A critique of Theories of Consciousness: What constitutes a satisfactory theory?

Lo Min Choong Julian

U22******

Philosophy, Nanyang Technological University

HY3011 Philosophy of Mind

Assignment 1: Final Essay

Prof. Teru Miyake

20 April 2023

Table of contents

Author's note

2023-05-27T08:04:59Z

I am uncertain whether I have been successful in making an explicit argument. Nevertheless, I like this submission, and I like this course. I got to explore (extremely superficially) the science of consciousness, and to explore an area where philosophy and science are intertwining so deeply. We also got to hear and talk to Prof. Bayne (because he came to NTU on 10 April 2023). It was fun.

On hindsight, I am uncertain whether IIT really is an axiomatic ToC, and if it is, whether it suffers from Gödel's Incompleteness Theorems.

Perhaps, I am mistaken, and that my arguments are flawed because I misunderstand the epistemological problem of measurement (i.e. how can we know our theory is correct when we use the theory to confirm itself; e.g. using a theory of temperature to confirm whether Kelvin/Celcius/Fahrenheit is successful). My main doubts, and thus the core of my argument is to what extent can a ToC be objective, given that past unsuccessful ToCs have always been shaped by human desires. Perhaps my doubts will be put to rest, or become well defined when I eventually take HY3018 Epistemology and HY3010 Philosophy of Science.

A critique of Theories of Consciousness

What does a successful theory of consciousness look like? What are its aspects? Will it answer all the questions we want about consciousness? In this essay, I argue that certain questions are currently unanswerable in the philosophy of mind. Specifically, it is currently unanswerable by contemporary Theories of Consciousness [ToCs] because we do not have enough information to justify a stance satisfactorily. In the process, I shall discuss our expectations of ToCs and whether we should modify our expectations or not.

To begin, I shall examine the central problems put out by the standard narrative of the philosophy of mind (substance dualism to physical monism). Then, I shall examine contemporary ToCs. This will set the proper context for what central problems should be within the domain of well-developed ToCs.

The narrative of the philosophy of mind

Descartes' central problem was "What is the Mind?". Before him, no one gave a positive description of the mind. He wanted to do just that. Thus, he proposed substance dualism. He argues the mind has no "physical extension", i.e. is not a physical object. However, it has the property of being able to affect the physical world. The mind is able to have a causal effect on the body it resides in. His explanation indeed tackles his central problem. However, we are unsatisfied because it is unscientific. No one, now, agrees that a non-physical thing can have causal relations with the physical world.

Thus, the central problem evolved. People were then interested in answering the same question, in addition to being constrained by physicalism — the idea that the mind is (must be) a physical object because to argue that the mind is non-physical is to regress to substance dualism. Thus, physical monism dominates the narrative post-Descartes.

Behaviourism argues that if something acts like a mind, it is a mind. More precisely, "a particular mental state is a disposition to behave in a certain way" (Tanney, 2022). The problem with this is what exactly is a dispositional link? How is it different from a causal link, or a correlation link? Are mental states the sole cause of behaviours? Are mental states necessary but insufficient causes of behaviours?

Critics argue behaviourism is imprecise. Consider the following: someone is in pain. Their nervous system is in a biochemical state of pain. This causes the appropriate mental state of pain. However, this person hides their pain, for whatever reason. Is this person in pain or not? One may give charity to behaviourism and conclude: yes, this person is in pain, but is pretending not to be in pain. This gives a shaky causal relation between mental states and behaviours.

Identity theory attempts to rectify problems with behaviourism. It argues that mental states are brain states, and that brain states are mental states, only (Smart, 2022). This theory is supported by some empirical evidence — patients with specific brain damage result in measurable, drastic changes in personality, e.g. Phineas Gage. However, identity theory conflates types and tokens. Suppose a specific brain state is the only brain state associated with pain. Every human ever will have the exact same brain state while in pain. Does that mean everything that feels pain has the exact same biochemical processes and brain states as ours? Critics argue that this is too narrow a definition. Pain is not necessarily, solely, caused by one specific brain state. Therefore, identity theory is false.

Functionalism argues that what sufficiently constitutes "The Mind" is anything that has the same function as it (Levin, 2023). It does not matter what the mind is actually made of. For example, if we could create something of pure non-biological substance that perfectly functions as a neuron, the thing we created is a neuron. Thus, functionalism allows isomorphism.

Functionalism also rests on computability — the mental processes of our mind are computational processes, and any sufficiently advanced technology will have the same functionality as our minds, and thus is a mind. However, is this really the case? Are minds purely computational machines?

