Qualia as Fundamental Observational Limits: A Computational Epistemological Approach to Consciousness

Zied Ben Houidi

Abstract

Drawing on recent advances in neurosciences and artificial intelligence, I argue for a new epistemological and ontological shift that aligns, in its initial line of inquiry, with the tradition of George Berkeley's empirical idealism, but ultimately departs from it in its conclusions: Sounds, images of objects, shapes, lines, points, words, ideas, what we call matter, emotions, logic, theories, energy, waves, everything: Everything that can be (more rigorously has been) *said* or *lived* is an *observation* that leaves us with a characteristic mental impression. It can be a (i) live observation (e.g. perception), (ii) a replay of prior observations (e.g. visual imagery) or (iii) a replay of the mixture of prior independent observations (i.e. imagination). (We can observe that) Some observations are undefinable (e.g. a dot or point) and (that) others are definable by the invocation of related prior observations. I postulate that knowledge is stored in the same format in which it was acquired, as observations, and it is re-observed when re-invoked. That's why it's reusable and useful (for future similar observations). Words, in the forms of observable visual signs (letters) or auditory signals (sounds), are also observables that observably can act as triggers to cause the observable mental visualization of prior observables (that have been by time and learning tied together). "Will" is an observable that observably results from the observable strength of an upcoming observable action or thought. Self is another observable, and action too. We are nothing but observers: "I am acting" is an observation of my current status, and "I want to act" is another one. All observables range from "undefinable" to more or less definable by the invocation of prior non-definable observables. More interestingly, concepts such as "defining" or "explaining" or even "logic" itself are also observables that observably lie in the middle of this chain of observables, which limits their abilities and applicability, with dramatic consequences on eternal philosophical questions such as the mind-body or free-will, but also on hard sciences, mathematics, logic and physics.

Keywords: Artificial intelligence; Natural intelligence; Consciousness; Qualia; Philosophy of mind; Idealism; Materialism

1 Summary

As humans, we have long sought to understand the mysteries of the world and our place in it. We have asked questions about the nature of reality, the nature of consciousness and mental states such as pain and joy, the nature of will and whether it is free, and what is the relationship between the mind and the body, if they are distinct entities at all. Many have provided answers throughout history, putting forward convincing arguments, but none of them seems to satisfy us today, for the debate is still wide open.

On one end of the spectrum of answers, we have materialism, which holds that reality is ultimately based on matter or physical objects, as opposed to being mental or spiritual in nature. Throughout history, materialists have argued that everything including our experiences and thoughts are ultimately reducible to physical processes. Materialism, which is survived by modern variations such as physicalism, is still the dominant perspective in philosophy and science, drawing its strength from the success of scientific theories in explaining a wide range of physical phenomena, including brain processes.

On the other end, we have idealism, which opposes that reality is fundamentally mental or spiritual in nature. Throughout history, idealists have defended a wide range of views, from Plato's idea that the material world is an imitation of an eternal, abstract realm of perfect forms, to Kant's philosophy that the mind actively shapes and structures our experience of the world, and that the world as it appears to us is a product of our mental faculties. Although a minority, idealist positions continue to be skillfully defended by contemporary thinkers, either through definite views such as panapsychisms (Chalmers et al. 2015), or more integrative versions such as property dualism. Such positions owe their strength to what we feel as the existence of mental properties, or qualia (**?**),[1] that cannot be explained in terms of mechanics or particular arrangements of atoms. Such "explanatory gap" (Levine 1983) is a still-standing old argument known also as Leibniz's gap (Cummins 2010, Leibniz 1989). It refers to the impossibility to explain subjective perception leveraging only mechanical processes involving, says Leibniz, "figures and motion". Today, this position is supported by arguments such as the knowledge argument (Mary's room thought experiment) (**?**) and the older inverted spectrum argument. This insurmountable problem stands as the "hard problem of consciousness" (Chalmers 1995).

These are only the two extreme ends of the spectrum as there exists a plethora of other philosophies. For example, refusing to adopt one view or the other, monism asserted that reality is a single, unified whole that cannot be divided into separate parts. Dualism occupied a middle ground, holding that the mind and body are distinct onto-

[1] subjective, first-person experiences of sensory phenomena such as pain, pleasure, and color.

logical entities that interact with each other. More recently, property dualism admits that there is only physical substance but that there are distinct mental and physical properties. Panapsychism, resp. panexperientalism, asserts in its simplest form that consciousness, resp. experience, is a fundamental aspect of the universe and is present in all things, including inanimate objects. Emergentism, in contrast, asserts that higher-level phenomena such as consciousness emerge from lower-level phenomena, such as physical processes in the brain. This view opposes both those who hold that consciousness is a fundamental aspect of the universe, and physicalists who defend that higher-level mental states can be strictly reduced to physical states of the brain. But there, even within physicalists, we find variations: type physicalism, or reductive materialism, holds that every type of mental state (e.g. seeing a red color) corresponds to a type of physical events in the brain (a unique neural activity pattern). Token identity physicalism, known as anomalous monism, accepts that every individual instance of a mental state corresponds to a physical state in the brain, but argues that the "mental is anomalous", e.g. different mental states may be realized by the same physical state.

