Research Article

Andrés Gómez-Emilsson*, Chris Percy

The "Slicing Problem" for Computational Theories of Consciousness

https://doi.org/10.1515/opphil-2022-0225
received August 3, 2022; accepted October 25, 2022

Abstract: The "Slicing Problem" is a thought experiment that raises questions for substrate-neutral computational theories of consciousness, including those that specify a certain causal structure for the computation like Integrated Information Theory. The thought experiment uses water-based logic gates to construct a computer in a way that permits cleanly slicing each gate and connection in half, creating two identical computers each instantiating the same computation. The slicing can be reversed and repeated via an on/off switch, without changing the amount of matter in the system. The question is what do different computational theories of consciousness believe is happening to the number and nature of individual conscious units as this switch is toggled. Under a token interpretation, there are now two discrete conscious entities; under a type interpretation, there may remain only one. Both interpretations lead to different implications depending on the adopted theoretical stance. Any route taken either allows mechanisms for "consciousness-multiplying exploits" or requires ambiguous boundaries between conscious entities, raising philosophical and ethical questions for theorists to consider. We discuss resolutions under different theories of consciousness for those unwilling to accept consciousness-multiplying exploits. In particular, we specify three features that may help promising physicalist theories to navigate such thought experiments.

Keywords: computationalism, causal structure, functionalism, physicalism, theories of consciousness, IIT, binding problem, Slicing Problem, philosophy of mind, thought experiment

1 Introduction

This article presents a new thought experiment, the "Slicing Problem," that targets substrate-neutral computational theories of consciousness.

Thought experiments have long been used to challenge, refine, and develop theories of consciousness,[1] remaining a live area of novel research as of 2022.[2] As theories of consciousness multiply with little consensus emerging,[3] diverse forms of enquiry may usefully be marshalled into this disputed area of philosophy, being increasingly urgent for practical, scientific, and political questions. Many ethicists and laypeople incorporate some notion of consciousness into topics that are highly charged from a political,

1 Praëm and Steglich-Petersen, "Philosophical Thought Experiments."
2 For example, Gidon et al., "Does Brain Activity Cause Consciousness? A Thought Experiment."
3 Seth and Bayne, "Theories of Consciousness."

* **Corresponding author: Andrés Gómez-Emilsson,** Qualia Research Institute, San Francisco, United States, e-mail: algekalipso@gmail.com
Chris Percy: College of Arts, Humanities and Education, University of Derby, Derby, United Kingdom
ORCID: Chris Percy 0000-0003-0574-9160

legal, or technical perspective, such as the treatment of animals[4] or the nature of machine sentience.[5] As these debates progress, they suffer from difficulties and disagreements around the definitions, components, and measurement techniques for consciousness, which are as closely intertwined with philosophical enquiry as they are with progress in the natural sciences.[6] Continued scrutiny of computational theories is particularly valuable given the recent focus in this area, such as investigations of working memory and other "computational correlates of consciousness"[7] and progress in computationally focused cognitive neuroscience.[8]

The Slicing Problem involves using water-based logic gates to construct a computer. This particular implementation, as demonstrated in Section 3, allows us to cleanly slice each gate and connection within the computer in half, creating two identical computers each instantiating the same computation. The slicing can be reversed and repeated via an on/off switch, without changing the amount of matter in the system. As the switch is toggled, theorists must consider whether the number of individual conscious entities in the system increases or stays the same and what differences might exist among the entities throughout. Depending on the preferred interpretation, we are led either to accept the existence of "consciousness-multiplying exploits" or to accept that boundaries between conscious entities are ambiguous and potentially arbitrary. Either interpretation may have counterintuitive implications for the philosophy of mind or for ethics based on the moral worth of individual, discrete conscious entities. The strength of the thought experiment lies in its full description of a theoretical mechanism by which any computational algorithm could be sliced, by explaining how pipes, universal logic gates, and clock signals can all be implemented and sliced using a single slicing mechanism, without requiring any complex disassembly or reassembly such as might open up space for shifts in a conscious experience that are not the explicit target of the thought experiment.

The structure of this article proceeds as follows: Section 2 describes the types of theories of consciousness that are most directly targeted by the Slicing Problem. Section 3 sets out the Slicing Problem itself, explaining how a Turing-complete water-based logic system can be constructed in a way that permits clean slicing into two computationally identical, causally separated systems without any increase in mass. Section 4 presents the different interpretation options available within a substrate-neutral computationalist paradigm and briefly explores the questions they raise for philosophy of mind and meta-ethics. Section 5 explains how alternative stances on consciousness might resolve the Slicing Problem without permitting consciousness-multiplying exploits, including the potential for non-computational theories and theories that specify a physical substrate within which information processing might occur. The conclusion reflects on how this thought experiment relates to other similar challenges to computationalism and summarizes the article.

2 Which theories are targeted by the Slicing Problem

Computationalism, or the computational theory of mind, describes a broad family of views that argue cognition and consciousness can arise as a result of computation,[9] noting that computation typically involves some form of structured information processing by which inputs or symbols are manipulated. The computationalist theories most closely targeted by this thought experiment can be isolated with respect to Marr's "three levels" framework for analyzing information processing systems, in that they specify at most the first two levels of the framework.[10] We draw on this framework as a narrative aid, acknowledging

4 For example, Burghardt, "Ethics and Animal Consciousness."
5 For example, Bostrom and Yudkowsky, "The Ethics of Artificial Intelligence."
6 Browning and Veit, "The Measurement Problem in Consciousness."
7 Reggia et al., "Modeling Working Memory."
8 For example, Zednik, "Computational Cognitive Neuroscience."
9 Rescorla, "The Computational Theory of Mind."
10 Marr, *Vision*, 25; Marr and Poggio, "From Understanding Computation;" Poggio, "The Levels of Understanding Framework."

that debates remain about the nature of information processing, the semantic content in Marr's approach[11] and connections to representational and computational theories.[12]

Within Marr's framework, the first level of information processing is the computational level, specifying what the system can achieve (such as multiplying two numbers together). The second level, algorithmic, specifies the representational schemes that explain the semantic referents of inputs and outputs and which algorithm(s) perform the computation (e.g., Karatsuba's algorithm for multiplication). The third level, implementation, specifies the physical structure(s) that realize the algorithm (e.g., a wooden computer powered by a mechanical hand-crank).

Some functionalist theories apply computationalism in a way that largely remains at the first level of Marr's framework. They describe what the computation achieves but remain neutral about which algorithms and which physical infrastructure implements it. Putnam's early thinking on machine functionalism draws heavily on this type of thinking, albeit Putnam later revised his views.[13] High-level descriptions of predictive processing frameworks also fit closely to this level where they prioritize what predictions occur and how they are used and updated, rather than the specific calculus that generates a prediction.[14] Global Workspace Theory and Higher-Order Thought Theory[15] also typically relate the consciousness-generating features to functions potentially implementable via different algorithms on different substrates, even if they may devote significant attention to specifying how a particular human brain substrate might implement them.[16]

Other computationalist theories of consciousness specify broad categories of the algorithm, potentially placing some restrictions at the second level of Marr's framework. Examples include enriched forms of Turing-style computation such as the representational theory of mind[17] and connectionists who favor algorithms modelled on neural networks.[18] Sometimes the algorithmic universality or Turing completeness is emphasized such that the category itself may be of low importance. Other computationalist theories may place far tighter restrictions on the permitted algorithms or specific algorithmic features required for computation to give rise to consciousness. Causal structure theories typically fall into this category and include Integrated Information Theory (IIT)[19] and Recurrent Processing Theory.[20]

In principle, a computational theorist might also specify the type of physical substrate a computation must take place in, reaching into the third level of Marr's framework. In practice, however, for most computationalists, substrate-neutrality is an attractive feature of their model. We explore the possibility of substrate-specific information processing in Section 5.

The Slicing Problem targets those theories in which state consciousness arises as a result of particular activities that can be fully described by the computational or algorithmic levels of an information processing system but are neutral regarding the implementation level. A physical substrate may be required but provided it has the necessary features to implement the algorithm then consciousness can arise; its other physical characteristics are irrelevant.

