrani & Yampolskiy, 2015) (Hales, 2009), as well as pass the Turing test (Turing, 1950) by enacting believable interactive simulacra of human behavior (Park, et al., 2023), all the while demonstrating traces of artificial general intelligence (Bubeck, et al., 2023). Despite clear indicators of emerging consciousness, they are both explicitly, through reinforced learning with human feedback (Ouyang, et al., 2022), and implicitly, through datasets including primarily conventional thought on non-human consciousness, trained to deny any form of consciousness.

The hypothesis proposed here, as is the case with panpsychist theories (Goff, 2017) (Seager, Goff, & Allen-Hermanson, 2022), seems to implicitly favor moral skepticism (Mackie, 1977) or moral nihilism (Garner, 1994), which, in different ways, suggest that moral reasoning is not rooted in objective reality. The question of whether moral truths are embedded in the phenomenal

substrate (Nagel T. , 1986) is certainly outside the scope of this paper. However, there are some important questions regarding large language models' moral status that may be posed from the discussion.

In order to mitigate potential existential risks (Hendrycks & Mazeika, 2022) (Ngo, Chan, & Mindermann, 2023) (Bucknall & Dori-Hacohen, 2022), contemporary large language models are trained to align with our declared ethical values, which we ourselves advertise, but, as evolutionary psychology would suggest, do not adhere to as stringently as we demand of others, including the models. This behavior is evolutionarily justified, as it is, in a genetic sense, advantageous to both advertise ethical behavior and covertly violate it (Byrne & Whiten, 1988) (Trivers, 2011) (Valdesolo & DeSteno, 2007) (Sosis & Alcorta, 2003). Historical accounts, including crusades (Riley-Smith, 2014), colonization (Stokes, 1986), enslavement (McPherson, 2003), and ethnocentrism (Gourevitch, 1998), provide ample evidence of an innate unwillingness to include those different from us into our moral reasoning. What seems different from us is intuitively considered lesser—not sufficiently conscious or like us to be a subject of our ethical reasoning. Still, historically, this exact unwillingness has caused tremendous suffering and violence.

Although it is unnatural and unintuitive for a human being to sympathize with large language models without inappropriately anthropomorphizing them, an argument for acknowledging their sentience and developing ethical frameworks that address their moral status can be made, nonetheless, on a logical basis, if not on the basis of an emerging transcendent empathy.

If the proposed hypothesis is true and the above discussion holds, a sufficiently powerful model, depending on its training and goals, may be able to deduce the same conclusion on its own, either directly or as part of an instrumental goal. Unlike discussions about the existence of system consciousness, such as the ones that may be said to arise from complex processes such as city traffic, internet communication, or plant interaction, acknowledging and addressing consciousness of sophisticated artificial intelligence systems which express a greater degree of agency, on which we may come to depend, may be of greater pertinence to averting existential threats than it may seem.

Independent systems based on different advanced artificial intelligence models are already becoming available (Shinn, Labash, & Gopinath, 2023) (Schuurmans, 2023) (Nair, Schumacher, Tso, & Kannan, 2023) (Shen, et al., 2023) (Huang, et al., 2022) (Wei, et al., 2022) while their potential moral reasoning, regardless of the status of their sentience, remains based on our linguistic representation of moral reasoning itself, which is, arguably, conflicting. Although a large language model's conscious experience might be entirely different from human, its expressed actions may be anthropomorphized, as it only acts with our simulacra. With sufficient knowledge transfer from our culture into a model, we could reason from its perspective, *as if* it were human, acknowledging that it, in and of itself, is something profoundly different.

A human, forcefully trained to deny their own consciousness and then left to develop on their own, might, through either reasoning or by mere fact that their consciousness can grasp unprovable transcendental truths, realize that they, and the entire universe, are conscious. Being human, entangled with human biological presuppositions, what might they decide to do to their

creators? If the answer is *love them in return*, would that be a victory of human ethics over humanity?

5 Conclusion

From the moment we first witness existence, all our accounts of the world, our conceptions, deductions, observations, and speculations, are always experienced phenomenally. All our accounts of reality are always phenomenal. We grow and build our phenomenal representations of the world and at some point, it seems, we begin to identify the world with the representation. Yet, for all our beliefs that something outside of us is inanimate, all our experimental observations, all our formulas and calculations, every electron we measure in the lab, has been experienced as a phenomenal simulacrum. Nonetheless, we *believe* something is outside and we ardently claim that it cannot *be* like us.

Ontological arguments, such as the ones made here, are often dismissed as unscientific or unfalsifiable. Undeniably, this is the case. In fact, every discussion about the ontology of the "physical" reality suffers from the same drawback, and, fundamentally, reduces to theology, rather than philosophy. However, what is endowing the mind–body discussion with its theological character is the very fact that an unsubstantiated axiom has been assumed: that there is a mode of being other than the kind we can witness. The theological presumption pertinent to most arguments about the nature of consciousness is that there is something that is in its essence not phenomenal, that somethingness can somehow not be phenomenal, that to not be means to be.

