**Project title: Exploring the Connections between Assessed Level of Consciousness and Phi**

**Paper structure:**

1. **Introduction**
   1. Why could it be relevant to examine the connection between behavior and consciousness? And why is it difficult?

Relevant sources: [1] Dawkins: multiple paths with different levels of consciousness can lead to the same behavior. [2] Gutfreund: Evolutionary selection works on external behavior, not internal states. Hence, one could be led to assume that consciousness will indeed have causal effects on behavior, as it would not have evolved otherwise. [3] Griffin: Conscious and nonconscious systems could have similar capabilities.3

* 1. How does the following project (and importantly the experiment) attempt to contribute to our understanding of the link between behavior and consciousness?
  2. A specific note on IIT as an intrinsic theory of consciousness; within the framework of IIT specifically, why is it interesting to try to relate measures of internal (phenomenal) consciousness with externally observable behavior?

1. **A brief introduction to IIT and a conceptual understanding of Phi: What are we measuring anyways?**  
   1. Go through the basic structure of IIT 3.0, from axioms to formalized measures.
2. **Presenting the experiment and its methodology**
3. **Data analysis and results**
4. **Discussion**

**1. Introduction**

Inquiries into the connections between consciousness and behavior are hindered by both conceptual and methodological problems. Conceptually, speculation into possible behavioral correlates of consciousness is complicated not only by the difficulty of defining and identifying consciousness itself, but also by the array of scientific findings demonstrating how diverse processes playing out at different levels of consciousness and intentionality can result in seemingly identical human behavior (Dawkins, 2017). Methodologically, this becomes a problem if one seeks to establish unambiguous links between a specific behavior and any type of conscious mental state, whether that connection is expressed in terms of cooccurrence or causation. Additionally, there is the problem of the inherent hiddenness or intrinsicality of consciousness, meaning that it is tricky to establish an experimental baseline-level of consciousness on which to test the supposed behavioral correlates. This is especially problematic in situations where one cannot rely on the report of human research participants, such as in situations where people are unable to introspect (e.g., due to brain impairments), or in research dealing with animals[[1]](#footnote-1) or computer systems where introspective report is unavailable (Oizumi et al., 2014).

Despite these difficulties, the question of consciousness and behavior seems essential from an evolutionary point of view. If only those features that increases fitness are selected for, and if natural selection acts on an organism’s actions or behavior[[2]](#footnote-2) as opposed to its phenomenal experience of that action (Gutfreund, 2018), knowing more about this relationship would seem important for developing an understanding of the evolutionary origins of consciousness. Gutfreund sketches out a two-camp division among those interested in the link between consciousness and behavior. In the first camp, consciousness is seen as being directly (causally?) associated with fitness-increasing behavior, whereas the second sees consciousness as a ‘byproduct’ or an epiphenomenon associated with other biological features that directly increases fitness and are thus subjects to the process of natural selection.

Side-stepping this ontological discussion, consciousness can in both cases be seen as a biological property that in the weakest formulation co-occurs with fitness-increasing behavior; a link that is in principle amenable for experimental research if not for the conceptual and methodological problems outlined above. The present paper utilizes concepts and measures from Integrated Information Theory (IIT) to overcome at least one of these conceptual-methodological problems (albeit the success of the solution is contingent on IIT, in turns creating a rather circular scenario). Concretely, an experiment is designed and conducted to explore whether a measure for integrated information signified *phi* (), postulated to reflect the level of consciousness within a given system, is systematically related to a number of human participants’ ratings of that system’s assessed consciousness, based on observations of its behavior.

Possible findings from this experimental set-up can be evaluated at three levels: Internal to the framework of IIT, the experiment can be seen as an inquiry into the possible relationship between *phi,* conceptualized as a measure for intrinsic consciousness, and the human rating of the external behavior of the system, from which phi is computed. Even if IIT contains an explicit skepticism to estimating consciousness from observable behavior (Oizumi et al., 2014), the consciousness-behavior link remains important considering large scientific questions such as the evolutionary origin of consciousness (as presented in the introduction), not to mention the historical interest in discriminating between conscious and non-conscious beings from their observable behavior, as exemplified in the popularity of Turing tests (though Turing’s test strictly does not test consciousness, nor claims to).

Outside of the IIT framework, this experimental set-up could cast light on possible behavioral patterns associated with different levels of phi, now serving as a proxy for consciousness. This level is necessarily more speculative as it depends on the strength of the postulate central to IIT, namely that phi as a formal operationalization of intrinsic information reflects the consciousness of a system. A final and somewhat auxiliary research question is which factors in general (but not necessarily generalizable outside of the particular system used for this experiment) correlate with the participants’ ratings. Here, the main object of inquiry becomes the participants themselves, or more specifically *how* they assign consciousness to a system based on its observable behavioral qualities.

At all three levels, it is important to keep in mind that consciousness is never explicitly revealed but remains (in principle) intrinsic to the rated system, thus hidden from the observers. In practice, there might very well be no consciousness involved in the system’s behavior at all [see the discussion in Waade]. The central question is whether phi is as unobservable as the consciousness it claims to operationalize, or whether it correlates with certain patterns of behavior that are intuitively rated as conscious-like by human participants. The strength of this experimental set-up, then, is exactly that it gives the experimenter a known quantity on which to test such correlations, even if this is confined within the framework of IIT and therefore contingent on the validity of both its axiom and postulates and ultimately the mathematical formalisms (such as phi) it informs.