Where the algorithmic specifications in a theory cannot technically be modelled using AND, XOR, and NAND gates or analog equivalents, the Slicing Problem as set out in Section 3 would not apply formally, but variants of it potentially would. Similarly, most role functionalist approaches are in the scope of the thought experiment, although noncomputational functionalist theories can also be imagined and are not the target

---

11 For example, discussion in Richtie, "The Content of Marr's Information-Processing Framework."

12 For example, Shagrir, "Structural Representations and the Brain."

13 Levin, "Functionalism."

14 For example, Clark, "Whatever Next?"

15 Baars, *A Cognitive Theory of Consciousness*; Carruthers and Gennaro, "Higher-Order Theories of Consciousness."

16 Park and Tallon-Baudry, "The Neural Subjective Frame."

17 Fodor, *The Language of Thought*.

18 Marblestone et al., "Toward an Integration of Deep Learning and Neuroscience;" Kriegeskorte, "Deep Neural Networks."

19 Oizumi et al., "IIT."

20 Doerig et al., "The Unfolding Argument;" Lamme, "Towards a True Neural Stance on Consciousness."

of this thought experiment.[21] The thought experiment is also relevant to realizer functionalism, except where a theory articulates why a specific physical substrate is necessary for consciousness to arise. As such our target theories map approximately to those following Chalmers' description of computational sufficiency, which "stat[es] that the right kind of computational structure suffices for the possession of a mind, and for the possession of a wide variety of mental properties."[22]

For this article, it is helpful to provide illustrative basic definitions of the consciousness aspects and information-processing activities used in this context, while noting that specific computational theorists will often invoke more specific definitions in their own theories which would typically inherit the challenges raised by the Slicing Problem at a more abstract level of computationalism.

The aspect of consciousness in question is phenomenal consciousness, taken here as possessing a first-person perspective or self-awareness, some locus at which awareness can take place, however much it may vary in terms of its other functions, agency, experience range, or persistence. This first-person perspective is a simple prerequisite for most meaningful definitions of consciousness. In the terminology of Seth and Bayne's recent review, this is "subjective awareness,"[23] with "something that it feels like" for an entity to be conscious.[24]

A similarly basic definition of information processing can be applied that fits the computationalist paradigm.[25] Information processing refers to some input, typically a defined space inscribed with symbols that can take a restricted range of values, which can then be manipulated in a variety of pre-specified ways to produce an intermediary or final output. Such manipulation can also be broadly defined, and may draw on intermediary stochastic inputs, memory writing, and memory access capabilities, as well as formal or mathematical logic.

Under this basic definition, many activities can be thought of as processing information, although it is typically necessary to posit some internal or external observer that can interpret the output in a way that the process becomes meaningful, i.e., the output must "represent" something to someone.[26] Internal observers of the semantic value of any given information can be understood as arising in the context of whichever theory of consciousness a reader wishes to adopt in reviewing the thought experiment. Computationalist perspectives typically hold that the right computations done in the right way will result in phenomenal consciousness, i.e., something which is capable of being aware of the sematic content of the information being processed and/or its output – capable at least in principle and drawing on other resources or functions as required.

For instance, tree rings can be thought of as the result of an information processing activity, which, despite its many inputs and complex processes in an individual tree, results (even if incidentally) in simple and broadly consistent output information. This information processing may be meaningful to a sufficiently informed human who can interpret it to estimate the age of a tree but may have little meaning to other entities unaware of the representative import of rings even if their sense organs are capable of perceiving the same rings and their minds capable of conceiving a tree's growth over time. In both cases, the same information processing has taken place within the tree system, but only some observers are capable of interpreting it.

21 Piccinini, "Functionalism, Computationalism, and Mental States;" Piccinini, "Computation and the Function of Consciousness."

22 Chalmers, "A Computational Foundation for the Study of Cognition," 324.

23 Seth and Bayne, "Theories of Consciousness."

24 Nagel, "What is it Like to be a Bat."

25 For example, Shannon and Weaver, *The Mathematical Theory of Communication*.

26 Dretske, *Knowledge and the Flow of Information*; Grice, *Studies in the Ways of Words*.

3 The Slicing Problem for a computationalist mind

3.1 Building a water computer

In order to present the Slicing Problem, we will need a physical computer that can simulate a Turing machine[27] and that can be cleanly sliced into two identical, independent computing systems. Such a computer is not only theoretically plausible but also describable in terms of specific mechanisms.

Imagine a physical computer that uses water for computation.[28] Instead of wires that carry electricity like in a digital electronic computer, pipes would carry water in our computer. During computation, a stream of water would represent a "1" and an absence of water would represent a "0." Standard logic gates are constructed, each of which takes potential streams from two input pipes and outputs a single steam based on the values of the two inputs. For instance, an AND gate outputs a "1" if and only if both of its inputs are "1." Instead of logic gates composed of transistors, we use physical pipes and barriers. Figure 1 illustrates an AND gate implemented in a water system. If both input streams are switched on, the water flows collide, leading to water gathered in the cup around the output pipe, leading to a stream flowing out of the output, indicating a value of "1." Subsequent mechanisms can siphon off any excess water so that the output stream remains the same volume as the input streams. If only one input stream is switched on, the water flows past the output pipe cup and can simply exit the system via a drain pipe or be otherwise redundant for the mechanism. In this case, or if no water enters the system, no water exits the output pipe, indicating a value of "0."

In a second example, Figure 2 generates an XOR gate. In this case, the drain and output pipes are reversed. If both input pipes are on, the two water flows will collide, exit via the drain pipe and the output pipe will register a "0." If a single input pipe is on, the flow will pass the drain pipe cup and flow out of the output pipe, registering a "1," as required by the XOR gate schematic.

Having created AND and XOR gates, they can be combined to construct a NAND gate, which is a universal logic gate and can be combined with itself in large combinations to eventually compute any Boolean function.[29] A NAND gate ("NOT-AND") always outputs a stream of water unless both of its input streams are switched on. A NAND water gate can be constructed in the usual manner: first, pass the two inputs through an AND gate, then pass the resulting single output stream into an XOR gate alongside a permanently on stream: NAND(input1, input2) = XOR(1, AND(input1, input2)).

If a clock signal is additionally desired, we can use a giant swimming pool as a water source and attach a valve controlled by an electric motor that opens and closes at a fixed frequency to release water. If a purely water-based system is sought after, we could use a Pythagorean cup that is filled by a constant stream such that the cup empties at precise, regular intervals. There are many other, often mechanistically simpler ways for water mechanisms to replicate the function of example logic gate systems. However, for the illustration that a Turing-complete system can in principle be constructed (barring finite memory limitations), it suffices to have demonstrated the potential for repeatedly interacting NAND gates. In practice, a purely NAND-based water system would be extremely large for most calculations of interest and likely prone to malfunction. Significant error correction protocols would be needed, which can be similar in principle to those used in modern computing and telecommunications.

For simplicity, the presentation has focused on a digital-equivalent set-up, with discrete states and steps. However, theories of consciousness that draw on implementation-agnostic analog computations can also be targeted, as the setup can be transformed into an analog equivalent with a volume of flowing water used as an analog input, allowing for the construction of the usual archetypal analog computing mechanisms. Indeed, allowing for arbitrary scale and run-time complexity, hypothetical frictionless pipes, and boosting motors to add new water and energy where required for the functioning of the mechanism can

27 Turing, "On Computable Numbers."

28 Blikstein, *Programmable Water*.

29 Sheffer, "A Set of Five Independent Postulates for Boolean Algebras."

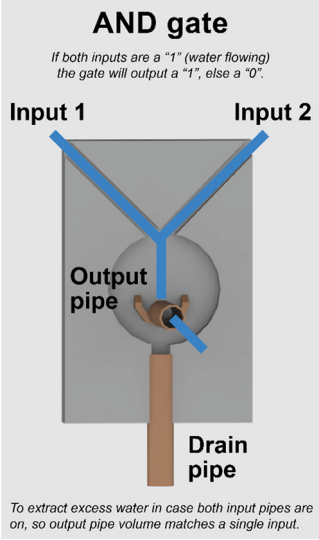

Figure 1: Example mechanism by which water stream inputs can serve as an AND logic gate.

