Are algorithms always arbitrary? Three types of arbitrariness and ways to overcome the computationalist's trilemma

Abstract

Implementing an algorithm on part of our causally-interconnected physical environment requires three choices that are typically considered arbitrary, i.e. no single option is innately privileged without invoking an external observer perspective. First, how to delineate one set of local causal relationships from the environment. Second, within this delineation, which inputs and outputs to designate for attention. Third, what meaning to assign to particular states of the designated inputs and outputs. Having explained these types of arbitrariness, we assess their relevance for algorithms from various computational theories of mind (CTM) seeking to account for phenomenal consciousness. Some CTM accounts can survive some of these arbitrariness challenges, but eventually each one examined faces a trilemma. Each is under pressure to dilute at least one of three important CTM desiderata: its causal relevance, its substrate neutrality, or its assignment discrimination (e.g. ability to avoid near-panpsychism). However, unlike previous research, we suggest the debate cannot be fully settled at the level of principle. Certain complex CTM algorithms may be salvageable, albeit not the algorithms specified to date. The paper specifies three areas of investigative work, as well as further theoretical avenues of research, which computationalists may wish to explore to escape the trilemma.

Key words: Computationalism; Consciousness; Computation Individuation; Boundary Problem

1. Introduction

One common way to think of algorithms is that they fully describe how a given abstractly-defined state is transformed over a finite sequence of discrete steps.[1] When considering any given physical structure as a candidate for implementing a particular algorithm, the available transformations are fixed by that mechanism's physical and causal structure, but there remain three arbitrary choices to define the algorithm:

- ***Meaning assignment***: The meaning or referent assigned to input/output states.
- ***Mechanism designation***: The features designated as inputs/outputs in a structure.
- ***Environment delineation***: The boundary between a structure and its environment.

This paper provides definitions and worked examples to explain these three types of arbitrariness, illustrating how the same physical structure can result in different algorithms being implemented.

An extensive literature has addressed the topic of algorithmic arbitrariness directly or indirectly under different names – such as triviality arguments, pancomputationalism, computation individuation/indeterminacy, and the boundary problem – and from different perspectives, such as computer science, cognitive science, thermodynamics, evolutionary biology, and philosophy of mind (citations throughout). The literature has not, however, contrasted these types of arbitrariness explicitly, which limits its traction on the topic of this paper: whether algorithms from computational theories of mind (CTM) can account for phenomenal consciousness.[2]

A computational functionalist perspective on phenomenal consciousness is often described as a "mainstream" view (e.g. Butlin et al., 2023:4), notwithstanding a long history of diverse arguments and counter-arguments (see overview in Rescorla, 2020). Such a claim is indirectly supported by occasional surveys of philosophers on related questions, such as functionalism being the most popular account of consciousness in the 2020 PhilPapers survey (Bourget & Chalmers, 2023; computationalism was not an option) and 67% of philosophers reporting they believed machines could become conscious (Francken et al., 2022).

CTM accounts are often attractive because they are claimed to provide at least the following three benefits:

1. ***Causal relevance***. Algorithms transform inputs into outputs, allowing new insights and actions to be initiated on the basis of information processing about external sensory data. The experience and utility of such information processing are regular

[1] This informal definition of algorithms and associated definitions of information follow Hill (2016) and Knuth (1973), but abstract the teleology given the topic focus on consciousness. While adequate for this purpose, we acknowledge definitional disagreements, e.g. Vardi (2012), and the implementation requirement for an adequately precise language in line with Primiero's (2020) notion of implementable abstract machines.

[2] Defined here as referring to having a first person perspective, such that we can say there is some subject of experience present (following Nagel, 1974), however minimal that experience or subject might be (Metzinger, 2020). Note that phenomenal consciousness is distinct from the likely role of algorithms and information processing in mental cognition, which might form part of what we experience phenomenally. Nonetheless, we do cite arguments about arbitrariness as applied to mental cognition, where they can be applied also to phenomenal consciousness.

features of human experience.
(e.g. attractive to those hoping to explain why evolution might have had reason to select for consciousness, e.g. Georgiev, 2024)

2. ***Substrate-neutrality***. Any physical substrate suffices for executing the CTM algorithm and generating consciousness, provided it meets minimum mechanical standards to implement the causal structure (e.g. Miłkowski, 2016). Given the ability to construct universal Turing Machines out of very simple mechanisms, such as NAND logic gates, simple physical systems can often be compounded to implement any Turing-simulable algorithm (e.g. Woods & Neary, 2007; Fages et al., 2017; Lachmann & Sella, 1995).
(e.g. attractive to those concerned about biological chauvinism or hoping to upload their mind to a computer one day, e.g. Sandberg & Bostrom, 2008)

3. ***Assignment discrimination***. CTM provides a seemingly clear motivation for assigning consciousness to a modest subset of the physical structures that exist, notably whether they execute the relevant algorithm (also called extensional adequacy in Sprevak, 2018). In particular, it avoids panpsychism or near-panpsychism, defined as implying a great plurality of additional adjacent conscious entities for each entity a theorist might wish to designate conscious.
(e.g. attractive to those who want animals and machines to be capable in principle of consciousness, but concerned about theories allowing rivers, cells, cities, or solar systems to be conscious, e.g. Rosenberg, 2004)

This paper demonstrates that the arbitrariness arguments pressure CTM accounts to abandon or significantly dilute at least one of these benefits: the "trilemma" of this paper. While some CTM accounts restrict the types of algorithms in scope to escape some types of arbitrariness, no current account is identified that escapes all of them. Many major positions that can be interpreted from a CTM perspective (including but not limited to those in Butlin et al., 2023) have questions to answer: representational, higher order, recurrent processing, global workspace, predictive processing, embodiment, and integrated information theories (IIT).

For all three types of arbitrariness, this paper points to specific theoretical and investigative work that could be done to strengthen CTM accounts, helping to inform the future research agenda. In this respect, we disagree with previous literature that either declares CTM as a whole class of theories rendered impossible (or at least extremely unlikely) given different arbitrariness critiques (e.g. Eliasmith, 2002; Piper, 2012; Shagrir, 2012) or declares CTM victorious given definitional restrictions on computation (e.g. Rescorla, 2014, 2020; Miłkowski, 2017).

One powerful meta-solution to these arbitrarinesses in the computation individuation literature is to introduce observer-dependency, some external bestower of semantics (Hemmo & Shenker, 2022b[3]). Schweitzer (2019), for instance, is happy to concede that computation is

[3] Note that their solution to prevent infinite regress applies only to computation and not to the CTM accounts in scope for this paper. They explicitly state that their mind-brain identity theory "does not identify the mind with a

observer-dependent in his discussion of cognition capabilities. Sprevak (2018) refers to the diversity of useful perspectives in scientific methods to argue for a similar pluralism in accounts of computation, which works where scientific methods have us as external observers. For mental cognition or for biological functions more generally (e.g. Bongard & Levin, 2023), this strategy can work by leaning on the first person perspective that brains generate to be that bestower of semantics, bracketing off how exactly it does so.

The meta-solution, however, fails for phenomenal consciousness, since such a subjective perspective is the very phenomenon we wish to account for, leading to an infinite regress. For instance, Shagrir (2020) acknowledges this in his defence of semantic computation: "Laptops operate on semantic properties that are defined, at least partly, by the user of the machine. Presumably, the content of the computations that take place in our brain is not defined by the interpretation of an external observer." Shagrir is spared from providing a solution, since his paper's target is not phenomenal consciousness.