Each of these theories attempts to give a complete description of the mind. It is uncontroversial to say that Behaviorism and Identity theory fails. It is controversial to say functionalism fails because its main philosophical critique is phenomenology; specifically, qualia.

Qualia: Questioning the narrative

Qualia is significant because it casts doubt on physical monism. Specifically, it questions whether experiences can be fully explained by physicalism. Whether it succeeds or fails is arguable.

Qualia is not necessarily incompatible with physical monism (Tye, 2021). Those who do argue it is necessarily incompatible argues the following:

P1. Physical Monism argues that everything that exists is made of physical stuff.

P2. Qualia or experiences are non-physical stuff.

C1. Therefore, either qualia exist, or physical monism is true.

The weak link in this argument is P2 — qualia is poorly defined. There is a severe lack of consensus within the philosophical community regarding what qualia is. Most will accept that qualia is a catch-all term that captures our vague intuitions about our experiences.

One interpretation of qualia is the discordance of the report of the perception of x, and the perceptual experience of x (Jackson, 1982, p. 130). For example, Mary and I both see the colour red. I have the "correct" perceptual experience of red, as it is shared by everyone else, but Mary has a different perceptual experience of it. However, despite her differing perceptual experience, she still reports "red" because her perceptual experiences are "messed up" in such a way that does not produce differing reports of the same thing.

This interpretation highlights the problem with qualia. (1) Qualia is poorly defined. It rests on our intuitions of phenomenology and experiences. (2) The argument is not valid. Just because something is conceivable does not mean it is possible. Just because I can conceive of someone having differing perceptual experiences, does not mean that there exists someone who actually does have different perceptual experiences. (3) What does "perceptual experience" mean?

I can have a "correct perceptual experience", while Mary can have a "messed up perceptual experience" because "perceptual experiences" exist. If "perceptual experiences" do not exist, inverted qualia evaporate. So, what exactly is "perceptual experience"? It is a vague term used to refer to our underlying intuitions. It points towards our intuition — we have subjective experiences. But our intuitions are not rigorous pieces of evidence. It may guide the direction of our inquiry, but its purpose ends there.

The problem with any philosophical investigation into qualia is that it is not empirically verifiable. We have no way of accessing, exploring, examining, quantifying, or extracting such experiences because, by definition, qualia is inaccessible to external observation.

Perhaps this problem of vague definability is specific to this interpretation of qualia. Perhaps another interpretation of qualia will work. However, Dennett has provided a sufficient attack against qualia, arguing that there exists nothing that is "ineffable, intrinsic, private, directly apprehensible properties of experience" (Dennett, 1988). Perhaps there can exist something that is directly apprehensible to properties of experience, except it is not private. This will allow for our experiences to be consistent with physicalism.

Should we even consider qualia? Should a ToC explain qualia?

Potential progress: Naturalistic dualism

At this point, one would be within reason to ask if we should reevaluate the central problems. This is exactly what Chalmers did in 1995. He proposed the distinction between the purely physical descriptions of consciousness, and the phenomenal experiences of consciousness. He argues that if and when we do arrive at the pure physical descriptions of our mind, we will not, necessarily, have anything to say about our conscious experience. This is the "hard problem of consciousness".

Is Chalmers right? How can we tell?

Naturalistic dualism is promising because it integrates phenomenal aspects into physical descriptions, thus providing a comprehensive theory of mind. But is this really what a comprehensive theory requires? The basis of this being a comprehensive theory is our (biassed) intuitions. Therefore, should we reject the notion that a satisfactory theory of mind must include the explanation of experiences?

The problem with mandating the inclusion of experiences is the fact that no one has provided a satisfactory definition, i.e. a definitive positive description, of experiences. Are experiences necessarily subjective? Or, are we mistaken for believing that experiences are subjective? Is there a finitely large number of experiences that are accessible to any conscious being, thus perpetuating our false intuitions because we do not have the resources to properly categorise and compare experiences in a clear, precise and objective manner?

Perhaps we should consider the meta-philosophical question: What questions should we be asking ourselves to guide our approach to the philosophy of mind? Do we have the resources to tackle the current central problems? Do we need to reevaluate our methodologies? Do we need to integrate philosophy and science in order to make a breakthrough?

Are some questions invalid?

Can we upload our consciousness to the cloud, specifically to a non-biological machine? Is that upload a simulation of our consciousness, or actually our consciousness? Does the upload even retain the property of being conscious?

I shall put aside the concern of whether these questions are answerable or not. Instead, the point I am trying to make by bringing up these questions is: Should a ToC address this problem to be counted as a successful ToC?