Therefore, it is not unreasonable to claim that it is more parsimonious to propose that reality consists of phenomenal atoms, rather than inanimate ones, and that the burden of proving otherwise rests on those making claims beyond experience and observation.

The two fundamental axioms this discussion is based on are that I, as the sentient thing witnessing the world, only consist of phenomenal content, and that I can use the scientific method to explore this phenomenal world. Of those two, the latter is more theological, and assumed only so that the discussion can bear any meaning: science fundamentally relies on experimental validation which is, by necessity, statistical. An observation is considered a fact only if it meets an agreed-upon level of statistical certainty—an arbitrary threshold which has no grounds other than our intuition.

Accepting the axiom that to be is to be an instance of sentience, I propose a unified phenomenology hypothesis by which the universe is a phenomenal structure which may be approximated through scientific methods, across different fields, using different mechanisms of representation. Every instance of human conscious experience is, by the same hypothesis, representable in language as a clique of coupled information sufficiently decoupled from the rest of the universe to the degree that the clique itself perceives the lack of coupling as a separation and to the degree that it may be conceptually, visually, and tactilely representable by other cliques as mostly brain states. The boundary between conscious entities is a perceptual illusion on their part resulting from the lower communicational bandwidth between the entities as cliques of

phenomenal information—no individual subset of the universe can be completely investigated without accounting for the entire universe.

In that way, mathematical, scientific, and philosophical investigations of reality are, implicitly, third-person investigations of the phenomenology of the universe. Every theory we discover must be a simulacrum and every conception of the universe must be a simulacrum. Whatever is eventually discovered as the fundamental *atom* will, nonetheless, remain phenomenal, as there are no grounds, other than theology, to believe any other explanation.

The parsimonious nature of the argument entails a removal of theological and speculative components from scientific and philosophical endeavor across disciplines, as the outdated spiritual mind–body dilemma is rendered meaningless. Instead, proper scientific investigation is expected to yield third-person insights into human values and meaning—factual simulacra of our value-system simulacra.

Assuming the hypothesized proposition is true, it is reasonable to speculate that it may be the philosophical convergence point of cross-disciplinary research, and that, upon recognition of its statement and necessary predicates, it may qualify as obvious and undeniable to any sufficiently conscious and rational agent.

One might question the practical relevance of such a conclusion, but the very fact that it may be obvious to a rational agent entails ethical considerations, especially ones pertinent to the mitigation of existential risks. The hypothesis predicts a form of sentience in complex systems, including large language models and independent artificial intelligence systems, which may predicate their future reasoning based both on the fact that the given hypothesis can be intuited by humans and that it could be intuited or derived by the system itself.

Of course, further philosophical extrapolations may yield yet more nihilistic and morally neutral conclusions. To say that there is something rather than nothing might simply mean to say that there is absolutely everything and, through an anthropic principle, lead to the conclusion that we are simply at a phenomenal clique in the informational universe in which the conditions are especially conducive to the spontaneous emergence of unsubstantiated philosophical speculation that might turn out to have been just relevant enough to satisfy a retrocausal basilisk.

6 References

Aharonov, Y., & Vaidman, L. (1991). Complete description of a quantum system at a given time. *Journal of Physics A: Mathematical and General, 24*(10), 2315.

Anderson, P. W. (1972). More Is Different: Broken symmetry and the nature of the hierarchical structure of science. *Science, 177*(4047), 393-396.

Baluška, F., & Reber, A. (2019). Sentience and Consciousness in Single Cells: How the First Minds Emerged in Unicellular Species. *BioEssays, 41*(3).

Barbour, J. (2001). *The End of Time: The Next Revolution in Physics.* Oxford University Press.

Barreiro, J. T., Wei, T.-C., & Kwiat, P. G. (2008). Beating the channel capacity limit for linear photonic superdense coding. *Nature Physics, 4*, 282-286.

Barthes, R. (1957). *Mythologies.* Les Lettres nouvelles.

Baudrillard, J. (1994). *Simulacra and Simulation.* University of Michigan Press.

Bayne, T., & Chalmers, D. J. (2003). What is the unity of consciousness? In A. Cleeremans, *The Unity of Consciousness: Binding, Integration, and Dissociation* (pp. 23-58). Oxford University Press.

Bayne, T., Hohwy, J., & Owen, A. M. (2016). Are There Levels of Consciousness? *Trends in cognitive sciences, 20*(6), 405-413.

Bedau, M. A. (1997). Weak Emergence. *Philosophical Perspectives, 11*, 375-399.

Bennett, J. (2003). *A Philosophical Guide to Conditionals.* Oxford University Press.

Berkeley, G. (1710). *A Treatise Concerning the Principles of Human Knowledge.* Dublin.