Observer-dependency is also an unattractive position because it suggests I can change the fact of your consciousness based on my external perspective: not just my epistemological uncertainty about your consciousness but the actual ontological status of your consciousness as you are experiencing it. In the Appendix, we consider three ways to salvage observer-dependency, concluding each requires significant further work to gel with CTM.

Section 2 provides additional definitions for the paper's scope. Section 3 provides examples to explain the three types of arbitrariness. Section 4 provides examples of how potential CTM accounts might address the trilemma for each type of arbitrariness.

2. Scope definition

This paper focuses on CTM accounts within a set of metaphysical assumptions, assumed for the purposes of discussing CTM rather than defended in this paper.

First, algorithms have to be executed in a physical substrate to have any causal effect - an algorithm existing as a concept or a blueprint is a description only. Causality is approximately interpreted in a counterfactual sense (e.g. Pearl, 2000). Algorithms reliably transform inputs to outputs, such that a counterfactual change in the inputs leads reliably to the appropriate counterfactual change in outputs. Algorithmic steps can be specified at a low level (e.g. machine code, see Stallings, 2015) or a high level (e.g. more complex but still unambiguous transformations such as "produce the prime factors of X"), provided there exists a clear language for translating the latter into the former, potentially multiple forms of the former that achieve the same function. Some CTM accounts might require examination at the base level of the causal structure (e.g. IIT CTM), whereas others might be content with higher level specifications (e.g. input/output functionalist CTM accounts). All levels are in scope.

Second, the physical substrate is not itself conscious; consciousness emerges instead from the relevant algorithms being executed on it (e.g. Rescorla, 2020). Such a position relates

computation, but rather with some particular set of *macrovariables of type b* of the brain (and perhaps body). These macrovariables are yet to be discovered by brain science and cognitive science."

naturally but not exclusively to physicalist perspectives, particularly flat physicalism (Hemmo & Shenker, 2022a). These and any other definitions are contested and subject to edge cases, but working definitions are sufficient for this paper, being supported by the worked examples of computations (e.g. for overviews see Frisch, 2023; Stoljar, 2024).

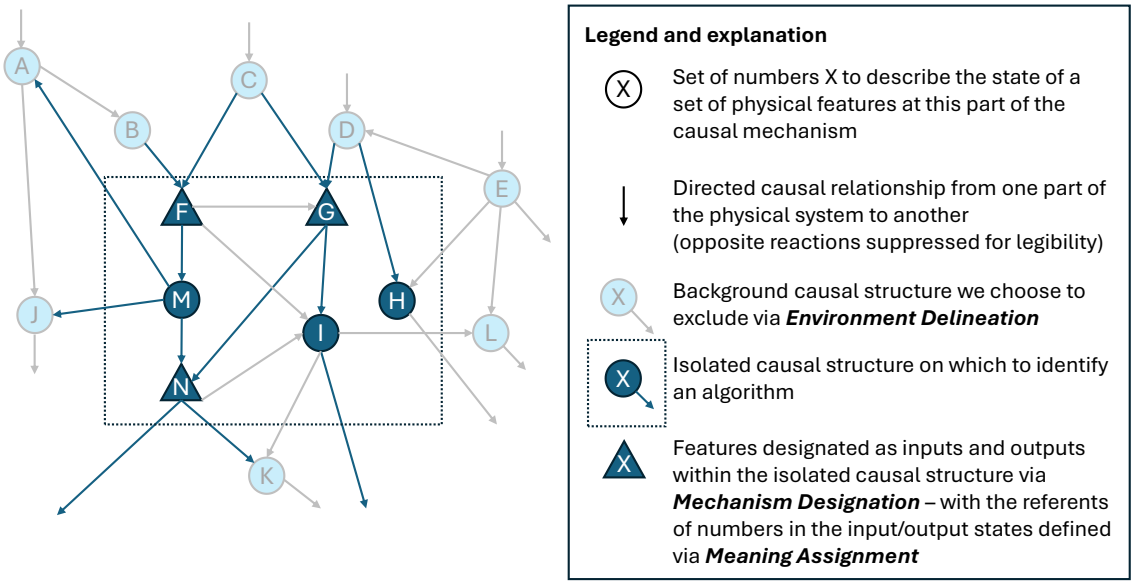
There are also two scope restrictions on the types of CTM account that are the primary target of this paper. First, the description of the algorithm and the physical structures it supervenes on must be finite. This does not exclude loops or an indefinite run time, provided in such cases consciousness is assumed to be present during the loop or execution run time (rather than "emerging" only when the algorithm completes). Second, the algorithm of interest can in principle be simulated on a Turing Machine working in discrete and distinguishable steps, recognising that in practice physical substrates often present more efficient alternatives than the classic machine's head, tape, state register, and instruction table.

The potential to relax these two scope restrictions (finite; Turing-simulable) and two metaphysical assumptions (execution-requirement; physicalism) is discussed in the Appendix. None of the relaxations provide much succour to CTM: either we dilute one of the trilemma benefits or introduce the interaction problem so challenging to dualism (e.g. Ludwig, 2003; Westphal, 2016). In any case, such assumptions are implicit or explicit in standard CTM accounts and retain in scope algorithms executed on modern digital computers, including the main CTM accounts in Butlin et al. (2023).

3. Three types of arbitrariness

Three relevant types of arbitrariness exist for any given physical structure implementing a particular in-scope algorithm: meaning assignment, mechanism designation, and environment delineation. Figure 1 illustrates how a specific algorithm with two sets of inputs (on nodes F and G) and one set of outputs (node N) is implemented on a structure carved out from an underlying interconnected causal substrate, including additional transformations taking place via one hidden node (M).

Figure 1. Arbitrariness illustration in a physical system modelled as a causal graph



3.1. Environment delineation arbitrariness

We start by applying a version of Rosenberg's boundary problem argument, as foreshadowed by Schrödinger (1951), to physically-supervenient algorithm implementation (Rosenberg 2004, see also Gómez-Emilsson & Percy, 2023).

A first-order environment delineation is, in practice, typically spatio-temporal. For instance, we might talk about a laptop based within particular spatial boundaries for the next 60 minutes, while I run algorithms on it. However, the algorithms of interest supervene on causal mechanisms in physical structures rather than space-time coordinates, so spatio-temporal delineation is an initial convenience, not the end of the story.

Algorithms are "location-neutral", much as they are "substrate-neutral" – and in both cases only within certain scopes. Algorithms need to be in a physical space to be executed, but many locations would suffice: moving the laptop to another room does not in general stop it from working (but moving it to the centre of the sun would).

Likewise, I could be running the algorithms of interest on a water computer, a transistor-based computer, or many other structures (Gómez-Emilsson & Percy, 2022). It needs to be implemented in a physical mechanism with the necessary accessible and manipulable causal structure, but otherwise any mechanism will do. It doesn't matter how the necessary input gets encoded in the system, provided it can get in, and it doesn't matter exactly how the specified transformation happens, provided it does, etc. Of course, some transformations might be specified more precisely in one algorithm than others, in which case they must be implemented in a way that respects that precision.

Since algorithms supervene on causal mechanisms, we need to use causal structures to specify their boundaries rather than space-time coordinates. In practice, this involves isolating a set of substrate modalities of interest, potentially also specifying some sensitivity thresholds. Both of these acts are arbitrary in an observer-dependent manner, as argued below, rather than innate in physics (potential ways to view them as "innate" are discussed in §4.3).

3.1.1. The physical world is richly causally interconnected

The physical world is causally interconnected via different mechanisms, whether electromagnetic, gravitational, mechanical, or others. With effort, we can create partially and temporarily causally-isolated systems, but the vast majority of physical systems available for implementing algorithms are causally integrated with their environment. Those of interest to us as users are by definition causally integrated with us (since we manipulate and interpret them). We are in turn richly and continuously integrated with our environment lest we perish (oxygen, temperature, nutrition, etc).