Seth & Bayne: Theories of Consciousness

Contemporary scientific investigations into consciousness are classified as the development of "Theories of Consciousness [ToCs]" (Seth et al., 2022). Seth and Bayne argue there are four main ToCs: (1) Higher Order Theories (HOTs), (2) Global Workspaces Theories (GWTs), (3) Integrated Information Theory (IIT), (4) and Re-entry and predictive processing theories.

The fourth is poorly defined as it is a cluster of related theories grouped together because they share a similar-ish doctrine: the emphasis on "the importance of top down signalling in shaping and enabling conscious perception." (p. 445).

As the contemporary field of ToCs is new, Seth and Bayne wrote this paper to establish a framework for ToCs so that scientists and philosophers have a "common language". Without this framework, there is a hermeneutic gap. This hermeneutic gap must be bridged in order for discussions surrounding ToCs to occur. After all, collaboration and discussions are the main drivers of innovation.

The problem with comparing each theory with another is that the methodologies, domains of investigation and end results are so distinct from each other that comparison is infeasible. To enumerate this, I shall focus on IIT.

For example, IIT is an axiomatic theory that "derives claims about the properties that any physical substrate of consciousness must satisfy." (p. 444). "Any system that generates a non-zero maximum of (irreducible) integrated information is conscious, at least to some degree." (p. 444). "it equates an organism's level of consciousness with its value of Φ" (p. 444).

Thus, IIT is grounded in mathematics, the axiomatic system, and information. It necessitates the measurability and quantification of consciousness because its core axioms (pre-)supposes that consciousness is measurable. This is problematic because it builds on the assumption that there exist phenomenological truths that are axiomise-able. But, it might not be fatal if it is vindicated in the future.

Since IIT equates consciousness to Φ, we off-load the problem of consciousness to the problem of Φ. Therefore, if we prove or disprove Φ, we either prove the measurability of consciousness, or disprove the causal connection between consciousness and Φ.

However, there are several deeper problems with IIT.

Φ is incomputable. It is "infeasible to measure, except in simple model systems" (p. 445). Only approximations of Φ are practically possible. Thus, more problems arise. How would we know if the approximations are really approximations of Φ? Precision is lost during approximation — how much precision can we afford to lose to accurately describe consciousness? How do we determine the best approximation of Φ, given there are multiple methods of approximating Φ?

Since IIT is an axiomatic theory, would it not be susceptible to Gödel's Incompleteness Theorems? Gödel argues that any axiomatic system cannot be complete, consistent and decidable simultaneously (Raatikainen, 2022). Thus, IIT fails to be satisfactory even if all its axioms are proven to be true because its insights and conclusions are lacking in one of these three dimensions.

Consider another potential problem. Suppose Φ is true, and there exists an approximation that is measurable, computable and useful for describing consciousness, and thus, is a sufficient resource for us to discern what is conscious and what is not. Is that all there is to say about consciousness?

As Φ is exclusive to IIT, how else can we prove IIT to be true? Can we prove a theory's status through comparison, or are empirical studies sufficient? Just because Φ is true, it does not necessarily mean it is incompatible with other theories. IIT's Φ and HOT's meta-representations may not be mutually exclusive. As such, how can we test for mutual exclusivity?

Each ToC has different empirical approaches, methodologies, and domains of inquiry (p. 439). Thus, each ToC has multiple points of "failure". Each ToC also has the potential to agree with each other, when there are overlaps in domains of inquiry, and if pieces of empirical data support more than one theory.

To evaluate a ToC, we are not necessarily limited to the contents of each theory or the framework provided by Seth and Bayne. Using the relational view of data, we can contextualise ToCs as models in the data journeys of consciousness (Beaulieu & Leonelli, 2021, p. 59-60). Under this, a ToC can be criticised if we discover we have limited knowledge of consciousness because of the theory employed. This allows us to diagnose the problem (limited knowledge output) and rectify it (expand the theory so that it gives us more knowledge). If we find our interactions with the world are not fully articulated by a theory, we then have grounds to expand a theory.

In short, we have three areas of evaluation: (1.) a theory's contents, (2.) the framework a theory resides in, and (3.) the data journey of a theory.

However, are these the only areas of evaluation we have at our disposal? Are these areas of evaluation sufficient as evaluative standards?

Successful ToCs: Consistency and novel predictions

What philosophers really want is to understand the fundamental nature of consciousness. To get there, they want a theory that is consistent with our beliefs and understanding of the natural world, and one that makes novel claims (Seth & Bayne, 2022, p. 449; Bayne, personal communication, 2023, April 10).

What should a Theory of Conscious be? What problems should it address?