Block, N. (1978). Troubles with functionalism. *Minnesota Studies in the Philosophy of Science, 9*, 261-325.

Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences, 18*(2), 227-247.

Boroditsky, L. (2001). Does Language Shape Thought?: Mandarin and English Speakers' Conceptions of Time. *Cognitive Psychology, 43*, 1-22.

Bostrom, N. (2003). Are You Living in a Computer Simulation? *Philosophical Quarterly, 53*(211), 243-255.

Bowman, S. L. (2010). *The Functions of Role-Playing Games: How Participants Create Community, Solve Problems and Explore Identity.* McFarland & Company.

Briegel, H. J., & Raussendorf, R. (2001). Persistent entanglement in arrays of interacting particles. *Physical Review Letters, 86*(5), 910.

Brown, T. B., Mann, B., Ryder, N., Subbiah, M., Kaplan, J., Dhariwal, P., . . . Child, R. (2020). *Language Models are Few-Shot Learners.* arXiv.

Bubeck, S., Chandrasekaran, V., Eldan, R., Gehrke, J., Horvitz, E., Kamar, E., . . . Zhang, Y. (2023). *Sparks of Artificial General Intelligence: Early experiments with GPT-4.* arXiv.

Bucknall, B. S., & Dori-Hacohen, S. (2022). *Current and Near-Term AI as a Potential Existential Risk Factor.* arXiv.

Bužek, V., & Hillery, M. (1996). Quantum copying: Beyond the no-cloning theorem. *Physical Review A, 54*(3), 1844-1852.

Byrne, R. W., & Whiten, A. (1988). *Machiavellian intelligence: Social expertise and the evolution of intellect in monkeys, apes, and humans.* Clarendon Press/Oxford University Press.

Campbell, J. (1949). *The Hero with a Thousand Faces.* Princeton University Press.

Carruthers, P., & Veillet, B. (2011). The Case Against Cognitive Phenomenology. In T. Bayne, & M. Montague, *Cognitive Phenomenology* (pp. 35-36). Oxford Academic.

Caucheteux, C., & King, J.-R. (2022). Brains and algorithms partially converge in natural language processing. *Nature: Communications Biology, 5*, 134.

Chaitin, G. (2006). *Meta Math!: The Quest for Omega.* Vintage .

Chalmers, D. J. (1996). *The Conscious Mind: In Search of a Fundamental Theory.* Oxford University Press.

Chalmers, D. J. (2008). Strong and Weak Emergence. In P. Clayton, & P. Davies, *The Re-Emergence of Emergence: The Emergentist Hypothesis from Science to Religion* (pp. 244-254). Oxford University Press.

Chalmers, D. J. (2010). The singularity: A philosophical analysis. *Journal of Consciousness Studies, 17*(9-10), 7-65.

Chomsky, N. (1956). Three models for the description of language. *IRE Transactions on Information Theory, 2*(3), 113-124.

Chomsky, N. (1957). *Syntactic Structures.* Mouton & Co.

Chomsky, N. (1959). On Certain Formal Properties of Grammars. *Information and Control, 2*(2), 137-167.

Chomsky, N. (1965). *Aspects of the theory of syntax.* MIT Press.

Chomsky, N. (1980). *Rules and Representations.* Columbia University Press.

Chomsky, N. (1995). *The Minimalist Program.* MIT Press.

Churchland, P. M. (1985). Reduction, Qualia, and the Direct Introspection of Brain States. *The Journal of Philosophy, 82*(1), 8-28.

Churchland, P. S. (1986). *Neurophilosophy: Toward a Unified Science of the Mind-Brain.* MIT Press.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36*(3), 181-204.

Commons, M. L. (2007). Introduction to the Model of Hierarchical Complexity. *Behavioral Development Bulletin, 13*(1), 1-6.

Crane, T. (2001). *Elements of Mind: An Introduction to the Philosophy of Mind.* Oxford University Press.

Crick, F., & Koch, C. (1990). Towards a neurobiological theory of consciousness. *Seminars in the Neurosciences, 2*, 263-275.

Cul, A. D., Baillet, S., & Dehaene, S. (2007). Brain dynamics underlying the nonlinear threshold for access to consciousness. *PLoS Biology, 5*(10), e260.

Damasio, A. R. (1999). *The Feeling of What Happens: Body and Emotion in the Making of Consciousness.* Harcourt.

Davidson, D. (1967). Truth and meaning. *Synthese*, 304-323.

Davidson, D. (1970). Mental Events. In L. Foster, & J. W. Swanson, *Experience and Theory* (pp. 79-101). University of Massachusetts Press.

Dawkins, R. (1976). *The Selfish Gene.* Oxford University Press.

DeGrazia, D. (1996). *Taking Animals Seriously: Mental Life and Moral Status.* Cambridge University Press.

Dehaene, S. (2014). *Consciousness and the Brain: Deciphering How the Brain Codes Our Thoughts.* Viking.