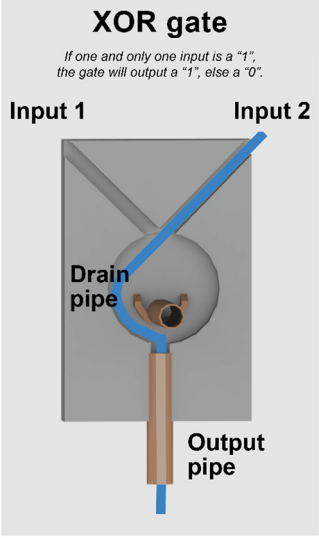

Figure 2: Example mechanism by which water stream inputs can serve as a XOR logic gate.

produce a Turing-complete water gate mechanism. Such a mechanism can in turn model the input–output mappings of artificial neural networks that give rise to modern artificial intelligence systems (e.g., the large language models[30]) or the input–output mappings of a quantum circuit, albeit not the latter's underlying mechanism, hence theories of consciousness that lean on specific physical mechanisms not necessarily being targeted by this thought experiment. We similarly note that stochastic computation can be addressed with the same sources of functional randomness as available in modern computers, such as far-out digits of pi being effectively random from the perspective of a user and assumed to be causally and statistically disconnected from whatever calculation is being conducted.[31]

30 Brown et al., *Language Models are Few-Shot Learners*.

31 Whether this is truly random raises a broader set of philosophical questions, as summarized by Eagle, "Chance versus Randomness."

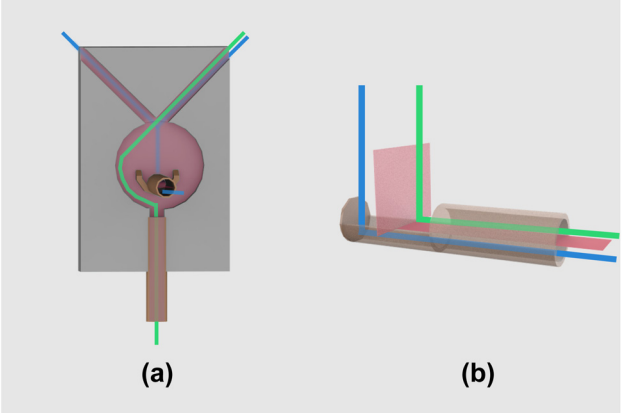

Figure 3: A visualization of how slicing would occur on an AND gate.

Any substrate-neutral computationalist theory of consciousness can in principle be modelled on the water computer system. Any specified algorithms for implementing a given computation can also generally be modelled, provided the algorithm specification is sufficiently abstract that it can be mapped to archetypal logic gates and memory systems, or otherwise be modelled in a water mechanism. In other words, by demonstrating that the Slicing Problem exists with a water computer, it exists in principle for any system claimed to be conscious in ways that can be isomorphically modelled via particular computational and algorithmic set-ups.

3.2 Slicing a water computer

Our entire water computer can be thought of as a physical system P. We can choose any computation C in a computational theory of consciousness that is associated with some conscious state S and have our water computer instantiate that computation. To a supporter of this computational theory, when P performs computation C, it possesses conscious state S. For IIT and some other theories, computation C may consist of many individual steps c_0, c_1, ... c_n, each with their own state of consciousness s_0, s_1, ... s_n and may require certain types of step (such as the recurrent, self-referencing steps that mark out a system as integrated into IIT), though this detail is unimportant for the rest of the argument.

We now provide a way to cleanly slice the computer in half. Down the middle of the swimming pool, and at the edge of each pipe and logic gate, we make a thin slit. In that slit, we place a thin, rigid sheet that plugs the pipes and logic gates (thus preventing the water from leaking from the computer). The sheet is such that when you push it into the computer, it divides the water into two independent streams (Figure 3). This sheet would be similar to a water control gate, except that rather than blocking the flow of water, the sheet cuts it in half along its direction of flow. We wire a system in place so that we can control the position of the sheet by pressing a button we call "slice/unslice." This way, we can simply push the "slice/unslice" button once and the sheet will descend into the flow, slicing the computer in half. If we press the button again, the sheet is pulled from the computer, which merges both halves into a whole stream again.

The water logic gates are redundant in the *Z*-direction – indeed, this is what makes Figures 1 and 2 easy to depict on two-dimensional paper. The redundancy creates space for a clean slice that preserves two copies of each logic gate's and each pipe's functionality. If we consider our reader's eye view of the water logic gate in Figure 1 as looking at the *XY*-plane, where the AND pipe is protruding out in the *Z*-axis, then the thin sheet would slice a water logic gate in half by stretching flat across the *XY*-plane and existing halfway into the volume (in the *Z*-direction) where the water flows.

What we are left with is a physical system P divided into two causally independent physical subsystems A and B, each performing computation C. Even though the entire water computer as a physical system has

functionally the same mass as before and even though P is functionally and computationally equivalent to what it was just a moment before, A and B are causally independent: we can add food coloring to one side of the computer and only that side will change color, leaving the water on the other side perfectly clear.

In Figure 3, a water logic AND gate is sliced with a thin red membrane; most slicing occurs flat in the *XY* plane, i.e., the same dimension as the paper this is written on. In the left-hand image (a), green water is entering on the front side of the slice (from the reader's perspective) through one input while blue water is entering on the back side of the slice through both inputs. The two flows of water are causally separated by the red-slicing membrane. The green water flows through its section of the drain pipe, leaving the output pipe at "0" (as intended, since only a single input flow is switched on). In the right-hand side of Figure 3(b), we show the output pipe in more detail for the different cases where both sets of green and blue input flows are fully switched on, i.e., both produce an output value of "1," to illustrate how a slicing membrane can partition the output pipe into two causally separated sections, one fed by water on the back side of the *XY* red membrane (the blue water) and one fed by the front side (the green water).

For the avoidance of doubt, the system partitioning as described here is distinct from the type of partitioning discussed by Tononi et al.[32] in an IIT context who note that: "in IIT, information must be integrated. This means that if partitioning a system makes no difference to it, there is no system to begin with." Their account refers to partitioning the mechanisms themselves, such that the functions on either side of the partition are different sets from each other. In contrast, the thin sheet that we are adding to our water computer is not separating a subset of mechanisms from another but creating two causally independent systems that have the exact same causal structure.

4 Different computational interpretations of slicing

What happens once the computer is sliced? From a computational perspective, we see two top-level possibilities, differentiated by a type-token distinction as common elsewhere in philosophy.[33]

Under the type interpretation, if one consciousness entity corresponds to one unique type of computation taking place, then the post-sliced system has a single conscious entity. While each side of the slice has causally separate structures, they each have identical inputs, processing activities, and outputs, corresponding to an identical type of information processing, and hence a single conscious entity. Under the token interpretation, each side of the slice is a different instance or token of the same computation, and each token has its own consciousness. Both interpretations have implications that need to be accounted for by the theorist. Bostrom discusses the feasibility of each alternative in a computational scenario,[34] ultimately endorsing what we call here a token interpretation while pointing out others who favor a type of interpretation.[35]

4.1 Implications for the token interpretation (consciousness doubles)

Under the token interpretation, slicing the water computer gives rise to two independent computers, each with its own instance of computation C and accordingly, each with its own instance of conscious state S. If we adopt this view, we could then take either half and slice it again. In turn, we can iteratively slice the computer over and over, and in this way, multiply consciousness almost arbitrarily (up to the

32 Tononi et al., "Integrated Information Theory," 457.
33 Wetzel, "Types and Tokens."
34 Bostrom, "Quantity of Experience."
35 Citing, for instance, Zuboff, "One Self."

miniaturization limit imposed by the physical practicalities of moving liquid through the system). When we press the "slice/unslice" button again, we collapse multiple conscious states back into a single conscious state.

This implication may be counterintuitive to some theorists because it suggests that bringing an experience with rich and complex phenomenology into existence can in some circumstances require nothing more conceptually complicated than inserting a thin sheet of material in a physical system, not diminishing the engineering achievement that such a system would represent.

We call this a *consciousness-multiplying exploit*: an operation that increases the amount of identical conscious entities in a physical system without any material change in mass of the overall system, relying only on a conceptually simple manipulation within the existing system. Once set up, no complex disassembly or restructuring is required to implement the consciousness-multiplying mechanism – simply a clean slice through all of its parts. Of course, the overall system must be set up with all the additional mass and functionality of the slicing mechanisms in advance. As such, the system is larger than a simpler version of the system that conducts the same computation without the slicing functionality, but from a computationalist perspective, the key issue is typically what computations a system can deliver, not whether it does so in an "efficient" manner.