More generally, at least some causal interactions extend past any practical boundary we might impose. We can mostly block external mechanical interactions by suspending a system in a vacuum. With increasing difficulty, we might block external electromagnetic field interactions with suitable casings. Attempts to block the gravitational influence of external

objects on geodesics influencing our isolated system will need to wait for future technology. Notwithstanding current or future human technologies, if there is a set of algorithms executed in the human brain which constitute the first person perspective in our minds, they cannot be considered to operate in a causally-isolated system.

3.1.2. Algorithm implementation relies on selecting a subset of causal interactions

In practice, we implement algorithms not by creating a wholly causally-isolated physical system that implements that algorithm but by choosing to focus on only some causal interactions of interest, designing a mechanism around that subset of causal interactions, intervening selectively on the subset of causally-available inputs, and reading selectively from the subset of available outputs. The system must be sufficiently isolated with respect to the target causal interactions such that only the user-manipulated inputs enter the system with sufficient force to affect the outputs we care about – other causal inputs must be safely ignorable within certain sensitivity bounds.

For instance, a Grandfather Clock focuses on mechanical causal interactions within a certain spatiotemporal area, but ignores electromagnetic interactions from nearby objects that are nonetheless taking place all around and within it. The location of the hands is not perfectly determined by the clockwork. For instance, it is slightly influenced by those electromagnetic interactions, as well as mechanical interactions from dust particles in the air, etc. Nonetheless, when wound up, the hands are reliably close enough to the relevant numbers that we can read off the desired answer. Likewise, a digital computer focuses on electrical interactions between logic gates, minimising unwanted causal interactions from electromagnetic field effects below a sensitivity threshold via shielding and other technologies. Quantum computers hope to focus on certain quantum mechanical interactions, but ignore gravitational interactions.

The incredible advances of recent centuries have involved designing such mechanisms that are ever more complex, while still enabling input manipulability and output sensitivity relative to a human user.

This rich and ever-extending network of causal interactions means that delineating any one physical system from its environment is an arbitrary act. No matter what point you identify as an input, you could have chosen some prior input that causally influences it. No matter what point you identify as an output, you could identify some subsequent output that it influences instead. This is the “environment delineation” source of arbitrariness. Some choices of delineation may feel obvious to a given user, but they are not fundamental in physics. They are “observer-dependent”, with no intrinsic motivation for privileging one delineation that enables one set of algorithms over another that enables a different set of algorithms.

3.1.3. Illustrative example: Laptop-Robot

One natural choice for a laptop’s environment delineation is the keyboard as an input to my laptop’s algorithm and the monitor as its output. We cannot isolate a single physical law or substrate modality as several are already involved, e.g. the mechanics of my hands typing the keyboard and the photons ejected from the monitor. However, imagine an automatic robot

whose robotic eyes can read the output of the monitor and conduct some further action – perhaps the monitor displays a maths question, which the robot reads as an image and translates into mathematical operations to conduct internally, before speaking out the answer. Suppose further that the human user entering the maths question to the laptop is blind. Even if the answer were also displayed on the first monitor, the user only hears the answer spoken by the robot.

From the user's perspective, the algorithm of relevance is no longer the original laptop (with keyboard + monitor) but the whole system of laptop + robot, since it is the spoken word of the robot that completes the algorithmic circuit for the user. The fact that the algorithm works through different substrate modalities at different points (electronic, visual, mechanical, auditory) is irrelevant. After all, a primary benefit of algorithms is substrate-neutrality.

Some might attempt to dissolve this example by being strict about different substrate modalities, perhaps asserting there are strictly two different algorithms taking place here (one in the laptop, one in the robot) rather than a single integrated algorithm. Unfortunately, the laptop itself betrays the attempt since it likely already uses different substrate modalities internally to answer the maths question. Perhaps the laptop has a hard disc drive using magnetism or fibre optic components using light pulses, in addition to the ubiquitous electrical interactions in transistors. We discuss the potential for more robust versions of this strategy to save certain CTM accounts from environment delineation arbitrariness in §4.3.

3.2. Mechanism designation arbitrariness

Let us assume we have found an acceptable environment delineation. We isolated the relevant substrate modalities to a level of sensitivity adequate for the purpose. This lines up to certain spatial coordinates, but those coordinates are defined by the causal structures.

We now have a particular physical structure on which to implement an algorithm, which has a boundary between the "inside" and the "outside". Inputs come from outside and outputs go to the outside, but we will not worry further about the inside/outside delineation. Inputs are the only place where the environment exerts a relevant causal influence over the system (relevant in being on the isolated substrate and above any applicable sensitivity thresholds).

A syntactical or structural type of arbitrariness must now be addressed - mechanism designation - described also in other terms in Eliasmith (2002) who explains that any given physical system can be described in terms of a near infinity of different virtual machines or Turing machines.

Even having defined a boundary, any algorithm complex enough to be of interest for a non-panpsychic CTM will have more than one input and more than one output. Provided outputs are causally downstream of inputs, we can make many choices to define what the algorithm is doing. Notably, which option from the power set of candidate input variables/nodes and which option from the power set of candidate output variables/nodes should be designated of interest.

For instance, if there are three input nodes which can each take certain values and subsequently interact via a causal mechanism, we can choose to pay attention to all three, any one, or any combination of them. The power set of subsets over three inputs provides 8 options, although the empty set can arguably be ignored as it amounts to ignoring the mechanism and going about your business elsewhere.

A candidate solution discussed further in §4.2 is simply to examine all inputs and outputs simultaneously. Two issues from a practical perspective are discussed here, leaving the philosophical implications until later. First, there may be impractically many candidate inputs (the options may not even be disjoint for some mechanisms). Second, the algorithm you want might only be implemented once you isolate a subset of inputs or outputs.

As an example for the first case, imagine we are implementing an algorithm exploiting the Brownian motion of gas in a given space. One set of input candidates might be the density of gas molecules initialised in three spatially demarcated zones on one edge of the space. But those three zones could be spatially demarcated in many different ways, by carving up the space differently up to the instruments' sensitivity. Many attempts to map to a physical mechanism result in such an explosion of options. For instance, the behaviour of electricity conducting through a complex metal may produce manipulable inputs and accessible outputs in terms of variables such as voltage, temperature, electromagnetic field topology, and many other macrostates, as well as an incredibly large number of microstates. Indeed, since we can work off state macrovariables, there are typically infinitely many options for each given state microvariable, as argued by Hemmo & Shenker (2019).

In the second case, let us continue with the Grandfather Clock example. We restrict the outputs to the hands and clockface, perhaps excluding the broader potential output contact surface via a sensitivity threshold for the human visual system. The standard algorithm implemented is dividing a half day into 12 equally spaced segments, either by looking at when the minute hand points directly up or when the hour hand moves along number increments (or the algorithm counts to 12 over a fixed period of time, if you prefer). However, there is another reading, arguably more natural given there are two hands. Each time the two hands overlap exactly divides the half day into 11 equally spaced segments. Yet, if you ask someone with access to the clock to take medical pills 11 times evenly spaced, they may struggle to calculate this or at least not hit on this easy approach (if not previously aware of it and not mathematically inclined). In that sense the availability or not of an algorithm in a system depends on the knowledge and decisions of an observer. Different ways of looking at the clockface can produce many more algorithms beyond these two simple time division algorithms.