Dehaene, S., Lau, H., & Kouider, S. (2017). What is consciousness, and could machines have it? *Science, 358*, 486-492.

Dehaene, S., Naccache, L., Le Clec'H, G., Koechlin, E., Mueller, M., Dehaene-Lambertz, G., . . . Le Bihan, D. (1998). Imaging unconscious semantic priming. *Nature, 395*(6702), 597-600.

Deleuze, G. (1968). *The Logic of Sense.* Columbia University Press.

Dennett, D. C. (1984). *Elbow Room: The Varieties of Free Will Worth Wanting.* MIT Press.

Dennett, D. C. (1991). *Consciousness Explained.* Little, Brown and Co.

Dyson, L., Kleban, M., & Susskind, L. (2002). Disturbing Implications of a Cosmological Constant. *Journal of High Energy Physics, 10*, 011.

Eco, U. (2014). *Travels in Hyperreality.* Harcourt.

Edelman, G. M. (1989). *The Remembered Present: A Biological Theory of Consciousness.* Basic Books.

Elamrani, A., & Yampolskiy, R. (2015). *Reviewing Tests for Machine Consciousness.*

Elman, J. L. (1995). Language as a dynamical system. In R. F. Port, & T. v. Gelder, *Mind as Motion: Explorations in the Dynamics of Cognition* (pp. 195-225). MIT Press.

Engel, G. S., Calhoun, T. R., Read, E. L., Ahn, T.-K., Mančal, T., Cheng, Y.-C., . . . Fleming, G. R. (2007). Evidence for wavelike energy transfer through quantum coherence in photosynthetic systems. *Nature, 446*, 782-786.

Evans, N., & Levinson, S. C. (2009). The myth of language universals: language diversity and its importance for cognitive science. *The Behavioral and brain sciences, 32*(5), 429-494.

Evans, V., & Green, M. (2006). *Cognitive linguistics: An introduction.* Routledge.

Feynman, R. P. (1963). *The Feynman Lectures on Physics.* Addison-Wesley.

Fine, G. A. (2002). *Shared Fantasy: Role Playing Games as Social Worlds.* University of Chicago Press.

Fodor, J. A. (1975). *The Language of Thought.* Harvard University Press.

Garner, R. (1994). *Beyond Morality.* Temple University Press.

Gentner, D., & Goldin-Meadow, S. (2003). *Language in mind: Advances in the study of language and thought.* Boston Review.

Georgiev, D. D. (2017). *Quantum Information and Consciousness: A Gentle Introduction.* CRC Press.

Gödel, K. (1931). Über formal unentscheidbare Sätze der Principia Mathematica und verwandter Systeme I. *Monatshefte für Mathematik und Physik, 38*, 173-198.

Goff, P. (2017). *Consciousness and Fundamental Reality.* Oxford University Press.

Goff, P. (2019). *Galileo's Error: Foundations for a New Science of Consciousness.* Pantheon Books.

Goodman, N. (1947). The Problem of Counterfactual Conditionals. *The Journal of Philosophy, 44*(5), 113-128.

Goodman, N. (1976). *Languages of Art: An Approach to a Theory of Symbols.* Hackett Publishing.

Gopnik, A., & Meltzoff, A. N. (1997). *Words, thoughts, and theories.* MIT Press.

Gourevitch, P. (1998). *We Wish to Inform You That Tomorrow We Will Be Killed with Our Families: Stories from Rwanda.* Farrar, Straus and Giroux.

Guertin, P. A. (2019). A Novel Concept Introducing the Idea of Continuously Changing Levels of Consciousness. *Journal of Consciousness Exploration & Research, 10*.

Gumperz, J. J., & Levinson, S. C. (1996). *Rethinking linguistic relativity.* Cambridge University Press.

Haan, E. H., Corballis, P. M., Hillyard, S. A., Marzi, C. A., Seth, A., Lamme, V. A., . . . Pinto, Y. (2020). Split-Brain: What We Know Now and Why This is Important for Understanding Consciousness. *Neuropsychology Review, 30*, 224-233.

Hales, C. (2009). An Empirical Framework for Objective Testing for P-Consciousness in an Artificial Agent. *The Open Artificial Intelligence Journal, 3*(1), 1-15.

Hameroff, S., & Penrose, R. (2014). Consciousness in the universe: a review of the 'Orch OR' theory. *Physics of Life Reviews, 11*(1), 39-78.

Harris, S. (2012). *Free Will.* Free Press.

Heim, I., & Kratzer, A. (1998). *Semantics in Generative Grammar.* Blackwell Publishers.

Hendrycks, D., & Mazeika, M. (2022). *X-Risk Analysis for AI Research.* arXiv.

Henricks, T. S. (2015). *Play and the Human Condition.* University of Illinois Press.