As an example, we can consider IIT, a theory of consciousness that specifies a causal structure but typically not a physical substrate.[36] In the language of IIT, the water logic gates in the Slicing Problem can constitute mechanisms and hence may give rise to a nonzero amount of consciousness (e.g., 2D grid of XOR gates or an expander graph, as well as more complex structures).[37] These mechanisms can then be sliced to give rise to two causally separated entities, each with the same consciousness volume (Φ) as the original. This would lead naturally to a token interpretation of the Slicing Problem, given that it is individual sets of causal connections themselves that are important for instantiating conscious entities with a certain volume of Φ.[38]

If a computational theorist would prefer that such consciousness-multiplying exploits should not be possible, at least not in the form described here, then they must adopt the type interpretation.

4.2 Implications for the type interpretation (consciousness is unchanged)

By using a different interpretation, one can solve the appearance of an entirely new experience by arguing that both sides of the computer are really *the same computer and as such correspond to a singular locus of identity having the same experience*. If we compare system A and system B by their information content, both computers are indistinguishable: they are doing the same operations at the same time (or, in an IIT sense, have the same causal structure and are capable of the same operations) – and as such ultimately contain the exact same information. With this view, particular experiences are identified with their corresponding computation, in contrast to the previous view in which experiences are identified with the individual instantiations of the computation. Such a position cannot lean on the spatial proximity of each half of the water computer alone, as once the two physical subsystems A and B are causally separate, additional mechanisms can be imagined to separate them arbitrarily far in space while maintaining them as remaining computationally and algorithmically identical to each other.

For type-theorists of computational consciousness, further implications arise when we consider the ongoing evolution of these new causally independent computational systems. Different target theories might lead to different implications, but we suggest each theory must adopt at least one of three positions, effectively corresponding to a consciousness equivalent of the Sorites paradox considering the relationship

36 Oizumi et al., "IIT."

37 Tononi, *Why Scott*.

38 Lombardi and López, "What Does 'Information' Mean in Integrated Information Theory?," 12.

between additional grains of sand and the appearance of a "heap of sand".[39] Each available option leads to the same phenomenon of consciousness-multiplying exploits or to boundary ambiguities that are hard to reconcile with a bounded, unitary phenomenology of consciousness,[40] noting that such unity can be complex – for instance, it may involve both background structure and foreground content connected via an attentional mechanism.[41]

The first option is that any infinitesimal change in one of the systems would create a new consciousness, given that we now have two non-identical computations C and C′, which give rise to two corresponding conscious states S and S′, necessitating two discrete conscious entities to experience each one. However, such an infinitesimal change is itself a consciousness-multiplying exploit as defined above: a conceptually minor, even trivially reversible step that creates a new entity with a potentially rich phenomenology. As with token interpretation, it is straightforward to slice the computer many times and perturb each of the slices in a slightly different way and get many different conscious experiences as a result.

The second option is that as the differences between system A and system B increase over time, they slowly become different computations and therefore different experiences, with no formal step change at any point. In such a case, system A and system B would start out producing the same single conscious state S. Then, as B is modified, B would add a corresponding fraction of consciousness to the whole system. However, this position must account for both questions of symmetry and the link between experience fractions and entity discreteness. Regarding symmetry, if B is changed with respect to A, either can be described as a slightly different version of the other. Is it A or B contributing the additional amount of consciousness to the system as a whole? More profoundly, what does a "fraction of an experience" mean with respect to a discrete number of conscious entities?[42] If our water computer is implementing a human mind and we slightly change the areas corresponding to visual experience, would the modification produce additional visual qualia experienced by the same unitary entity as before? As the modifications compound, does the original entity experience some gradual decoupling experience after which there is eventually a second separate entity, but during which there is some nebulous co-existence? How can this be reconciled with the normal, waking experience of consciousness as both unitary and bounded?

The third option is that the new consciousness appears after some threshold of differences. However, in this case, the consciousness-multiplying exploits still exist. They exist just before and just after the change that pushes the system past the specific threshold, as opposed to just before and after the slicing button is toggled.

4.3 Why these implications might matter

Some theorists may take issue with the discontinuities in consciousness implied by most of the Slicing Problem interpretations. For instance, Chalmers remarks in regard to a different thought experiment:

> On the second hypothesis, the replacement of a single neuron could be responsible for the vanishing of an entire field of conscious experience. If so, we could switch back and forth between a neuron and its silicon replacement, with a field of experience blinking in and out of existence on demand. This seems antecedently implausible, if not entirely bizarre. If Suddenly Disappearing Qualia were possible, there would be brute discontinuities in the laws of nature unlike those we find anywhere else.[43]

Computationalists may also find some readers concerned at the violation of a conservation principle: some intuition that there should be a connection (however complex) between the volume of total phenomenal

39 For example, see discussion in Dolev, "Why Induction Is No Cure For Baldness."
40 Bayne, *The Unity of Consciousness*.
41 Wiese, "Attentional Structure and Phenomenal Unity."
42 Note that Bostrom, "Quantity of Experience," argues in favour of fractional qualia – see discussion in the Appendix.
43 Chalmers, "Absent Qualia," 315.

experiences and the volume of corresponding physical substrates, rather than the marginal toggling of expanded experience seemingly possible by the slicing mechanism.

However, as with most thought experiments, it is possible to bite the bullet. Theorists may accept that consciousness-multiplying exploits exist even without any material change in the mass of the system, giving rise to sharp discontinuities in nature, and/or that the apparent discreteness of unitary consciousness experience does not exist. We provide brief reflections here on the potential ethical relevance of the former, given the available literature on the importance or otherwise of unitary experience.[44]

More generally, we acknowledge that positions on ethics vary widely and our intention here is not to argue that computational theorists who accept these implications have an irreconcilable ethical dilemma; rather we suggest they have a philosophical duty to respond to it. They may do so in a range of ways, whether accepting the ethical implications directly or adopting/modifying ethical theories which do not give rise to the issue (e.g., by not relating ethics to the experiences of discrete conscious entities or by specifying unitary consciousness as necessary but not sufficient for moral value). They might also deny the thought experiment, perhaps by rejecting its computationalist premise or objecting to how it proceeds to identify the implications. In any case, reflections on the thought experiment help to elucidate a position in one area of philosophy and the constraints that might arise for positions in other domains to avoid internal contradiction.

Some ethical accounts discuss consciousness as a necessary prerequisite for being a moral patient,[45] discussing the moral status of algorithms and information processes.[46] More generally, Bostrom discusses the ethical implications that follow depending on whether conscious entities are multiple or singular with respect to running identical computations.[47] If conscious entities of moral importance can be arbitrarily multiplied or reduced in number, especially without any diminishing in the volume of their conscious activity (however that might be defined), then one possible moral duty might be not to reduce their number, leaning on moral injunctions against murder. A second moral duty might be not to increase the number of entities if the original computations were structured such that the conscious entity was suffering from negative experiences.

Real world parallels to the rapid and reversible consciousness-multiplying exploits described above are currently hard to imagine, but an approximate, less efficient equivalent of the reverse act is somewhat easier to contemplate. Placing someone under general anesthetic or in a medically induced coma – or perhaps indeed when we go into deep sleep – is a typically reversible process that is easy to describe in terms of "button-toggling" terminology, even if its physical mechanisms might only be poorly understood today and might ultimately prove very complex. Some people's ethics may struggle to see such medical or natural circadian acts as having any negative moral value at all, even if in some cases that negative value can be argued outweighed by a future positive value, such as a life-saving operation conducted under general anesthetic or sleep enabling someone's future functioning.

Under the type interpretation, there is a further ethical implication that some theorists may need to consider, depending on their preferred ethical stance. Imagine three people Alice, Bob, and Carol, who are all modelling very similar, but slightly different computational states, and as such collectively represent three distinct conscious entities. If all three are suffering, it might be ethically attractive (under some ethical paradigms) to nudge their computational states to be identical. As a result, the three physically separate systems now give rise to only a single conscious state. There is now one discrete "locus of awareness" that is suffering rather than three. Depending on the utilitarian calculus used to assess the options, it may even be better to increase the suffering of one of the people initially only suffering mildly so that their computational model becomes identical to the others, collapsing their consciousnesses into one. Such meta-ethical implications could be explored further in future work.