Far from being exhaustive, these are only few examples of the many opposing views whose still-active existence is telling about the difficulty of the problem. As a new glimmer of hope, recent advances in artificial intelligence (AI) and neuroscience entertained a promise to finally give definite answers to such eternal questions, by studying the neural correlates of consciousness, or shedding light on the foundations of intelligence. Ultimately, instead of a resolution, they brought more puzzles, further reviving the flames of the debate.

First, in the field of AI, the target of creating a machine that can truly understand and think like a human seems getting closer, at least in appearance, with spectacular performance achieved by large language models (Lemoine 2022).[2] But a question persists: beyond performance, do large language models understand? Despite decades of debate, John Searle's Chinese Room thought experiment (Searle 1980) still stands to remind us that mechanically following a set of rules may not equate to true understanding, echoing again the question of the nature of understanding and consciousness.

The same applies for the various neurosciences which are still faced, despite spectacular advancements, with the challenge of understanding consciousness and subjective experience. In general, an overwhelming evidence has shown a tight correspondence between particular experiences and precise neural activation patterns, and this, across various modalities such as perception, touch, emotions etc. A seminal work is that of Hubel and Wiesel (Hubel & Wiesel 1959) that demonstrated the existence of simple cells, i.e. neurons that retinotopically respond selectively to edges that are oriented in a specific direction. Crucially, if these cells do not learn such patterns on-time during early stages

[2]Some AI engineers were seriously wandering whether an AI such as Lamda is sentient (Lemoine 2022).

of development, then the subject will lack the corresponding subjective ability. This goes also the other way around: the artificial activation of neurons can cause subjective experience. For example, visual hallucinations such as phosphenes were experimentally induced through transcranial stimulation of the visual cortex (Kammer 1998, Kammer et al. 2005, Antal et al. 2003). However, not everything that fires leads to a conscious or subjective experience, even when it has a computationally obvious function. Blindsight (Weiskrantz et al. 1974) is one of such discoveries that brought new puzzles to the problem: it refers to the phenomenon by which individuals who are cortically blind (e.g. due to partial damage to the primary visual cortex) are still able accurately judge visual stimuli presented in their blind field, without any conscious awareness. This means that the brain can successfully process visual information without conscious awareness. If this is the case, then one valid question is why is consciousness even needed? The global workspace theory (Baars 2005) descriptively accommodates this apparent paradox by postulating that some computational processes are unconscious, handing over information to others which are. But in the end, what makes a process conscious and others not, remains a mystery.

In this paper, reconciling the above views, I attempt to redesign our cartography of knowledge by putting again under the spotlight the fact that it is ultimately thanks to our thoughts (and feelings), let them be the result of a physicalist brain or a substance on their own, that we barely say anything about the problems above. As opposed to the ontological one, my epistemological "cogito" will read as follows: Everything that has been said, our collective knowledge, all our theories are (what most of us have learned to name) human thoughts of some sort, or what I call *observations that leave us with unique mental impressions*. This axiom has high chances of being accepted by both the physicalist for whom the mental states are the product of the brain, and by the idealist for whom the thoughts are ontologically separate.

In accepting that everything that has been said is some sort of ideas, I am not going to conclude, following traditional forms of idealism, that everything that exists is fundamentally mental or spiritual in essence. In what I call an epistemological empirical idealism, I will only accept that everything we ever said or lived is some sort of experience, observations that leave us with impressions. Learning from recent advances in neurosciences and artificial intelligence, I argue for a new epistemological and ontological shift that aligns, in its initial line of inquiry, with the tradition of George Berkeley's empirical idealism, but ultimately departs from it in its conclusions:

Sounds, images of objects, shapes, lines, points, words, ideas, what we call matter, emotions, logic, theories, energy, waves, everything: Everything that can be (more rigorously has been) *said* or *lived* is an *observation* that leaves us with a characteristic mental impression. It can be a (i) live observation (e.g. perception), (ii) a replay of prior observations (e.g. visual imagery) or (iii) a replay of the mixture of prior independent

observations (i.e. imagination). (We can observe that) Some observations are undefinable (e.g. a dot or point) and (that) others are definable by the invocation of related prior observations. I postulate that knowledge is stored in the same format in which it was acquired, as observations, and it is re-observed when re-invoked. That's why it's reusable and useful (for future similar observations).