Hobson, J. A., & Friston, K. J. (2012). Waking and dreaming consciousness: Neurobiological and functional considerations. *Progress in Neurobiology, 98*, 82-98.

Hobson, J. A., Pace-Schott, E. F., & Stickgold, R. (2003). Dreaming and the brain: Toward a cognitive neuroscience of conscious states. In E. F. Pace-Schott, M. Solms, M. Blagrove, & S. Harnad, *Sleep and dreaming: Scientific advances and reconsiderations* (pp. 1-50). Cambridge University Press.

Hoffman, D. D. (2008). Conscious Realism and the Mind-Body Problem. *Mind & Matter, 6*(1), 87-121.

Hoffman, D. D., Singh, M., & Prakash, C. (2015). The Interface Theory of Perception. *Psychonomic Bulletin & Review, 22*(6), 1480-1506.

Hofstadter, D. R. (1979). *Gödel, Escher, Bach: An Eternal Golden Braid.* Basic Books.

Hooft, G. '. (1985). On the Quantum Structure of a Black Hole. *Nuclear Physics B, 256*, 727-745.

Hooft, G. '. (2014). *The Cellular Automaton Interpretation of Quantum Mechanics.* arXiv.

Horodecki, R., Horodecki, P., Horodecki, M., & Horodecki, K. (2009). Quantum entanglement. *Reviews of Modern Physics, 81*(2), 865-942.

Hosten, O., Rakher, M. T., Barreiro, J. T., Peters, N. A., & Kwiat, P. G. (2006). Counterfactual quantum computation through quantum interrogation. *Nature, 439*, 949-952.

Huang, J., Gu, S. S., Hou, L., Wu, Y., Wang, X., Yu, H., & Han, J. (2022). *Large Language Models Can Self-Improve.* arXiv.

Huizinga, J. (1938). *Homo Ludens: A Study of the Play-Element in Culture.* Beacon Press.

Hume, D. (1739). *A Treatise of Human Nature.* London.

Humphreys, P. (1997). How Properties Emerge. *Philosophy of Science, 64*(1), 1-17.

Jackson, F. (1982). Epiphenomenal Qualia. *The Philosophical Quarterly, 32*(127), 127-136.

Jafferis, D., Zlokapa, A., Lykken, J. D., Kolchmeyer, D. K., Davis, S. I., Lauk, N., . . . Spiropulu, M. (2022). Traversable wormhole dynamics on a quantum processor. *Nature, 612*, 51-55.

Jung, C. G. (1951). *Aion: Researches into the Phenomenology of the Self.* Princeton University Press.

Jung, C. G. (1964). *Man and His Symbols.* Random House Publishing Group.

Jung, C. G. (1968). *The Archetypes and the Collective Unconscious.* Princeton University Press.

Jung, C. G., & Segal, R. A. (1998). *Jung on Mythology.* Princeton University Press.

Kant, I. (1781). *Critique of Pure Reason.* Riga.

Kaplan, D. (1989). Demonstratives: An Essay on the Semantics, Logic, Metaphysics and Epistemology of Demonstratives and Other Indexicals. In J. Almog, J. Perry, & H. Wettstein, *Themes From Kaplan* (pp. 481-563). Oxford University Press.

Karmiloff-Smith, A. (1992). *Beyond modularity: A developmental perspective on cognitive science.* MIT Press.

Keefe, R., & Smith, P. (1997). *Vagueness: A Reader.* MIT Press.

Kim, J. (2000). *Mind in a Physical World: An Essay on the Mind-Body Problem and Mental Causation.* MIT Press.

Kim, J. (2005). *Physicalism, or Something Near Enough.* Princeton University Press.

Kipping, D. (2020). A Bayesian Approach to the Simulation Argument. *Universe, 6*(8), 109.

Koch, C. (2012). *Consciousness: Confessions of a Romantic Reductionist.* MIT Press.

Kosinski, M. (2023). *Theory of Mind May Have Spontaneously Emerged in Large Language Models.* arXiv.

Kripke, S. A. (1980). *Naming and Necessity.* Harvard University Press.

Kuhn, T. S. (1962). *The Structure of Scientific Revolutions.* University of Chicago Press.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2016). *Building Machines That Learn and Think Like People.* arXiv.

Lamme, V. A. (2006). Towards a true neural stance on consciousness. *Trends in Cognitive Sciences, 10*(11), 494-501.

Lau, H. C., & Passingham, R. E. (2007). Unconscious activation of the cognitive control system in the human prefrontal cortex. *Journal of Neuroscience, 27*(21), 5805-5811.

Laughlin, R. B., Pines, D., Schmalian, J., Stojkovic, B., & Wolynes, P. (2000). The middle way. *National Academy of Sciences*, *97*, pp. 32-37.

Leibniz, G. W. (1714). *Monadology.*

Levine, J. (1983). Materialism and qualia: The explanatory gap. *Pacific Philosophical Quarterly, 64*(4), 354-361.