---

44 For example, Bayne, *The Unity of Consciousness*; Bayne and Chalmers, "'What is the Unity of Consciousness?."

45 Muehlhauser, "Report on Consciousness and Moral Patienthood."

46 Tomasik, *Do Artificial Reinforcement-Learning Agents Matter Morally*.

47 Bostrom, "Quantity of Experience."

5 Alternative accounts of consciousness

A different response to this thought experiment is to deny the premise: the Slicing Problem is not in fact possible because the types of computational theories of mind it addresses do not adequately account for the conditions under which a first-person perspective can arise. In this section, we focus on physicalist theories that acknowledge a role for computation, at least in some broad form of information processing, but require it to be implemented in a certain way to give rise to consciousness, which may relate to physical constraints in implementation, e.g., a particular physical mechanism or substrate, or relate to spatio-temporal specifications, e.g., a required intensity of information processing.

For completeness, we first briefly consider alternative ways of denying the premise of the Slicing Problem. For instance, a theory may deny that computation leads to discrete conscious entities, no matter what form that computation might take place, whether continuous or discrete, whether stochastic or determined, or so on. One example draws on an open individualism interpretation of the question of identity.[48] Such an interpretation could assert there is only one conscious entity in the universe; if we feel ourselves to be bounded and discrete, separate from other entities or the universe as a whole, it is simply a reflection of impoverished perceptual filters or processing that disguises some greater truth. We note in passing that empty individualism would not be an adequate response to the Slicing Problem. While empty individualism denies the persistence of a single conscious entity over time, the Slicing Problem focuses instead on the number and nature of conscious entities that might exist in an individual moment in time. Even in empty individualism, we might desire a theory that can explain whether a particular pattern, existing just for a flash, gives rise to two first-person perspectives or one, even if these perspectives are illusory in the sense they would disappear and never return once the moment passes or the pattern shifts infinitesimally.

A third approach might allow computation and its physical realization to play a role in consciousness, perhaps even a necessary role, but in either case, not a sufficient role.[49] Something more would be needed beyond that computation and its physical realization. We might need some other substance not part of the computation's realization (such as a soul) or some functionality that is either unrelated to computation or cannot be reduced to information processing as defined here. This third category is hard to explore further in this article at present, as it often results from abductive reasoning that at least some mental phenomena elude the physical, the computational, or the functional, but without yet being able to specify in detail what theory might address them, often falling back on a residual, dwindling category of the unexplained, or even the forever inexplicable.

5.1 Physicalist theories for addressing the Slicing Problem

By definition, the Slicing Problem applies only to theories of consciousness that rely at most on the first two levels of Marr's framework for information processing systems: the computational and algorithmic layers. The third layer is the implementation layer, the physical substrate and physical mechanisms used to implement the algorithm. There is a wide and growing range of theories of consciousness that specify a physical implementation to different levels of detail.[50]

In a trivial sense, almost any physicalist theory could avoid the Slicing Problem as currently formulated, by asserting that they do not use water gates. Similarly, any theory that axiomatically requires the human brain or a similar structure to instantiate consciousness would avoid the problem. However, an

48 Kolak, *I Am You*.

49 For example, discussions in Fodor, *LOT2*.

50 For instance, McFadden, "CEMI Field Theory;" Pockett, "Consciousness is a Thing;" Barrett, "An integration of IIT;" Hameroff and Penrose, "Orch OR;" Pearce, *Physicalism*; and Tegmark, "Consciousness as a State of Matter."

axiomatic rejection of a thought experiment is analytically uninteresting, so we suggest a more fruitful line of enquiry is exploring the criteria a physicalist theory would want to satisfy to prevent the consciousness-multiplying exploits arising under a token interpretation of the Slicing Problem, assuming that some slicing mechanism could in principle be applied to whatever physical (non-water) system is instantiating consciousness in a particular theory. In other words, we want to describe the features of a system where some form of physical slicing – potentially highly involved and distributed over several mechanisms – would be incapable, even in principle, of creating two causally separate entities with the same conscious experience as the original.

As a starting point for future enquiry, we suggest physicalist theories that can account for three key features that might respond to the Slicing Problem challenge with reduced concern for consciousness-multiplying exploits. In the first instance, we suggest that at least one category of physicalist theory appears promising in this respect: those that identify consciousness with the energy and configuration of a region of a physical field, potentially a highly complex combination of fields with certain persistent spatio-temporal patterns.

The first feature is a physically grounded way to determine how conscious a system is. We need to define what about a physical structure generates consciousness, and ideally provide some mechanistically plausible or at least consistent account of how it does so.

Field theories might identify particular physical properties such as the volume of energy or electrical charge in a field to be the sorts of things that count toward consciousness, with tentative insights pointing to the importance of local field potentials in neural activity giving rise to consciousness.[51] In this sense, any slicing of the system – even assuming it could be physically implemented – would necessarily result in smaller systems that have individually less total energy and hence lower volumes of consciousness than the original whole. Whatever survives a slicing mechanism cannot be described as two duplicates of the original – you may well end up with two discrete loci of awareness, but total consciousness has been preserved or at least not increased in some way, such that there are less severe discontinuities than in the Slicing Problem via a substrate-neutral computationalist lens.

The second feature the physical theory must account for is a physically grounded way to resolve the binding problem, i.e., why awareness appears to have a broadly unitary and bounded phenomenology.[52] There is an eventual edge to our awareness, even if it is a fading nebulous edge like the boundaries of a cloud. Within that awareness, there is typically a sense of a single experience that unifies many individual instances of sense perception and information processing. We do not normally sense each "pixel" of the world individually, but rather we see a unified visual field.

Fields in physics are unified by default, so the challenge collapses to answering how fields might develop relevant boundaries. Complex and persistent field patterns might naturally lose consistency as they encounter neighboring, disordered fields from the ambient environment, permitting the identification of a nebulous boundary. It may also be possible to specify more precise boundaries, pursuing the topological features of fields in physics more generally,[53] which allow the boundary around an experience to be frame-invariant.[54]

With respect to the slicing problem, as a result of how they resolve the binding problem, it is possible that complex fields theorized to give rise to consciousness simply do not have a redundant dimension along which they can be sliced.[55] This would mean any two entities that follow a slicing exercise cannot be running the same original computations in two discrete entities. Even if a redundant dimension or some complex slicing mechanism can be identified for some field-based conscious entities, the first feature ensures that the two resulting functionally identical computations are lesser in some way than the original:

---

51 Pockett, *The Nature of Consciousness*; McFadden, "CEMI Field Theory."

52 Chalmers, "The Combination Problem for Panpsychism."

53 Gross and Kotiuga, *Electromagnetic Theory and Computation*.

54 Gómez Emilsson, *Solving the Phenomenal Binding Problem*.

55 See, e.g., Lehar, *Harmonic Resonance Theory*, with respect to Chladni figures.

perhaps they operate at a slower speed, higher error rate, or lower level of detail; perhaps they give rise to reduced intensity of qualia. In either case, conscious-multiplying exploits are much reduced.

The first two features are necessary requirements for a theory of consciousness, being a subset of the eight subproblems of consciousness.[56] The third feature is not a formal requirement (since consciousness may be purely epiphenomenal), but would increase confidence in the account for those adopting a nonepiphenomenal perspective. The third feature asks the theory to articulate a reason why consciousness as described in the theory is used by evolved organisms. In other words, why does its model of consciousness provide some benefit or relevance in a functional sense for the goals that the human system has been incentivized by evolutionary processes to achieve? In some cases, this benefit is likely to relate to functions that can be represented in computational terms, even if the human brain instantiates them in some noncomputational manner. For field theories of consciousness, it is possible that the holistic behavior of some fields can be recruited for field computing tasks as discussed by MacLennan,[57] perhaps enabling the speeding up of certain analog computing tasks, but significant further investigation is required into this topic.