If a more computationally-familiar example is preferred, consider the "xor-and-adder" mechanism detailed in Table 1. Via an assumed environmental delineation, we have a physical system sufficiently causally isolated to have two inputs we can manipulate and two outputs we can read, with inner causal structures to implement the logic in Table 1. Nonetheless, we can still choose which inputs and outputs to use. The non-empty power set of output nodes has three options to examine, each of which implements an algorithm with a

different causal structure: (i) using only output A implements a XOR logic gate; (ii) using only output B implements an AND gate, and (iii) using both implements a two 1-bit adder with carry (where Output A is the least significant bit and B is the carry).

Table 1. A logical table mapping to a two 1-bit adder with carry

Input A	**Input B**	**Output A**	**Output B**
0	0	0	0
0	1	1	0
1	0	1	0
1	1	0	1

Options also exist via the inputs. For instance, if we hold B fixed at 1 and manipulate only A, then Output A provides a "flip" mechanism and Output B an "identity" mechanism. If we manipulate only A but B is influenced by some extraneous causal activity in the environment, unknown or orthogonal to the user's activity but following a fixed probability distribution, then examining Output A or B will reliably produce values according to two different probability distributions, potentially of value to the user.

There is no obvious way to say which of these functions is the correct or inherent one in either the "grandfather clock" or the "xor-and-adder." The mechanism simply does what it does from input state to end state. Users privilege some arbitrary structure according to their needs.

3.3. Meaning assignment arbitrariness

Let us assume we have solved both the environment delineation and mechanism designation problems. We now have a specified causal mechanism which takes a specified set of (manipulable) inputs and transforms them into a specified set of (accessible) outputs. The question remains, what do those inputs and outputs mean? What values should be assigned to them? Whether they "need" to mean anything for CTM is discussed in §4.1, but for now, let's specify the arbitrariness that can exist at this step.

Seminal papers on this topic have focused on this type of arbitrariness. Searle (1990) invites us to consider a nearby wall. Given the wall's sheer scale and complexity at the atomic level, there should be some way of assigning 0s and 1s to different bits of it such that it maps onto the 0s and 1s in a Word Processing program, at least at a particular moment in time. And with sufficient complexity, it could map increasing iterations of a given Word Processing program, paralleling what is happening on my laptop right now.

In some ways, this is similar to a Library of Babel argument (Borges, 2000 [1941]). Assuming a mapping between natural numbers and letters or logical operations (e.g. ASCII code or Peano arithmetic), somewhere in the expansion of π lies every book ever written and every algorithm that could be imagined for a Turing Machine. Physical systems are replete with features corresponding to very many very long natural numbers. Such numbers may be limited from denumerable infinity by Planck lengths and ontological uncertainties, but nonetheless we should expect very many algorithms to be present in one form or another,

especially given the huge choice of different mapping functions to letters or logical operations.[4]

If those early examples seem a little fanciful, definitions have since been refined to remove much of their bite, requiring that a computation has some causal, counterfactual, or dispositional structure (e.g. Blackmon, 2013; Chalmers, 1996; Chrisley, 1994, Piccinini & Maley, 2021).[5] The algorithms identified in Searle's Wall or very long measurements are not causally manipulable – you would need to find different meaning assignments or different locations in the wall/measurement corresponding to each specific and further manipulation of an algorithm. While it is, in principle, in the power of an observer to read such programs into a wall or lengthy natural numbers, this "offloads" considerable sensory, investigating, and information processing requirements onto the observer. Assigning the values to carry out such algorithms is little different to simply carrying out the algorithms yourself, such that these examples are no longer felt to carry much threat to CTM.

However, other cases remain where different values can be arbitrarily assigned to physical features of a defined input/output causal structure to produce different algorithms. For instance, one set of examples arbitrarily collapses different value ranges of physical features in order to represent a 0 or 1 to be used as an output. For instance, if the physical feature has values x, y, and z, we could assign x and y to 1 (and z to 0) or just x to 1 (and y, z to 0) – or several other combinations.[6] The examples of tri-stable dual gates in Shagrir (2012; he credits Black for the original idea) are of this type, i.e. a gate that does AND or OR operations depending how you assign values. Likewise, the novel examples given in Hemmo and Shenker (2021) are chosen because they generate both reversible and irreversible logical operations in the same mechanism, providing the authors traction on Landauer's principle and the second law of thermodynamics.

A more quickly presented example concerns base representation. Take any set of 0/1 input nodes which lead to a set of 0/1 output nodes in a causal mechanism. If the 0/1 values are interpreted in binary, you have one algorithm being implemented. If the 0/1 were to be interpreted instead in decimal, you have a different algorithm. It might be that one of those algorithms seems more "naturally useful" to a specific human observer, but examples can be imagined where these choices are debatable. All options exist equally subject to an observer's preference. In terms of the innate system, either none of those representations apply or all of

[4] A related form of arbitrariness that similarly offloads information processing to an observer arrives when recognising that infinitely many algorithms produce the same input/output mapping on a constrained value space (as any finite physical substrate is restricted to in practice), because the same is trivially true of mathematical operations. For instance, add any one of infinitely many steps in an algorithm that is undone/inverted later, e.g. minus one and plus one, or apply various modulo readings of the data. However, if you only care about inputs/outputs (e.g. functionalism) or care also about the detail of the causal mechanisms (e.g. mechanistic approaches), this arbitrariness can be ignored.

[5] Note that even stricter definitions of computation have not been accepted as they rule out, for instance, running virtual machines (e.g. Sprevak, 2018; Joslin, 2006). For instance, a requirement that there be some physical, locational, or structural similarity between the algorithm's purpose (or even referents) and its implementation moves too far beyond the substrate-neutrality that makes algorithms useful.

[6] Note that this type of arbitrariness can also be implemented via adding a physical structure into the mechanism which implements the given value collapse, effectively translating it into a "mechanism designation" type.

them, encompassing at least denumerably infinite options given the denumerably infinite number of bases to choose from.

These examples show the difference between meaning assignment and mechanism designation. Regardless of how you collapse x,y,z into 0/1 or what base you interpret 0/1s in, the internal causal mechanism is the same, even if the addition of novel causal structures can sometimes translate one into the other.

4. Relating illustrative CTM accounts to arbitrariness types

The examples in section 3 have demonstrated that at least some forms of arbitrariness exist, in that any given physical mechanism above a very minimal level of complexity can be seen to implement more than one algorithm, with no obvious observer-independent way to privilege one over another. Section 4 asks how these different types of arbitrariness affect particular CTM accounts' ability to maintain the three desiderata from the introduction: causal relevance, substrate neutrality, and assignment discrimination.

4.1. Surviving meaning assignment arbitrariness

Among other examples from §3.3, arbitrary meaning assignment of base and units to an algorithm's outputs of 1/0/0 might refer to 100 apples, 4 permutations, or ¼, or many other options. With value collapse, other variants are also possible, such as translating AND gates into OR gates. With no innate way of privileging one reading based on the mechanism's inner workings alone, we either conclude that all or none are taking place innately. "None" is the more plausible answer as there are in general an infinite set of such options. If any difference in information is to be read across the options, then our finite system would have infinite capabilities, which seems implausible. If there is no difference in information between the interpretations, then they may as well not exist. However, why would this matter for computational accounts of phenomenal consciousness?

The absence of semantic meaning is a challenge for some perspectives on representational CTM theories of consciousness, where they require the intentionality or aboutness of a calculation to be essential for generating phenomenal consciousness (e.g. discussion around virtualism in Bourget, 2010; Searle, 1983). It is also hard to see how higher order theories labelling "first-order states as accurate representations of reality" or conducting "general belief formation" can do so without semantic reference for representations, reality, or beliefs (following Butlin et al., 2023, p31; but noting some HOT theorists argue against construing HOT as computational, e.g. Rosenthal, 2021). In such cases, assignment discrimination is lost in a severe manner – nothing is now conscious, unless we can tolerate observer-dependence (see Appendix).