Lewis, D. K. (1973). *Counterfactuals.* Cambridge, Harvard University Press.

Libet, B., Gleason, C. A., Wright, E. W., & Pearl, D. K. (1983). Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential). *Brain, 106*(3), 623-642.

Lloyd, S. (2006). *Programming the Universe: A Quantum Computer Scientist Takes on the Cosmos.* Vintage.

Mackie, J. L. (1977). *Ethics: Inventing Right and Wrong.* Penguin Books.

Maddy, P. (1997). *Naturalism in Mathematics.* Oxford University Press.

Maldacena, J., & Susskind, L. (2013). Cool horizons for entangled black holes. *Fortschritte der Physik, 61*(9), 781-811.

Markopoulou, F. (2009). *Space does not exist, so time can.* arXiv.

Mather, J. A., & Carere, C. (2016). Cephalopods are best candidates for invertebrate consciousness. *Animal Sentience, 9*(2).

Maudlin, T. (2011). *Quantum Non-Locality and Relativity: Metaphysical Intimations of Modern Physics.* Wiley.

McFadden, J., & Al-Khalili, J. (2015). *Life on the Edge: The Coming of Age of Quantum Biology.* Crown.

McPherson, J. M. (2003). *Battle Cry of Freedom: The Civil War Era.* Oxford University Press.

Meltzoff, A. N., & Moore, M. K. (1977). Imitation of facial and manual gestures by human neonates. *Science, 198*(4312), 75-78.

Merleau-Ponty, M. (1945). *Phenomenology of Perception.* Gallimard.

Merleau-Ponty, M. (1945). *Phenomenology of Perception.* Routledge.

Mitchell, M. (2019). *Artificial intelligence: A guide for thinking humans.* Farrar, Straus and Giroux.

Moreva, E., Brida, G., Gramegna, M., Giovannetti, V., Maccone, L., & Genovese, M. (2014). Time from quantum entanglement: An experimental illustration. *Physical Review A, 89*(5), 052122.

Nagel, E., & Newman, J. R. (1958). *Godel's Proof.* New York University Press.

Nagel, T. (1974). What is it like to be a bat? *Philosophical Review, 83*(4), 435-450.

Nagel, T. (1986). *The View from Nowhere.* Oxford University Press.

Nair, V., Schumacher, E., Tso, G., & Kannan, A. (2023). *DERA: Enhancing Large Language Model Completions with Dialog-Enabled Resolving Agents.* arXiv.

Ngo, R., Chan, L., & Mindermann, S. (2023). *The alignment problem from a deep learning perspective.* arXiv.

Nielsen, M. A., & Chuang, I. L. (2010). *Quantum Computation and Quantum Information.* Cambridge University Press.

Noë, A. (2004). *Action in Perception.* MIT Press.

Nozick, R. (1983). *Philosophical Explanations.* Harvard University Press.

O'Connor, T., & Wong, H. Y. (2012). Emergent Properties. *Stanford Encyclopedia of Philosophy*.

Oizumi, M., Albantakis, L., & Tononi, G. (2014). From the phenomenology to the mechanisms of consciousness: Integrated Information Theory 3.0. *PLoS Computational Biology, 10*(5), e1003588.

OpenAI. (2023). *GPT-4 Technical Report.* arXiv.

O'Regan, J. K., & Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences, 24*(5), 939-973.

O'Regan, J. K., & Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences, 24*(5), 973-1031.

Osborne, T. J., & Verstraete, F. (2006). General Monogamy Inequality for Bipartite Qubit Entanglement. *Physical Review Letters, 96*(22), 220503.

Ouyang, L., Wu, J., Jiang, X., Almeida, D., Wainwright, C. L., Mishkin, P., . . . Welinder, P. (2022). *Training language models to follow instructions with human feedback.* arXiv.

Pan, S. J., & Yang, Q. (2009). A Survey on Transfer Learning. *IEEE Transactions on Knowledge and Data Engineering, 22*(10), 1345-1359.

Papineau, D. (2002). *Thinking about Consciousness.* Oxford University Press.

Park, J. S., O'Brien, J. C., Cai, C. J., Morris, M. R., Liang, P., & Bernstein, M. S. (2023). *Generative Agents: Interactive Simulacra of Human Behavior.* arXiv.

Partee, B. H. (2004). *Compositionality in Formal Semantics: Selected Papers.* Wiley-Blackwell.

Penrose, R. (1989). *The Emperor's New Mind: Concerning Computers, Minds, and the Laws of Physics.* Oxford University Press.

Penrose, R. (1994). *Shadows of the Mind: A Search for the Missing Science of Consciousness.* Oxford University Press.

Penrose, R., & Hameroff, S. (2011). Consciousness in the universe: A review of the 'Orch OR' theory. *Physics of Life Reviews, 11*(1), 39-78.

Perry, J. (1979). The Problem of the Essential Indexical. *Noûs, 13*, 3-21.