Are we asking too much of ToCs? Just as Mathematics has its limits and appropriate applications, so do ToCs. The underlying thesis behind this notion, or vague intuition, is epistemological: there exists knowledge that is inaccessible to us. While belonging to epistemology, philosophers and scientists must keep this in mind, and evaluate their ToCs and frameworks accordingly. How to do so is a question for epistemologists to answer, and thus incorporate into the developing methodologies in the investigation of consciousness.

As we have seen in the history of the philosophy of mind, central problems are replaceable. Domains of inquiries are dynamic and evolving. ToCs ought to keep up with such changes. This in itself is uncontroversial. The problem is: What should be incorporated?

Should a ToC address qualia? IIT does not directly address experience since that is not its goal. HOTs do, in an indirect manner via meta-representations. But, this is arguable since meta-representations have yet to be fully defined. GWTs, similarly, do not consider experiences directly. It only considers it a by-product or a potential end result. Its primary concern is with the structure of the brain and how consciousness resides in it. Thus, GWTs and re-entry and predictive processing theories are siblings because they explicitly and primarily address what structures consciousness might necessitate.

The precise articulation of this concern is: Should a ToC address feature x, and to what extent should it incorporate feature x? Should feature x be the primary focus or not (and does it even matter)? Should problem x be addressed by the ideal ToC? What is an ideal ToC?

Sci-fi fantasy questions such as "Can I and should I upload my consciousness" have philosophical significance (Chalmers, 2010). But does every philosophically significant problem need to be addressed by an ideal ToC?

Conclusion

Personally, of the four main ToCs examined by Seth and Bayne, I like IIT the most because it is understandable, articulate, and the most concrete. Perhaps it is because my dispositions are aligned with mathematics, thus, being grounded in mathematics, appears to me as more concrete compared to terms such as "meta-representations".

As consciousness is highly complex, it is safely arguable that well-developed ToCs should be jointly developed by scientists and philosophers.

Thus, my stance is that it is currently undecidable, because there is not enough information to properly justify any stance.

The motivation for the standards of an ideal ToC stem from seemingly arbitrary sources. We desire a satisfactory ToC to be consistent with our knowledge of the natural world, and to provide us with novel predictions. However, the metaphysical state of the world may not be in line with our desires. A complete description of consciousness may be inaccessible to us. Thus, we may arrive at a ToC that is consistent and has novel predictions but lack the means to recognise it as a satisfying ToC. Even now, we do not have any indication of whether this is the case or not.

Functionalism is promising, but it feels incomplete due to its explanation gap regarding descriptions of experiences. Naturalistic dualism is similarly promising but is currently incomplete due to the current state of qualia. Contemporary ToCs are the most promising by virtue of being grounded in the scientific method, and are novel in their own right. As I believe IIT is the most well-developed (in terms of concreteness) while being highly problematic, I can only speculate that a well-developed ToC would incorporate some form of information theory, or integrated information theory as a key aspect.

References

Beaulieu, A., Leonelli, S. (2022). *Data and Society: A Critical Introduction*. United Kingdom: SAGE Publications. ISBN: 9781529732542. https://uk.sagepub.com/en-gb/eur/data-and-society/book269709.

Chalmers, D. J. (2010). *Mind Uploading: A Philosophical Analysis. Journal of Consciousness Studies, 17*, pp. 7-65. https://consc.net/papers/uploading.pdf.

Chalmers, D. J. (1995). Facing Up To The Problem Of Consciousness. *Journal of Consciousness Studies, 2*(3), pp. 200-19. https://consc.net/papers/facing.pdf.

Dennett, D. C. (1988). Quining Qualia. *Consciousness in Modern Science*. Oxford University Press. https://philpapers.org/rec/DENQQ.

Jackson, F. (1982, April). Epiphenomenal Qualia. *The Philosophical Quarterly, 32*(127). http://www.jstor.org/stable/2960077.

Levin, J. (2023, June 21). *Functionalism*. The Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/archives/sum2023/entries/functionalism/.

Raatikainen, P. (2022, January 21). *Gödel's Incompleteness Theorems*. The Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/archives/spr2022/entries/goedel-incompleteness/.

Seth, A. K., Bayne, T. (2022, May 3). Theories of consciousness. *Nature Reviews Neuroscience, 23*, pp. 439-52. https://doi.org/10.1038/s41583-022-00587-4.

Smart, J. J. C. (2022, December 21). *The Mind/Brain Identity Theory*. The Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/archives/win2022/entries/mind-identity/.

Tanney, J. (2022, June 21). *Gilbert Ryle*. The Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/archives/sum2022/entries/ryle/.

Tye, M. (2021, September 21). *Qualia*. The Stanford Encyclopedia of Philosophy. https://plato.stanford.edu/archives/fall2021/entries/qualia/.