An alternative would be to apply "aboutness" to an abstract feature of a causal network, rather than a specific referent of that feature. Such a sleight of hand might work if it is considered adequate for "representing a representation" to refer simply to further processing of an input, resulting in that input becoming a conscious state. Unfortunately, this sleight of hand also sacrifices assignment discrimination from the other direction, resulting in near-panpsychism. If an input node is defined as a first order representation (e.g. the visual sensor

that responds to light in the environment), then the first subsequent node is already a second order representation, since it is operating on information coming from the first. If merely having two or three steps in an algorithm or nodes in a causal network is adequate for phenomenal consciousness, then it is truly ubiquitous in physical systems, e.g. a simple hand-held calculator or even a sand-timer trivially have many such prior-input-dependent steps.

One attractive escape route is to declare that only higher order representations of adequate complexity or spatiotemporal intensity are adequate, or perhaps only representations of adequately complex lower order or sufficiently pre-processed representations. Such an escape route must motivate why increasing complexity matters – indeed matters so much that it causes a new type of phenomenon to emerge (first person perspectives from "blind" interactions). A gradually sliding scale does not suffice to save assignment discrimination since we then have panpsychism or pan-protopsychism – some dim consciousness must be present almost everywhere. Rather we need to invoke a phase transition.

Phase transitions do exist in nature and do cause dramatically different phenomena to emerge: liquid water turning solid, matter to combust into stars, stars to collapse into black holes, etc. However, simply gesturing at "complexity" is inadequate to convince any but the already converted. The phase transition mechanism must be explained along with an account for why it should cause a first person perspective to appear. Importantly, specifying such a mechanism would not prevent near-panpsychism directly, it would only enable an investigative process to begin. Once we know what mechanisms to look for, we must then ask what physical structures might be capable of implementing them and how common they are.

A different argument is that a computation definitionally only takes place if some value is assigned to an output, otherwise it is a mere automation or mechanism (Shagrir, 2001; 2020; Sprevak, 2010). Superficially this sounds like a threat to all *computational* theories of mind, not just representational accounts since any computation therefore relies on an external observer's preference. However, CTM theorists can generally survive this by adopting a different definition of computation: the causal structure already provides significant algorithmic activity, even prior to assigning meaning to inputs and outputs. Such theories survive meaning assignment arbitrariness, but must still confront the remaining two types.

4.2. Surviving mechanism designation arbitrariness

Any algorithm supervening on a physical structure has a boundary with an external environment, from which it receives inputs and to which it provides outputs. Depending on which areas and aspects of that boundary are selected as inputs and outputs, algorithms with different causal structures can be implemented. Descriptions of algorithms or drawings of causal structure diagrams neatly abstract this arbitrariness away. The inputs simply are what they are specified to be. There are no fewer and no more than exactly those on the diagram.

Unfortunately, when implementing a proposed CTM algorithm in a physical system, choices must almost always be made (see §3.2). The exception is a very strict robust mapping condition such that the transformations in the physical systems map one-to-one to a single computation (Piccinini, 2015; Ritchie & Piccinini, 2018). Likewise, the environment

externalism account hopes that once all input/output interactions are considered simultaneously, only one algorithm would remain (see critique in Hemmo & Shenker, 2021, and discussion in Shagrir, 2001). But denying the freedom in mechanism designation while retaining algorithmic utility places even greater pressure on environment delineation. One option for investigative work, discussed in §4.3, is to identify CTM accounts and physical systems that meet these conditions, but the accounts in this paper have not yet demonstrated sufficient uniqueness to isolate human consciousness over simple physical systems.

If we must have some algorithm running, but cannot isolate one, our remaining option is to conclude all possible algorithms are being run on any given physical mechanism (e.g. as argued in Miłkowski, 2013). If we do this, is computational near-panpsychism unavoidable? Not necessarily.

In principle, a computational causal structure that generates consciousness might be very complicated. OpenAI's former chief scientist, Sutskever, has suggested "today's large neural networks are slightly conscious", reflecting on similarities to the Boltzmann Brain concept given the capabilities emergent from their considerable complexity (Heaven, 2023). Depending on how you measure it, human brains have 10^{11} neurons, 10^{14} connections, and far more degrees of freedom if we examine microtubules, astrocytes, and other cells in the central nervous system (Herculano-Houzel, 2009). Perhaps the causal structure that generates consciousness is not just enormous repetition and compounding of a simple causal mechanism, such as the challenge to higher order theories in §4.1, but actually a highly intricate causal mechanism itself.

This distinction matters because the intuition pointed to by theorists such as Shagrir (2012) and Hinckfuss' Pail (Sprevak, 2018) is that physical systems at the human scale are so vast that they almost surely capture all patterns. Excluding the meaning assignment type arguments around large numbers (discussed in §3.3), this intuition could be misleading. Physical systems are indeed vast, e.g. some 10^{27} atoms in a pail of water, resulting in an incredibly vast combinatorial space, but the necessary causal structure might also be incredibly rare. How often does something extremely rare occur in an extremely large space? Juggling two opposing infinities does not lead to safe intuitions (e.g. Jones, 2015). Specific instances need to be interrogated, not resolved en masse based on general principle.

More importantly, the vast combinatorial space of physical systems does not translate into a vast combinatorial space of ever more intricate causal mechanisms. Imagine a flat plane with a billion identical inert grains of sand that can be placed in any fashion but with no other external influences (no wind, no plants, etc). Grains can be arranged in a combinatorially vast number of ways. However, assuming the grains are smooth and incompressible, the number of causal mechanisms that can be physically executed is modest. Grains can be placed on each other and influence the resulting shape (gravity exists). Patterns and heaps of different sizes can be formed on the plane and disrupted, but each grain of sand can only fall to the surface once, so we cannot create causal mechanisms within the system that chain patterns together. How would a large number of interlinkable universal logic gates be produced?

What other options exist with such sand? We could "draw out" a Turing Machine or play Conway's Game of Life ourselves, but these rely on some meaning assignment from an external observer – the mechanism is not actually mechanistically causal. We may look to quantum interactions within the sand to find more complex causal mechanisms, but here still options are limited, since the quantum mechanical effects also limit our ability to chain together those internal interactions between different grains and reduce the reliability of the algorithm as a whole.

Where does this leave us? A vast combinatorial physical space of relative positions does not guarantee a vast combinatorial space of ever-more complex causal mechanisms. But in a natural world of more complex phenomena than grains of sand, we also cannot rule out that complex CTM algorithms may be executed in systems that we cannot interact with. Perhaps a tree or a pond has the necessary complexity; perhaps sections of the ocean or Earth's mantle have phenomenal consciousness. It would certainly not be trivial to rule it out. Rather than trading intuitions about vast combinatorial spaces, support for assignment discrimination in the face of mechanism designation arbitrariness is better seen as an investigative question. The first step, of course, is for CTM theorists to propose specific algorithms that are complex enough not to be trivially present in even simple, common physical systems. We must then investigate physical systems to see which might realise those algorithms.

At present, CTM theorists gesture towards complexity rather than specify it, meaning this question is not yet mature enough to be addressed. Butlin et al. (2023) have begun the exercise of extracting computational indicators of consciousness from leading CTM theories. So far, a minimalist interpretation of these indicators could be widely realised, allowing plenty of space for near-panpsychism.