6 Conclusion

The central intuition explored in the Slicing Problem is how the same computing system might reversibly give rise to singular or multiple conscious entities with identical phenomenal experiences, without any material change in their mass. The idea that equally rich phenomenal experiences might be so easily manipulated into and out of existence can run counter to physical intuitions about discontinuities in nature, as described by Chalmers,[58] or some conservation principle connecting the volume of total phenomenal experiences and the volume of corresponding physical substrates, even if the translation between the two may be far from linear.

The Appendix contrasts the Slicing Problem with other thought experiments identifying potentially counter-intuitive implications of computationalist-type approaches by referencing different intuitions or different mechanisms, including famous work by Searle, Block, Parfit, and Bostrom. Such thought experiments all draw on related, non-identical intuitions and mechanisms, but collectively represent an important challenge to different subsets of computational and functional theories. We suggest that such exercises are a valuable tool for refining theories of consciousness.

The value of a thought experiment corpus is less in exploring the intuitive reaction to one particular thought experiment and more in testing the consistency of a given theory and its interpretation across a range of different thought experiments. Indeed, while intuition and introspection remain useful – perhaps unavoidable – starting points and reference points for global assessments of a theory, the poor track record of individual instances of intuition in the history of philosophical and scientific enquiry urge caution at relying on any individual intuitive reaction. For instance, geocentrism and vitalism were arguably intuitive in the absence of evidence granted by modern scientific instruments and were accepted principles for long periods of time in certain mainstream communities.

Alongside internal consistency across different thought experiments, we agree with Seth and Bayne[59] that an additional valuable tool is an external consistency with observed features of systems we normally treat as archetypes of consciousness, notably ourselves. Examples include the effect of anesthetic on human consciousness, the clinical phenomenon of blindsight, and the apparent unity of consciousness we experience and can reliably disrupt. Other common tools are helpful but perhaps less rigorous, such as looking to self-report or observation of reactions from potentially conscious entities, since the more such

56 Johnson, *Principia Qualia*.
57 MacLennan, "Field Computation."
58 Chalmers, "Absent Qualia."
59 Seth and Bayne, "Theories of Consciousness."

entities differ structurally from humans the harder it becomes to give credence to self-report or observation, such that the method may become a scientific at the limit.[60]

To summarize, this article has presented the "Slicing Problem" thought experiment, by which a Turing-complete mechanism constructed out of water-based logic gates can potentially give rise to consciousness-multiplying exploits under computational theories of consciousness. We suggest that such consciousness-multiplying exploits lead to ethical questions and philosophical questions about the nature of consciousness that may help computational theorists reflect on their theories and interconnected philosophical positions.

This article ends by exploring how alternative conceptions of consciousness may address the Slicing Problem without giving rise to such blunt conscious-multiplying exploits. For instance, physical theories might avoid the dilemma by specifying a particular physical substrate or physical mechanism that is necessary for consciousness. We hope that a future research program will investigate the potential of physical theories, in particular field theories, to meet three features that help to address the Slicing Problem: a physically grounded mechanism that gives rise to a volume of consciousness, a physically grounded mechanism that satisfies the binding problem, and an account for why evolution might have co-opted those mechanisms into the modern human organism and potentially a much broader range of organisms.

Appendix: Contrast with other thought experiments

This appendix sets out the distinctions between the Slicing Problem thought experiment and other thought experiments that similarly explore potentially counter-intuitive implications of computationalist-type approaches by referencing different intuitions or different mechanisms.

For instance, Searle's Chinese Room[61] thought experiment challenges our intuitions about true understanding compared to mechanistic input–output functions, however, complex those functions might be. The account asks where the conscious understanding would exist (if indeed it can exist) in a closed room containing a non-Chinese speaker communicating with external Chinese speakers via complicated symbol manipulation rule books to generate output Chinese text in response to incoming Chinese text, similar in principle to today's large language models.

Block's China Nation or China Brain thought experiment[62] challenges intuition via a different mechanism, i.e., the intuition that consciousness needs to be situated in some system that appears closed and well-contained to us. Would a mind be produced by a vast nation of people passing messages to each other that functionally reflect the neuronal computations taking place in a brain? If so, what kind of a mind is this and what sort of space can it be considered to exist in?

Block also attacks functionalism with a thought experiment of two functionally identical people having inverted color qualia – if functionalism cannot account for a difference and yet one person sees red as green, this difference in qualia must exist outside functionalism.

In contrast to these three thought experiments, the Slicing Problem challenges possible intuitions that it should not be trivial to manipulate equally rich phenomenal experiences into and out of existence or that there should be some conservation principle connecting the volume of total phenomenal experiences and the volume of corresponding physical substrates.

The fractional qualia that emerge from Bostrom's "unreliable computer" account provide an alternative potential challenge to computationalist thinking.[63] He envisages a multitude of separate probabilistic (or "imperfectly reliable") mechanisms that potentially direct signals within copper wires to duplicate components in an overall computational system, thus emphasizing a different principle from this article which

60 For example, as discussed but rejected by Goff, *Galileo's Error*.
61 Searle, "Minds, Brains, and Programs."
62 Block, "Troubles with Functionalism."
63 Bostrom, "Quantity of Experience."

relies on the singular and unitary nature of the slicing mechanism and an initial nonduplicate structure. While the Slicing Problem leads to the discussion of consciousness-multiplying exploits, Bostrom's account leads to his conclusion that "subjects of experience," or discrete entities having a first person perspective, can be fractional. In other words, an experience can be qualitatively identical but experienced by 0.3 of a subject in the same way that two people experiencing the same pain is qualitatively identical but quantitatively larger than one person experiencing that pain. In the 2006 paper, Bostrom is willing to accept this conclusion, but others may take his discussion as a reason against the computationalist concept of consciousness that allows for his set-up and the resulting fractional entities. Physicalist theories that meet the requirements we set out in Section 5 have the potential to mitigate concerns about both consciousness-multiplying exploits and fractional qualia.

Several thought experiments separately challenge the persistence of personal identity over time, ranging from the classic Ship of Theseus-type accounts[64] to the teletransportation of Parfit.[65] If teletransportation results in two individuals identical at the atomic level, both the original at the origin and the copy at the destination, where does the original personal identity reside? Perhaps at the origin. Were the original to be destroyed in the act of copying, would the identity have moved to the destination? If so, what mechanism can account for how a first-person perspective transfers spatially in the latter case but not the former? Similar intuitions can be tested via Parfit's fusion thought experiment in which two persons' brains are connected together such that they work as one – whose identity is retained, if anyone's?

The Slicing Problem focuses on the number of entities that exist within a particular physical structure at a moment in time as it is manipulated and is distinct from the notion of identity persistence. The duplication of a human body-brain doubles the amount of mass in the system, so there is no intuitive concern if two consciousnesses emerge. The merging of two initially *different* human minds in the fusion experiment is likewise different from the duplication and dissolution of multiple *identical* minds in a single physical system. Likewise, Parfit's "divided minds" thought experiment suggests how two functionally different and causally separate entities might be created out of one physical system – this difference may well result in each experience being smaller in some way than the original, given the smaller amount of mass and mechanisms each has access to.

Importantly, the demonstration of a physical mechanism by which a water computer could implement the experiment removes a range of mechanistic objections that might be applied to the fusion experiment or related personal identity thought experiments. Such mechanistic objections potentially identify space for changes in identity or conscious experience during more complex processes of disassembly, connection, or reassembly. More broadly, the personal identity thought experiments can often be resolved by appealing to the notion that persistent identity is entirely illusory, as in empty individualism,[66] but there still remains a question of how many discrete entities exist in a particular pattern even if just for an infinitesimal moment in time. An eliminative perspective can of course resolve both simultaneously: if there is no such thing as a discrete first-person perspective (despite apparent introspective evidence to the contrary), then many thought experiments in the philosophy of mind are simply moot.

Another thought experiment line of attack on computationalism has focused on the possibility that inputs and outputs can be arbitrarily extracted from any reasonably sized physical system and arbitrarily mapped to different meaningful representations, such that the system could be interpreted by different external observers as implementing very many different computations, raising the question of "which one (s)" are giving rise to a discrete conscious entity.[67] Multiple realizability is another type of critique arguing that at least some mental experiences (such as "pain") cannot be identical to any single physical structure, since the experience can be realized on multiple different structures, whether different human brains or different animal brains.[68] One example attempt at resolution is to present the word "pain" as a convenient

64 Fowler, *Plato*.
65 Parfit, *Reasons and Persons*.
66 See Kolak, *I Am You*, for a description and rejection of this position.
67 Eliasmith, "The Myth of the Turing Machine."
68 Polger and Shapiro, *The Multiple Realization Book*.

umbrella category, like "cloud," with many unique instances each corresponding to a unique physical structure, just as there may be many different individual clouds but there is still communicative utility in an umbrella term for the category.