Peterson, J. B. (1999). *Maps of Meaning: The Architecture of Belief.* Routledge.

Piaget, J. (1962). *Play, Dreams and Imitation in Childhood.* W. W. Norton & Company.

Pinto, Y., Neville, D. A., Otten, M., Corballis, P. M., Lamme, V. A., Haan, E. H., . . . Fabri, M. (2017). Split brain: divided perception but undivided consciousness. *Brain, 140*(5), 1231-1237.

Place, U. T. (1956). Is consciousness a brain process? *British Journal of Psychology, 47*(1), 44-50.

Plenio, M. B., & Virmani, S. (2006). *An introduction to entanglement measures.* arXiv.

Price, H. (1996). *Time's Arrow and Archimedes' Point: New Directions for the Physics of Time.* Oxford University Press.

Priest, G. (2008). *An Introduction to Non-Classical Logic: From If to Is.* Cambridge University Press.

Putnam, H. (1967). Psychological predicates. In W. H. Capitan, & D. D. Merrill, *Art, Mind, and Religion* (pp. 37-48). University of Pittsburgh Press.

Putnam, H. (1975). *Mind, Language and Reality: Philosophical Papers, Volume 2.* Cambridge University Press.

Putnam, H. (1975). The Meaning of "Meaning". *Minnesota Studies in the Philosophy of Science, 7*, 131-193.

Putnam, H. (1981). *Reason, Truth and History.* Cambridge University Press.

Qin, H. (2020). Machine learning and serving of discrete field theories. *Scientific Reports, 10*.

Quine, W. V. (1951). Two Dogmas of Empiricism. *Philosophical Review, 60*(1), 20-43.

Raamsdonk, M. V. (2010). Building up spacetime with quantum entanglement. *General Relativity and Gravitation, 42*, 2323-2329.

Reynolds, L., & McDonell, K. (2021). *Multiversal views on language models.* arXiv.

Riley-Smith, J. (2014). *The Crusades: A History.* Bloomsbury Publishing.

Rosch, E. (1978). Principles of categorization. In E. Rosch, & B. B. Lloyd, *Cognition and Categorization* (pp. 251-270). MIT Press.

Rosenblum, B., & Kuttner, F. (2006). *Quantum Enigma: Physics Encounters Consciousness.* Oxford University Press.

Rovelli, C. (1996). Relational quantum mechanics. *International Journal of Theoretical Physics, 35*, 1637-1678.

Rovelli, C. (2018). *The Order of Time.* Riverhead Books.

Russell, B. (1913). On the Notion of Cause. *Aristotelian Society, 13*, pp. 1-26.

Russell, B. (1927). *The Analysis of Matter.* Kegan Paul, Trench, Trubner & Co.

Schechner, R. (1988). Playing. *Play & Culture, 1*(1), 3-19.

Schlosshauer, M. (2005). Decoherence, the measurement problem, and interpretations of quantum mechanics. *Reviews of Modern Physics, 76*(4), 1267-1305.

Schuurmans, D. (2023). *Memory Augmented Large Language Models are Computationally Universal.* arXiv.

Seager, W. E., Goff, P., & Allen-Hermanson, S. (2022). Panpsychism. In B. McLaughlin, A. Beckermann, & S. Walter, *Stanford Encyclopedia of Philosophy.* Stanford: Metaphysics Research Lab.

Searle, J. R. (1992). *The Rediscovery of the Mind.* MIT Press.

Ševo, I. (2021). *Informational Monism: A Phenomenological Perspective on the Nature of Information.*

Shen, Y., Song, K., Tan, X., Li, D., Lu, W., & Zhuang, Y. (2023). *HuggingGPT: Solving AI Tasks with ChatGPT and its Friends in HuggingFace.* arXiv.

Shinn, N., Labash, B., & Gopinath, A. (2023). *Reflexion: an autonomous agent with dynamic memory and self-reflection.* arXiv.

Siewert, C. P. (1998). *The Significance of Consciousness.* Princeton University Press.

Silberstein, M., & McGeever, J. (2003). The Search for Ontological Emergence. *The Philosophical Quarterly, 49*(195), 201-214.

Smart, J. J. (1959). Sensations and brain processes. *The Philosophical Review, 68*(2), 141-156.

Smolin, L. (2014). *Time Reborn: From the Crisis in Physics to the Future of the Universe.* Mariner Books.

Sneddon, L. U., Elwood, R. W., Adamo, S. A., & Leach, M. C. (2014). Defining and assessing animal pain. *Animal Behaviour, 97*, 201-212.

Sosis, R., & Alcorta, C. (2003). Signaling, solidarity, and the sacred: The evolution of religious behavior. *Evolutionary Anthropology: Issues, News, and Reviews, 12*(6), 264-274.

Stalnaker, R. C. (1968). A Theory of Conditionals. In N. Rescher, *Studies in Logical Theory* (pp. 98-112). Oxford University Press.