For instance, pending some novel phase transition to be specified (see §4.1), the "modules" defined in a global workspace theory sense as "specialised systems capable of operating in parallel" (ibid, p46) are common in modern computing systems and identifiable in nature. The "global broadcast" of information to multiple modules is also straightforward, met in a de minimis sense when any node's state is an input into at least two other nodes which do different operations on it. State-dependent attention may be harder to identify in nature, but is straightforward to point to in simple, early computing systems. Given the account of environment interactions via inputs and outputs, principles of embodiment (ibid, p43) are also universally met at a minimum level. Perhaps embodiment only ignites consciousness after, say, 10,000 or 10^{10} degrees of freedom in external input/output interactions – but such a phase transition is yet to be motivated. Until then, it remains an act of faith. Likewise, some minimal version of predictive processing principles can be implemented in a simple programme. The higher order theory requirement of "noisy perception modules" (ibid, p46) could simply be met by taking any existing input sensor and adding noise or degrading it.

Butlin et al. (2023) do, however, emphasise that this research is ongoing and it is unclear yet how many indicators need to be met to declare phenomenal consciousness present at a given confidence level. For instance, algorithmic recurrence is described as a necessary indicator

but a "weak condition that many AI systems already meet" (ibid, p21).[7] This paper is best seen as a call to action to continue their research programme and a motivation for specifying phase transitions, rather than drawing a terminal conclusion given the work to date.

4.3. Surviving environment delineation arbitrariness

Modern theories of physics describe the universe as richly causally interconnected, providing many structures for implementing algorithms. Such structures interconnect and extend indefinitely, switching between different, mutually influencing substrate modalities like mechanical interactions, electromagnetic interactions, and so on. As discussed in §3.1.2, human designers manage this arbitrariness in practice by specifying, from the outside, certain substrate modalities to pay attention to, often with accompanying sensitivity thresholds. However, CTM needs to specify such restrictions from the inside or innately. What are its options for doing so?

One option is to specify a feature of the substrate that creates an ontologically hard boundary. We need some qualitative structural difference between the entities that are bound together and the entities that are outside – a difference that is not simply a matter of external choice, perspective, or scale. Gómez-Emilsson and Percy (2023) discuss potential physical candidates for this, each of which would need further empirical investigation and modelling to test their suitability (e.g. quantum entanglement, resonance, electromagnetic field topology).

Unfortunately, even if it succeeds, this option dilutes the substrate neutrality desideratum. For instance, a Turing Machine might be able to model the effects of quantum entanglement with arbitrary accuracy, but none of that creates actual quantum entanglement within the Turing Machine itself. To execute something that relies on actual quantum entanglement requires limiting the algorithm to a substrate that exhibits it. Unless such entanglement is computationally necessary (i.e. arbitrarily accurate simulation can be shown to fail), the restriction is an arbitrary weakening of substrate neutrality.

A second option is to allow only causal mechanisms that supervene on the same substrate modality to count for forming a CTM algorithm, leaning on dissipation at the edges of one system before the next begins or transitions from one substrate modality to another to carve up the boundaries. For instance, sets of causal mechanisms solely on the electromagnetism substrate or solely on the mechanical substrate can be chained together to execute CTM, but not sets of mechanisms that use both. Even if willing to tolerate such dilution of substrate-neutrality, we now have one rule for “phenomenal consciousness” algorithms and one rule for all other algorithms that humans exploit. As described in §3.1.3, our current laptops and imaginable automated systems happily transition between causal substrates as a core part of their utility. Such a rule may be built into the fabric of reality, but it is hard to identify a natural motivation for it. One possible challenge is the reliance on substrate modalities to

[7] Recurrence is also prey to a Bartlett (2012) style critique. Since the algorithms operate locally step by step, what happens if we delete the loop step once it has been passed but before the information re-enters the algorithm at an earlier point? We have a condition for consciousness that either does not matter or matters even after it is deleted (or never used), depriving the mechanism of causal relevance.

have discrete types. While describable as such for scientific convenience, are the edges between modalities always ontologically strict?

The third option follows the favoured solution to mechanism designation arbitrariness: allow all the algorithms to exist simultaneously. Every possible way of carving up the causal structure of the universe into different substructures, using individual or multiple sets of substrate modality, all implement their implied algorithms. Unfortunately, the temporary delay on panpsychism of "investigation-pending" in mechanism designation does not apply here. Since we know at least some entities are conscious (or at least one: you, dear reader) and we know those entities are causally interconnected with their environment, the algorithms that generate consciousness also exist as subsets of an arbitrarily large number of algorithms within the environment.

Under this third option, both the Problem of Many Minds (Monton & Goldberg, 2006; see also Shagrir, 2012) and the Hard Problem of the Many (Simon, 2017) may apply, depending on the circumstances, opening the door to near-panpsychism. With effort, we may be able to rule out that a river is conscious under a particular future CTM account, pending the necessary investigative work called for in §4.2. But we cannot rule consciousness out for the algorithms supervening on the joint causal system of a person interacting with a river, being already complex enough by definition to implement a CTM in one of its components.

These problems extend incrementally but indefinitely from one phenomenally conscious algorithm into their environment, through the unending set of causal connections across different substrate modalities, threatening potential solutions to computational plurality. For instance, Shagrir (2020) discusses (but does not argue for) the maximal option for "picking one" algorithm out of those available, being the algorithm not implied by any other. Unfortunately, in this toy example, it is the person algorithm that is implied by the joint system of "river + person", not the other way round. Joslin (2006) proposes a "reverse engineering" solution to narrow down the options, but admits this only excludes the worst excesses of panpsychism; there would still be multiple algorithms implemented in most systems (just not all algorithms). Perhaps future research could follow this trail to a better answer.

Could we argue that the CTM algorithm and only that algorithm suffices for consciousness? For instance, as soon as you add any extra causal step of any type at its boundary, even if the rest of the algorithm's billions of steps remain just as functional as before, the new algorithm is no longer conscious. In other words, the consciousness resides only in the subpart that is exactly coterminous with the CTM algorithm. Perhaps. But this is also a risky concession that needs investigation before declaring victory. Under evolution and the presumed consciousness of our children and parents, we must allow that some additions or changes to a conscious physical causal structure (and the set of algorithms it therefore executes) do not remove its consciousness. A CTM theorist could perhaps argue that some additions are allowed and others are not, but it is hard to see how this could have a natural motivation without diluting substrate-neutrality or implying an unspecified phase transition. The idea also needs reconciling with the plausible identification of algorithmic redundancy in the

human brain. In any case, as elsewhere, this option would need to be specified and investigated to assess whether it avoids near-panpsychism.

IIT has the clearest argument for surviving this arbitrariness, but at considerable cost. The principle of exclusion assumes that the "most conscious" (i.e. highest phi) subset in a system extinguishes all the others (Albantakis et al., 2023).[8] However, in these cases, we lose both assignment discrimination and causal relevance. For the first, IIT's traditional exposition famously already allows for XOR gates and protons to be dimly conscious (e.g. Horgan, 2015). For the second, the fact of being the "most conscious" causal structure is epiphenomenal, in that it has no causal relevance above and beyond the causal mechanisms already being implemented. Placing causal weight on the mere presence (but not usage) of counterfactuals also risks undermining causal relevance (Bartlett, 2012).

Conclusion

This paper has provided definitions and examples for three types of arbitrariness faced by implementing algorithms on physical systems: meaning assignment, mechanism designation, and environment delineation. Different CTM accounts can adopt strategies to survive, but eventually at least one of CTM's causal relevance, substrate neutrality, or assignment discrimination desiderata is at risk.