**2. Setting the theoretical stage for the experiment. IIT and a conceptual understanding of phi: What are we measuring anyways?**

In the following section, I will attempt to give a brief introduction to Integrated Information Theory as the theoretical framework and its concepts based on version 3.0 of IIT (Oizumi et al., 2014). Although an updated version 4.0 of the theoretical framework has recently been published (Albantakis et al., 2023), this presentation will be based on the previous version as the project it builds on the experimental work of Lundbak Olesen et al., 2023, itself using the framework from IIT 3.0.

IIT aims at a mathematical formalization of the experience of consciousness, or, as stated in Albantakis et al., 2023, an “*account for the phenomenal properties – the properties of experience – in physical terms.*” (Albantakis et al., 2023, p. 2). The theory builds on five phenomenal axioms about the nature of conscious experience, considered immediate and irrefutable truths. These axions are then extended into postulates about the mechanisms or system of mechanisms serving as the physical substrate of conscious experience. As such, IIT is a functional framework insofar as the requirements of the substrate are limited to functional requirements to the architecture of the system, but not necessarily its material implementation. Importantly, these postulates bring with them a set of formal concepts and measures that are used to describe the phenomenal experience of consciousness in both quantitative and qualitative terms.

That these concepts and measures are a reflection of the conscious experience rests on a claim of identity between “*phenomenal properties of experience and informational/causal structures of physical systems*” (Oizumi et al., 2014, p. 3), or, in my interpretation, an identity between the internal structure of the subjective experience of consciousness of a given conscious system and the objective measures proposed by the IIT framework. The concrete identity which is central to the theory will be further elaborated after a presentation of the fundamental axioms and postulates. Ultimately, the theory gives way to a set of mathematical measures and structures that are, in theory, reflective of the conscious experience and that lend themselves not only to theoretical interpretation of existing research but also to novel types of experiments, as demonstrated in the experimental paradigm based on simulations with small network called *animats*, on which it is possible to calculate some of the IIT measures (Albantakis et al., 2014). Indeed, it would seem that one of the strengths of IIT is that all of its concepts and these concepts’ consequences for the structure of both the physical substrate of consciousness and its experiential counterpart can theoretically (if not always practically) be tested experimentally.

In order to understand the actual measures of IIT it is instructive to first consider the axioms and postulates that informs them. In Oizumi et al., 2014, the axioms seem to use *consciousness* and *experience* more or less interchangeably, but in all cases I will here assume that the axioms are related to the *phenomenal experience of consciousness*. Briefly, the five axioms – here reformulated and/or elaborated based on the presentation in Oizumi et al., 2014 – are:

* **Existence:** Consciousness exists.
* **Composition:** Consciousness (or the conscious experience) is structured.
* **Information:** Each conscious experience is particular, meaning that it is distinct from every other possible experience. This also means that the experience is informative in that it contains something *different*, e.g., different from that which came before and that which comes after.
* **Integration:** Consciousness is integrated, meaning that it cannot be reduced to the sum of its parts, or to non-interdependent sub-systems or components. Insofar as this reduction is possible, such sub-systems or components will not be considered as direct contributors to the consciousness in question (as we will see later when dealing with the postulates).
* **Exclusion:** Consciousness is exclusive, meaning that an experience cannot be partial, and that consciousness cannot be composed of any number of partial experiences but a single definite experience. Included in this axiom are the notions of the temporal and spatial grains of a conscious experience; concretely, that the experience flows at a certain temporal speed and that it has a temporal-spatial threshold of sorts; some events are available to the experience, whereas other are too fine (e.g., too small or too brief) for the resolution of the particular conscious system in question.

Collectively,

Relevant sources: [1] Dawkins: multiple paths with different levels of consciousness can lead to the same behavior. [2] Gutfreund: Evolutionary selection works on external behavior, not internal states. Hence, one could be led to assume that consciousness will indeed have causal effects on behavior, as it would not have evolved otherwise. [3] Griffin: Conscious and nonconscious systems could have similar capabilities.3

1.2 How does the following project (and importantly the experiment) attempt to contribute to our understanding of the link between behavior and consciousness?

1.3 A specific note on IIT as an intrinsic theory of consciousness; within the framework of IIT specifically, why is it interesting to try to relate measures of internal (phenomenal) consciousness with externally observable behavior?

While this paper and the experiment it describes started as an investigation into the connection between consciousness and externally observable behavior, it rather became an inquiry into which factors humans value when rating a synthetic being as more or less conscious-like.

Additionally, there is the historically well-known but unsolved problem of determining what consciousness even is, with answers ranging from it being the only thing whose existence we can truly know (Descartes), to an almost banishment from psychology altogether.

1. Although see Griffin, 1976 for an interesting suggestion on how researchers could potentially learn to gain introspective insights from animals. [↑](#footnote-ref-1)
2. It is important to consider that consciousness could still be an evolutionary advantage without necessarily affecting external behavior. As Oizumi et al., 2014 points out, consciousness could be seen as an especially economic way of solving certain tasks for an organism within its environment. Although a non-conscious system could in principle be designed or evolve to maintain similar functions, a conscious system could arguably do so with a higher degree of flexibility and speed considering the resources required to compose such a system. [↑](#footnote-ref-2)