Acknowledgments: The authors thank Andrew Zuckerman for discussion, research, and intellectual support in the early drafts of the article. Jeremy Hadfield for brainstorming discussions and participating in the early process of considering paper topics. Michael Johnson for the valuable work done in *Principia Qualia* identifying promising physicalist theories of consciousness and his arguments against functionalism, upon which this article builds. David Pearce for many helpful discussions about the causal relevance of phenomenal binding. Michael D. Smith and David Brooks for digital correspondence about water computers. Two anonymous reviewers and the editorial team for their constructive criticism and support throughout.

Funding information: The authors state no funding was received for this research.

Author contributions: The authors applied the SDC approach for the sequence of authors, in order of declining importance of contribution. A.G. came up with the idea of the slicing problem. A.G. and C.P. worked together to specify the implementation detail of the problem. C.P. reviewed the literature to ascertain the originality of the argument, relationship to other thought experiments, and positioning with respect to existing theories of consciousness. A.G. identified the different interpretations within a computationalist paradigm after the computer is sliced. C.P. identified the broader implications and alternative resolutions for the thought experiment. All authors contributed to writing the manuscript and testing and refining the arguments.

Conflict of interest: The authors state no conflict of interest.

Ethical approval: The conducted research is not related to either human or animal use.

Data availability statement: Data sharing is not applicable to this article as no datasets were generated or analyzed during the current study.

References

Baars, Bernard J. *A Cognitive Theory of Consciousness*. Cambridge, UK: Cambridge University Press, 1993.

Barrett, Adam B. "An Integration of Integrated Information Theory with Fundamental Physics." *Frontiers in Psychology* 5 (2014), 63. doi: 10.3389/fpsyg.2014.00063.

Barrett, Adam B. and Pedro A. Mediano. "The Phi Measure of Integrated Information is Not Well-defined for General Physical Systems." *Journal of Consciousness Studies* 26:1–2 (2019), 11–20.

Bayne, Tim and David Chalmers. "What is the Unity of Consciousness?." *The Unity of Consciousness: Binding, Integration, and Dissociation*, edited by Axel Cleeremans, 23–58. Oxford, UK: Oxford University Press, 2003.

Bayne, Tim. *The Unity of Consciousness*. UK: Oxford University Press, 2010.

Blikstein, Paulo. *Programmable Water*. Paulo Blikstein. (n.d.) http://alumni.media.mit.edu/~paulo/courses/howmake/mlfabfinalproject-old.htm.

Block, Ned. "Troubles with Functionalism." *Minnesota Studies in the Philosophy of Science* 9 (1980), 261–325.

Bostrom, Nick. "Quantity of Experience: Brain-Duplication and Degrees of Consciousness." *Minds & Machines* 16 (2006), 185–200. doi: 10.1007/s11023-006-9036-0.

Bostrom, Nick and Eliezer Yudkowsky. "The Ethics of Artificial Intelligence." In *Artificial Intelligence Safety and Security*, edited by Roman Yampolskiy, 57–69. Boca Raton, Fla., US: Chapman and Hall/CRC, 2018.

Brown, Tom B., Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared Kaplan, Prafulla Dhariwal, A. Neelakantan, et al. *Language Models are Few-Shot Learners*, 2020. arXiv preprint arXiv:2005.14165.

Browning, Heather and Walter Veit. "The Measurement Problem in Consciousness." *Philosophy Topics* 48 (2020), 85–108.

Burghardt, Gordon. "Ethics and Animal Consciousness: How Rubber the Ethical Ruler?" *Journal of Social Issues* 65 (2009), 499–521. doi: 10.1111/j.1540-4560.2009.01611.x.

Carruthers, Peter and Rocco Gennaro. "Higher-Order Theories of Consciousness." In *The Stanford Encyclopedia of Philosophy* (*Fall 2020 Edition*), (2020), edited by Edward N. Zalta. https://plato.stanford.edu/archives/fall2020/entries/consciousness-higher/.

Chalmers, David J. "Absent Qualia, Fading Qualia, Dancing Qualia." In *Conscious Experience*, edited by Thomas Metzinger, 309–28. Leiden, NL: Ferdinand Schoningh, 1995.

Chalmers, David J. "A Computational Foundation for the Study of Cognition." *Journal of Cognitive Science* 12:4 (2011), 325–359. doi: 10.17791/jcs.2011.12.4.325.

Chalmers, David J. "The Combination Problem for Panpsychism." In *Panpsychism*, edited by Godehard Bruntrup, and Ludwig Jaskolla, 179–214. Oxford University Press, 2016. doi: 10.1093/acprof:oso/9780199359943.003.0008.

Clark, Andy. "Whatever Next? Predictive Brains, Situated Agents, and the Future of Cognitive Science." *Behavioral and Brain Sciences* 36:3 (2013), 181–204. doi: 10.1017/s0140525x12000477.

Doerig, Adrien, Aaron Schurger, Kathryn Hess, and Michael H. Herzog. "The Unfolding Argument: Why IIT and Other Causal Structure Theories Cannot Explain Consciousness." *Consciousness and Cognition* 72 (2019), 49–59. doi: 10.1016/j.concog.2019.04.002.

Dolev, Yuval. "Why Induction Is No Cure For Baldness." *Philosophical Investigations* 27 (2004), 328–44. doi: 10.1111/j.1467-9205.2004.t01-1-00230.x.

Dretske, Fred. *Knowledge and the Flow of Information*. Oxford: Blackwell, 1981.

Eagle, Antony. "Chance versus Randomness." *The Stanford Encyclopedia of Philosophy* (*Spring 2021 Edition*), (2021), edited by Edward N. Zalta. https://plato.stanford.edu/archives/spr2021/entries/chance-randomness.

Eliasmith, Chris. "The Myth of the Turing Machine: The Failings of Functionalism and Related Theses." *Journal of Experimental & Theoretical Artificial Intelligence* 14:1 (2002), 1–8. doi: 10.1080/09528130210153514.

Fodor, Jerry A. *The Language of Thought*. Cambridge, Mass.: Harvard University Press, 1975.

Fodor, Jerry. *LOT2*. Oxford: Clarendon Press, 2008.

Fowler, Harold. *Plato in Twelve Volumes*, Vol. 9, *translated by* Harold N. Fowler. Cambridge, MA, Harvard University Press; London, William Heinemann Ltd., 1925. http://data.perseus.org/citations/urn:cts:greekLit:tlg0059.tlg009.perseus-eng1:139.

Gidon, Albert, Jaan Aru, and Matthew Larkum. "Does Brain Activity Cause Consciousness? A Thought Experiment." *PLoS Biol* 20:6 (2022), e3001651. doi: 10.1371/journal.pbio.3001651.

Goff, Philip. *Galileo's Error: Foundations for a New Science of Consciousness*. London, UK: Rider, 2019.

Gómez Emilsson, Andrés. *Solving the Phenomenal Binding Problem: Topological Segmentation as the Correct Explanation Space* [Video]. *YouTube* (2021) https://www.youtube.com/watch? v = g0YID6XV-PQ.

Grice, Paul. *Studies in the Ways of Words*, Cambridge: Harvard University Press, 1989.

Gross, Paul and P. Robert Kotiuga. *Electromagnetic Theory and Computation: A Topological Approach*. Cambridge: Cambridge University Press, 2004. doi: 10.1017/CBO9780511756337.

Hameroff, Stuart and Roger Penrose. "Consciousness in the Universe: A Review of the 'Orch OR' Theory." *Physics of Life Reviews* 11:1 (2014), 39–78. doi: 10.1016/j.plrev.2013.08.002.

Johnson, Michael E. *Principia Qualia*. 2016. https://opentheory.net/PrincipiaQualia.pdf.