Certain perspectives on representational CTM accounts fail to sustain assignment discrimination in the face of meaning assignment arbitrariness, as do prima facie interpretations of higher order theories. These theories have a possible escape route by motivating a phase transition mechanism: some level of complexity below which consciousness does not ignite. However, this escape route provides no guarantees and has a high barrier to entry. Theorists must specify the phase transition mechanism, unlocking investigative work to then examine its likely presence or absence in different physical structures. Without such work, we cannot be confident either way about near-panpsychism across everyday physical systems.

Causal structure and input/output functionalist CTM accounts are unscathed by meaning assignment arbitrariness, but they (as well as representational and higher-order theories) must still address mechanism designation arbitrariness. Several structures explicitly described in theory today typically admit near-panpsychism, such as algorithmic recurrence, global broadcast, or modularisation. However, these may be necessary rather than sufficient conditions. In any case, the theories are not yet mature enough to permit an assessment of the assignment discrimination desideratum. As with the phase transitions just mentioned, a second investigative avenue is proposed, of preparing specific computations for comparison against actual physical systems, continuing the valuable work begun by Butlin et al. (2023).

[8] The principle of exclusion also leads to a high-stakes inference that there is no system within which humans are nested which is more complex than us, including the universe in all its vastness and all its combinatorial subsets that include us (else our human-level consciousness would wink out and appear instead in the other system).

Until this work is done, such theories may or may not rule out the consciousness of sunflowers, streams, or solar systems.

All CTM accounts face the environment delineation challenge. Options for isolating particular causal mechanisms can work, but at the cost of diluting substrate neutrality. All possible algorithms could be allowed, although a strong prior given the flexibility of evolution would lead to near-panpsychism and the sacrifice of assignment discrimination. A third avenue for investigative work would be to specify sets of algorithms which both allow evolutionarily diverse entities to be conscious but prevent others. IIT's principle of exclusion avoids this type of arbitrariness, but at the cost of both causal relevance and assignment discrimination. Options to save CTM through changing the scope of the paper or accepting observer-dependence are discussed in the Appendix, but appear unlikely to succeed without changing CTM beyond recognition.

We close by wondering whether sacrificing assignment discrimination is so harmful, if it means causal relevance and substrate-neutrality are retained. The intuition against an unilluminating promiscuity of consciousness (Rosenberg, 2004, s4.8) was perhaps never well-grounded and we would lose little by giving it up (Roelofs, 2022).

The anti-panpsychism intuition may have its roots in our evolved "theory of mind", in the sense of the psychological ability that develops around age 1-3 to understand what the world looks like from another person's perspective (e.g. Call & Tomasello, 1998). This evolved ability has good evolutionary reasons to restrict the application of "other minds" to entities that can take actions in response to our actions that affect our homeostasis, notably animals, particularly other people. If so, there is good reason not to lean too heavily on its intuitions in philosophy or science. We have already learned the universe of objects is vastly larger and vastly weirder than the spaces and societies we evolved to interact with, perhaps too is the universe of minds. In any case, panpsychism and CTM share something else in common: a more severe challenge than the arbitrariness issues in this paper – the combination or binding problem, as described by Bartlett (2012) for algorithms and Chalmers (2016) for panpsychism.

Does such near-panpsychism introduce ethical concerns, with new numbers of moral wellbeing subjects that dwarf our own? Perhaps, but perhaps not concerns we need concern ourselves with. As Gottlieb & Fischer (2024) argue, if we cannot interact with these other subjects and have no way of knowing what activities produce negative or positive valence for them (or even if they experience valence), we are under little moral obligation to temper our actions on their account. In any case, this is no argument against the possible truth of near-panpsychism, only about what we might do if it were true. We commend such a near-panpsychist pluralism to the computationalists.

References

Albantakis, L., Barbosa, L., Findlay, G., Grasso, M., Haun, A. M., Marshall, W., ... & Tononi, G. (2023). Integrated information theory (IIT) 4.0: formulating the properties of phenomenal existence in physical terms. PLoS Computational Biology, 19(10), e1011465.

Barad, K. (2022). Agential realism - A relation ontology interpretation of quantum physics. In O. Freire (Ed.) The Oxford Handbook of the History of Quantum Interpretations. pp1031–1054. Oxford: OUP. https://doi.org/10.1093/oxfordhb/9780198844495.013.0043

Barnden, J. (2020). The Meta-Dynamic Nature of Consciousness. Entropy 22, no. 12: 1433. https://doi.org/10.3390/e22121433

Bartlett, G. (2012). Computational theories of conscious experience: between a rock and a hard place. Erkenntnis, 76, 195-209.

Blackmon, J. (2013). Searle's Wall. Erkenn 78, 109–117. https://doi.org/10.1007/s10670-012-9405-4

Bongard, J., & Levin, M. (2023). There's Plenty of Room Right Here: Biological Systems as Evolved, Overloaded, Multi-Scale Machines. Biomimetics, 8, 110. https://doi.org/10.3390/biomimetics8010110

Borges, J. (2000 [1941]). The Total Library: Non-Fiction 1922–1986. London: The Penguin Press

Bourget, D. (2010). The representational theory of consciousness [Dissertation]. Australian National University

Bourget, D., & Chalmers, D. J. (2023). Philosophers on Philosophy: The PhilPapers 2020 Survey. Philosophers' Imprint.

Butlin, P., Long, R., Elmoznino, E., Bengio, Y., Birch, J., Constant, A., ... & VanRullen, R. (2023). Consciousness in artificial intelligence: insights from the science of consciousness. arXiv preprint arXiv:2308.08708.

Call, J., & Tomasello, M. (1998). Distinguishing intentional from accidental actions in orangutans (Pongo pygmaeus), chimpanzees (Pan troglodytes), and human children (Homo sapiens). Journal of Comparative Psychology. 112 (2): 192–206. 10.1037/0735-7036.112.2.19

Chalmers, D. (1996). Does a rock implement every finite-state automaton? Synthese 108, 309–333. https://doi.org/10.1007/BF00413692

Chalmers, D. (2016). The combination problem for panpsychism. In G. Bruntrup and L. Jaskolla (Eds.) Panpsychism: Contemporary perspectivess. pp179–214. Oxford: Oxford University Press. 10.1093/acprof:oso/9780199359943.003.0008

Chrisley, R.L. (1994). Why everything doesn't realize every computation. Mind Mach 4, 403–420. https://doi.org/10.1007/BF00974167

Copeland, B. J. (2002). Hypercomputation. Minds and Machines, 12(4), 461–502.

Deutsch, D., & Marletto, C. (2015). Constructor theory of information. Proc. R. Soc. A.4712014054020140540 http://doi.org/10.1098/rspa.2014.0540

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

Fages, F., Le Guludec, G., Bournez, O., & Pouly, A. (2017). Strong Turing Completeness of Continuous Chemical Reaction Networks and Compilation of Mixed Analog-Digital Programs. In Feret, J., Koeppl, H. (eds) Computational Methods in Systems Biology. CMSB 2017. Lecture Notes in Computer Science, vol 10545. Springer, Cham. https://doi.org/10.1007/978-3-319-67471-1_7

Francken, J. C., Beerendonk, L., Molenaar, D., Fahrenfort, J. J., Kiverstein, J. D., Seth, A. K., & van Gaal, S. (2022). An academic survey on theoretical foundations, common assumptions and the current state of consciousness science. Neuroscience of consciousness, 2022(1), niac011. https://doi.org/10.1093/nc/niac011

Frisch, M. (2023). Causation in Physics. In E. Zalta & U. Nodelman (eds.) The Stanford Encyclopedia of Philosophy (Winter 2023 Edition). https://plato.stanford.edu/archives/win2023/entries/causation-physics/

Georgiev, D. (2024). Evolution of Consciousness. Life. 14(1):48. https://doi.org/10.3390/life14010048