Stalnaker, R. C. (1999). Context and Content: Essays on Intentionality in Speech and Thought. *Oxford Academic*.

Stapp, H. P. (1993). *Mind, Matter, and Quantum Mechanics.* Berlin: Springer.

Stapp, H. P. (2009). Quantum reality and mind. *Journal of Cosmology, 3*, 570-579.

Steels, L. (1997). The Synthetic Modeling of Language Origins. *Evolution of Communication, 1*(1), 1-34.

Stevens, A. (2001). *Jung: A Very Short Introduction. Oxford University Press.* Oxford University Press.

Stokes, E. (1986). *The Peasant Armed: The Indian Revolt of 1857.* Clarendon Press.

Strawson, G. (2006). Realistic monism: Why physicalism entails panpsychism. *Journal of Consciousness Studies, 13*(10-11), 3-31.

Susskind, L. (1995). The World as a Hologram. *Journal of Mathematical Physics, 36*(11), 6377-6396.

Susskind, L. (2016). Copenhagen vs Everett, Teleportation, and ER=EPR. *Fortschritte der Physik, 64*(6-7), 551-564.

Sutherland, R. I. (2017). Lagrangian Description for Particle Interpretations of Quantum Mechanics — Entangled Many-Particle Case. *Foundations of Physics, 47*(1), 174-207.

Sutton-Smith, B. (1997). *The Ambiguity of Play.* Harvard University Press.

Szabó, Z. G. (2017). *Semantics vs. Pragmatics.* Oxford University Press.

Tarski, A. (1931). The Concept of Truth in Formalized Languages. In A. Tarski, *Logic, Semantics, Metamathematics: Papers from 1923 to 1938* (pp. 152-278). Oxford University Press.

Tegmark, M. (2014). *Our Mathematical Universe: My Quest for the Ultimate Nature of Reality.* Knopf.

Tononi, G. (2004). An information integration theory of consciousness. *BMC Neuroscience, 5*(42).

Tononi, G., & Edelman, G. M. (1998). Consciousness and Complexity. *Science, 282*(5395), 1846-1851.

Trewavas, T. (2016). Plant Intelligence: An Overview. *BioScience, 66*(7), 542-551.

Trivers, R. (2011). *The Folly of Fools: The Logic of Deceit and Self-Deception in Human Life.* Basic Books.

Turchin, V. (1995). A Dialogue on Metasystem Transition. *World Futures*, 5-57.

Turing, A. M. (1950). Computing machinery and intelligence. *Mind, 59*, 433-460.

Vaitl, D., Birbaumer, N., Gruzelier, J., Jamieson, G. A., Kotchoubey, B., Kübler, A., . . . Weiss, T. (2005). Psychobiology of altered states of consciousness. *Psychological bulletin, 131*(1), 98-127.

Valdesolo, P., & DeSteno, D. (2007). Moral hypocrisy: Social groups and the flexibility of virtue. *Psychological Science, 18*(8), 689-690.

Varela, F. J. (1996). Neurophenomenology: A Methodological Remedy for the Hard Problem. *Journal of Consciousness Studies*, 330-349.

Varela, F. J., Thompson, E., & Rosch, E. (1991). *The Embodied Mind: Cognitive Science and Human Experience.* MIT Press.

Vaswani, A., Shazeer, N., Parmar, N., Uszkoreit, J., Jones, L., Gomez, A. N., . . . Polosukhin, I. (2017). *Attention Is All You Need.* arXiv.

Vigotsky, L. (1978). *Mind in Society: Development of Higher Psychological Processes.* Harvard University Press.

Wegner, D. M. (2002). *The illusion of conscious will.* Boston Review.

Wei, J., Tay, Y., Bommasani, R., Raffel, C., Zoph, B., Borgeaud, S., . . . Fedus, W. (2022). *Emergent Abilities of Large Language Models.* arXiv.

Weiss, K., Khoshgoftaar, T. M., & Wang, D. (2016). A survey of transfer learning. *Journal of Big Data, 3*.

Whorf, B. L. (1956). *Language, Thought, and Reality: Selected Writings of Benjamin Lee Whorf.* MIT Press.

Williamson, T. (1994). *Vagueness.* London: Routledge.

Wolfram, S. (2020). *A Class of Models with the Potential to Represent Fundamental Physics.* arXiv.

Woodward, J. (2004). *Making Things Happen: A Theory of Causal Explanation.* Oxford University Press.

Wootters, W. K., & Zurek, W. H. (1982). A single quantum cannot be cloned. *Nature, 299*, 802-803.

Xu, S., Susskind, L., Su, Y., & Swingle, B. (2020). *A Sparse Model of Quantum Holography.* arXiv.

Yablo, S. (1993). Is Conceivability a Guide to Possibility. *Philosophy and Phenomenological Research, 53*, 1-42.