Kolak, Daniel. *I Am You: The Metaphysical Foundations for Global Ethics*. Berlin, DE: Springer Science & Business Media, 2007.

Kriegesgorte, Nikolaus. "Deep Neural Networks: A New Framework for Modeling Biological Vision and Brain Information Processing." *Annual Review of Vision Science* 1 (2015), 417–46.

Lamme, Victor A. "Towards a True Neural Stance on Consciousness." *Trends in Cognitive Sciences* 10:11 (2006), 494–501. doi: 10.1016/j.tics.2006.09.001.

Lehar, Steven. *Harmonic Resonance Theory: An Alternative to the "Neuron Doctrine" Paradigm of Neurocomputation to Address Gestalt properties of perception*. 1999. http://slehar.com/wwwRel/webstuff/hr1/hr1.html.

Levin, Janet. "Functionalism." In *The Stanford Encyclopedia of Philosophy* (*Fall 2018 Edition*), (2018), edited by Edward N. Zalta. https://plato.stanford.edu/archives/fall2018/entries/functionalism/.

Lombardi, Olimpia and Cristian López. "What Does 'Information' Mean in Integrated Information Theory?" *Entropy* (*Basel, Switzerland*) 20:12 (2018), 894. doi: 10.3390/e20120894.

MacLennan, Bruce J. "Field Computation in Natural and Artificial Intelligence." *Information Sciences* 119:1–2 (1999), 73–89. doi: 10.1016/S0020-0255(99)00053-5.

Marblestone, Adam, Greg Wayne, and Konrad Kording. "Toward an Integration of Deep Learning and Neuroscience." *Frontiers in Computational Neuroscience* 10 (2016), 1–41.

Marr, David and Tomaso Poggio. "From Understanding Computation to Understanding Neural Circuitry." *A.I. Memo* 357 (1976), 1–22.

Marr, David. *Vision: A Computational Investigation into the Human Representation and Processing of Visual Information*. W. H. Freeman and Company, 1982.

McFadden, Johnjoe. "The CEMI Field Theory Closing the Loop." *Journal of Consciousness Studies* 20:1–2 (2013), 153–68.

Muehlhauser, Luke. "Report on Consciousness and Moral Patienthood." *Open Philanthropy Project* 357 (2017), 62–86.

Nagel, Thomas. "What Is It Like To Be a Bat?", *Philosophical Review* 83 (1974), 435–450.

Oizumi, Masafumi, Larissa Albantakis, and Giulio Tononi. "From the Phenomenology to the Mechanisms of Consciousness: Integrated Information Theory 3.0." PLoS Comput Biol 10(5): e1003588 (2014). https://doi.org/10.1371/journal.pcbi.1003588.

Parfit, Derek. *Reasons and Persons*. Oxford [Oxfordshire]: Clarendon Press, 1984.

Park, Hyeong Dong and Catherine Tallon-Baudry. "The Neural Subjective Frame: From Bodily Signals to Perceptual Consciousness." *Philosophical Transactions of the Royal Society B: Biological Sciences* 369 (2014), 20130208.

Pearce, David. *Non-Materialist Physicalism: An experimentally testable conjecture*. 2014. https://www.physicalism.com/.

Piccinini, Gualtiero. "Functionalism, Computationalism, and Mental States." *Studies in History and Philosophy of Science Part A* 35:4 (2004), 811–33. doi: 10.1016/j.shpsa.2004.02.003.

Piccinini, Gualtiero. "Computation and the Function of Consciousness." In *Neurocognitive Mechanisms: Explaining Biological Cognition*. Oxford, UK: Oxford University Press, 2020. https://oxford.universitypressscholarship.com/view/10.1093/oso/9780198866282.001.0001/oso-9780198866282-chapter-15.

Pockett, Susan. *The Nature of Consciousness: A Hypothesis*. Writers Club Press, 2000.

Pockett, Susan. "Consciousness is a Thing, Not a Process." *Applied Sciences* 7:12, (2017), 1248. doi: 10.3390/app7121248.

Poggio, Tomaso. "The Levels of Understanding Framework, Revised." *Perception* 41:9 (2012), 1017–23.

Polger, Thomas and Lawrence Shapiro. *The Multiple Realization Book*. New York: Oxford University Press, 2016.

Praëm, Sara Kier and Asbjørn Steglich-Petersen. "Philosophical Thought Experiments as Heuristics for Theory Discovery." *Synthese* 192 (2015), 2827–42. doi: 10.1007/s11229-015-0684-6.

Reggia, James A., Garrett E. Katz and Gregory P. Davis. "Modeling Working Memory to Identify Computational Correlates of Consciousness" *Open Philosophy* 2:1 (2019), 252–69. doi: 10.1515/opphil-2019-0022.

Rescorla, Michael. "The Computational Theory of Mind." *The Stanford Encyclopedia of Philosophy* (*Fall 2020 Edition*), (2020), edited by Edward N. Zalta. https://plato.stanford.edu/archives/fall2020/entries/computational-mind/.

Ritchie, J. Brendan. "The Content of Marr's Information-Processing Framework." *Philosophical Psychology* 32:7 (2019), 1078–99. doi: 10.1080/09515089.2019.1646418.

Searle, John R. "Minds, Brains, and Programs." *Behavioral and Brain Sciences* 3:3 (1980), 417–57. doi: 10.1017/S0140525X00005756.

Seth, Anil K. and Tim Bayne. "Theories of Consciousness." *Nature Neuroscience Neuroscience* 23 (2022), 439–52. doi: 10.1038/s41583-022-00587-4.

Shagrir, Oron. "Structural Representations and the Brain." *British Journal for the Philosophy of Science* 63:3 (2012), 519–45.

Shannon, Claude and Warren Weaver. *The Mathematical Theory of Communication*. Urbana, IL: University of Illinois Press, 1949.

Sheffer, Henry M. "A Set of Five Independent Postulates for Boolean Algebras, with Application to Logical Constants." *Transactions of the American mathematical society* 14:4 (1913), 481–8. doi: 10.1090/S0002-9947-1913-1500960-1.

Tegmark, Max. "Consciousness as a State of Matter." *Chaos, Solitons & Fractals* 76 (2015), 238–70. doi: 10.1016/j.chaos.2015.03.014.

Tegmark, Max. "Improved Measures of Integrated Information." *PLOS Computational Biology* 12:11 (2016), e1005123. doi: 10.1371/journal.pcbi.1005123.

Tomasik, Brian. *Do Artificial Reinforcement-Learning Agents Matter Morally*? 2014. [Preprint]. http://arxiv.org/abs/1410.8233 [Accessed April 30, 2021].

Tononi, Giulio. *Why Scott Should Stare at a Blank Wall and Reconsider* (*or, the Conscious Grid*). 2014. http://integratedinformationtheory.org/download/conscious_grid.pdf.

Tononi, Giulio and Christof Koch. "Consciousness: Here, There and Everywhere?" *Philosophical Transactions of the Royal Society B: Biological Sciences* 370:20140167 (2015). doi: 10.1098/rstb.2014.0167.

Tononi, Giulio, Mélanie Boly, Marcello Massimini, and Christof Koch. "Integrated Information Theory: From Consciousness to its Physical Substrate." *Nature Reviews Neuroscience* 17:7 (2016), 450–61. doi: 10.1038/nrn.2016.44.

Turing, Alan M. "On Computable Numbers, with an Application to the Entscheidungsproblem." *Proceedings of the London Mathematical Society* 2:1 (1937), 230–65. doi: 10.1112/plms/s2-42.1.230.

Wiese, Wanja. "Attentional Structure and Phenomenal Unity." *Open Philosophy* 5:1 (2022), 254–64. doi: 10.1515/opphil-2022-0197.

Wetzel, Linda. "Types and Tokens." In *The Stanford Encyclopedia of Philosophy* (*Fall 2018 Edition*), (2018), edited by Edward N. Zalta. https://plato.stanford.edu/archives/fall2018/entries/types-tokens/.

Zednik, Carlos. "Computational Cognitive Neuroscience." In *The Routledge Handbook of the Computational Mind*, edited by Sprevak, Mark and Colombo, Matteo, 357–69. New York: Routledge, 2019.

Zuboff, Arnold. "One Self: The Logic of Experience." *Inquiry*, 33 (1991), 39–68.