Words, in the forms of observable visual signs (letters) or auditory signals (sounds), are also observables that observably can act as triggers to cause the observable mental visualization of prior observables (that have been by time and learning tied together). "Will" is an observable that observably results from the observable strength of an upcoming observable action or thought. Self is another observable, and action too. We are nothing but observers: "I am acting" is an observation of my current status, and "I want to act" is another one. All observables range from "undefinable" to more or less definable by the invocation of prior non-definable observables.

More interestingly, concepts such as "defining" or "explaining" or even "logic" itself are also observables that observably lie in the middle of this chain of observables, which limits their abilities and applicability, with dramatic consequences on eternal philosophical questions such as the mind-body or free-will, but also on hard sciences, mathematics, logic and physics.

A key insight of this new cartography is that our understanding of the world is intimately shaped by the way in which we extract and transform these observations. Our ability to describe and understand the world is built on a hierarchy of increasingly complex features or observations. For example, in vision, the brain extracts simple features such as edges and lines, which are then combined to form more complex features such as shapes and objects. These complex features, in turn, are used to construct even more complex representations of the world, such as scenes and concepts. This hierarchical structure is reflected in the way we organize and store our knowledge, and it parallels neuroscience findings regarding the extraction of complex features in the brain.

I argue that it is thanks to such observations (which occur to us due to our innate abilities) that we later arrived at conclusions about the existence of the self, the outside world, and reality itself. As we did not create our own mental faculties, but rather discovered them as we became aware of them, we all observe the products and abilities of our intelligence. Before we were even conscious of philosophical conundrums, we observed our ability to create chains of causes and effects. This perspective allows us to reexamine the positions of both physicalists and idealists: the physicalist first observes thoughts, then infers that the brain is behind them, while the idealist observes that thoughts predate matter but struggles to fit mental states into a framework of "shapes and motions".

The main contribution of this paper is to outline how "fundamental" observations/qualia are, not as an essence, but as our limit and only way with which we can describe ourselves and the world: everything we say is a manipulation of such observations or qualia. From

this perspective, I argue that we face a fundamental limitation in our ability to explain and understand consciousness and qualia. If explaining means transforming one concept into another or showing a qualia to a qualia-capable receptor, then we are fundamentally unable to explain the nature of qualia itself. This leads to my central claim: we are, in a sense, "doomed." If explainability is the act of transforming one concept into another, then we are condemned to never fully understand the nature of consciousness and qualia. We have hit a fundamental limit of our computational brain.

This realization has profound implications for our philosophical inquiries and scientific endeavors. From this starting point, I derive the limits of explainability and understanding, distinguishing between two main types of explanation and understanding. The first is deep across-modality explanation, for example explaining a word by showing the corresponding image or situation. The second, language-related, is relative text-to-text, i.e., explaining a word by invoking other words. Both of these approaches, however, ultimately fall short when we attempt to explain the fundamental nature of qualia and consciousness.

To make a case for my point and support my theory, I will leverage recent advances in various neuroscientific disciplines as well as prowesses in artificial intelligence. The latter have shed light on interesting computational foundations of human and artificial intelligence alike, showing curiously striking similarities between both, which I will map in the paper. By studying the computational foundations of human and artificial intelligence, we can gain a deeper understanding of these fundamental limitations and their implications for our quest to understand consciousness and reality.

Crucially, we must recognize that explaining itself is an observable ability – a mental operation that our brains can execute. When we explain the concept of a dog, for instance, we can show an image of a dog, and this explanation is effective. We engage in a kind of "transfer learning," applying our explanatory abilities across different domains. However, when we attempt to apply this same ability to explain qualia itself, we find ourselves at an impasse. The mental operation of explanation exists and is observable, but it fails to transfer or apply effectively to the very foundation of our subjective experience. This failure is not due to a lack of effort or intelligence, but rather it points to a fundamental limitation in the nature of explanation itself when confronted with the bedrock of consciousness.

Understanding the nature of our thoughts and how they are formed, stored, and retrieved is crucial to understanding the philosophical questions we pose and the answers we propose. This approach allows us to examine not just the content of our philosophical inquiries, but also the very tools and processes we use to conduct these inquiries. By doing so, we may gain new insights into age-old questions and perhaps even reformulate them in more productive ways, while also recognizing the inherent limitations we face.

In this paper, I will first provide a synthetic interpretation of these findings in what

could be a theory of natural and artificial general intelligence. I'll show how research from various disciplines corroborates this theory. From there, I will draw a set of fundamental limitations that any thinking exercise or enterprise will have to confront, concluding with these limitations and their implications for our understanding of consciousness, reality, and the nature of thought itself. This exploration will underscore the paradoxical nature of our quest for understanding: even as we push the boundaries of our knowledge, we must confront the possibility that complete understanding of consciousness and qualia may forever remain beyond our grasp.

Draft Work in Progress

This document is a work in progress. The sections beyond this point are not included in this version.