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

Goldin, D., & Wegner, P. (2008). The Interactive Nature of Computing: Refuting the Strong Church–Turing Thesis. Minds & Machines 18, 17–38. https://doi.org/10.1007/s11023-007-9083-1

Gómez-Emilsson, A. & Percy, C. (2022). The "Slicing Problem" for Computational Theories of Consciousness. Open Philosophy, 5(1), 718-736. https://doi.org/10.1515/opphil-2022-0225

Gómez-Emilsson, A., & Percy, C. (2023). Don't forget the boundary problem! How EM field topology can address the overlooked cousin to the binding problem for consciousness. Front. Hum. Neurosci. 17:1233119. 10.3389/fnhum.2023.1233119

Gottlieb, J., & Fischer, B. (2024). The ethical implications of panpsychism. Australasian Journal of Philosophy. 1–15. https://doi.org/10.1080/00048402.2024.2350708

Heaven, W. (2023). Rogue superintelligence and merging with machines: Inside the mind of OpenAI's chief scientist. MIT Technology Review. https://www.technologyreview.com/2023/10/26/1082398/exclusive-ilya-sutskever-openais-chief-scientist-on-his-hopes-and-fears-for-the-future-of-ai/

Hemmo, M., and Shenker, O. (2019) The Physics of Implementing Logic: Landauer's Principle and the Multiple-Computations Theorem. Studies in History and Philosophy of Modern Physics, 68, 90-105.

Hemmo, M., & Shenker, O. (2021). A Challenge to the Second Law of Thermodynamics from Cognitive Science and Vice Versa. Synthese. Volume 199, pages 4897–4927.

Hemmo, M., & Shenker, O. (2022a). Flat physicalism. Theoria, 88(4), 743-764.

Hemmo, M., & Shenker, O. (2022b). The Multiple-Computations Theorem and the Physics of Singling Out a Computation. The Monist, Volume 105, Issue 2, April 2022, Pages 175–193, https://doi.org/10.1093/monist/onab030

Herculano-Houzel, S. (2009). The human brain in numbers: a linearly scaled-up primate brain. Frontiers in human neuroscience, 3, 31. https://doi.org/10.3389/neuro.09.031.2009

Hill, R.K. (2016). What an algorithm is. Philosophy & Technology, 29(1): 35–59.

Hoffman, D., Prakash, C., & Prentner, R. (2023). Fusions of Consciousness. Entropy, 25, 129. https://doi.org/10.3390/e25010129

Hofstadter, D. (2007). I Am a Strange Loop. New York : Basic Books

Horgan, J. (2015). Can Integrated Information Theory Explain Consciousness? Scientific American (1 Dec 2015). https://blogs.scientificamerican.com/cross-check/can-integrated-information-theory-explain-consciousness/

Jones, S. (2015). Calculus limits involving infinity: the role of students' informal dynamic reasoning. International Journal of Mathematical Education in Science and Technology, 46:1, 105-126, 10.1080/0020739X.2014.941427

Joslin, D. (2006). Real realization: Dennett's real patterns versus Putnam's ubiquitous automata. Mind Mach 16, 29–41. https://doi.org/10.1007/s11023-006-9009-3

Knuth, D. E. (1973). The Art of Computer Programming, second edition. Reading, MA: Addison-Wesley.

Lachmann, M., & Sella, G. (1995). The computationally complete ant colony: Global coordination in a system with no hierarchy. In Morán, F., Moreno, A., Merelo, J.J., Chacón, P. (eds) Advances in Artificial Life. ECAL 1995. Lecture Notes in Computer Science, vol 929. Springer, Berlin, Heidelberg. https://doi.org/10.1007/3-540-59496-5_343

Lahav, N., & Neemeh, Z. A. (2022). A relativistic theory of consciousness. Frontiers in Psychology, 12, 704270.

Ludwig, K. (2003). The Mind-Body Problem: An Overview. In S.P. Stich and T.A. Warfield (Eds) The Blackwell Guide to Philosophy of Mind. https://doi.org/10.1002/9780470998762.ch1

Marshall, S., Moore, D., Murray, A., Walker, S., & Cronin, L. (2022). Formalising the Pathways to Life Using Assembly Spaces. Entropy 24, no. 7: 884. https://doi.org/10.3390/e24070884

Metzinger, T. (2020). Minimal phenomenal experience: Meditation, tonic alertness, and the phenomenology of "pure" consciousness. Philosophy and the Mind Sciences, 1(I), 7. https://doi.org/10.33735/phimisci.2020.I.46

Miłkowski, M. (2013). Explaining the computational mind. Cambridge: MIT Press.

Miłkowski, M. (2014). Is the mind a Turing machine? How could we tell? In A. Olszewski, B. Brożek, & P. Urbańczyk (Eds). Church's Thesis. Logic, Mind, and Nature (pp.305-333). Copernicus Center Press

Miłkowski, M. (2016). Computation and multiple realizability. Fundamental issues of artificial intelligence, 29-41.

Miłkowski, M. (2017). Why Think That the Brain Is Not a Computer? Apa Newsletter Philosophy And Computers Spring 2017 Volume 16 Number 2 pp22-28

Montero, B.G. (2022). Mathematical platonism and the causal relevance of abstracta. Synthese 200, 494. https://doi.org/10.1007/s11229-022-03962-x

Monton, B., & Goldberg, S. (2006). The problem of the many minds. Minds & Machines 16, 463–470. https://doi.org/10.1007/s11023-006-9045-z

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

Pearl, J. (2000). Causality. Cambridge: Cambridge University Press.

Piccinini, G. (2015). Physical Computation: A Mechanistic Account, Oxford: Oxford University Press.

Piccinini, G., & Maley, C. (2021). Computation in Physical Systems. In E. Zalta (Ed). The Stanford Encyclopedia of Philosophy (Summer 2021 Edition). https://plato.stanford.edu/archives/sum2021/entries/computation-physicalsystems/

Piper, M. (2012). You Can't Eat Causal Cake with an Abstract Fork An Argument Against Computational Theories of Consciousness. Journal of Consciousness Studies, Volume 19, Numbers 11-12, 2012, pp. 154-190(37)

Primiero, G. (2020). On the Foundations of Computing. New York: Oxford University Press.

Rescorla, M. (2014). A theory of computational implementation. Synthese 191, 1277–1307. https://doi.org/10.1007/s11229-013-0324-y

Rescorla, M. (2020). The Computational Theory of Mind. In E. Zalta (Ed.) The Stanford Encyclopedia of Philosophy (Fall 2020 Edition). https://plato.stanford.edu/archives/fall2020/entries/computational-mind

Ritchie, J. B., & Piccinini, G. (2018). Computational implementation. In M. Sprevak and M. Colombo (Eds). The Routledge handbook of the computational mind, 192-204. Abingdon: Routledge. 10.4324/9781315643670-15

Robinson, H. (2023). Dualism. In E. Zalta & U. Nodelman (eds.) The Stanford Encyclopedia of Philosophy (Spring 2023 Edition). https://plato.stanford.edu/archives/spr2023/entries/dualism/

Roelofs, L. (2022). No Such Thing as Too Many Minds. Australasian Journal of Philosophy. 10.1080/00048402.2022.2084758

Rosenberg, G. H. (2004). The boundary problem for experiencing subjects. In G. Rosenberg (Ed). A place for consciousness: Probing the deep structure of the natural world. Oxford: Oxford University Press.

Rosenthal, D. (2021). Assessing criteria for theories. Cognitive Neuroscience, 12(2), 84-85.

Sandberg, A., & Bostrom, N. (2008). Whole brain emulation: A roadmap. Oxford: Future of Humanity Institute, University of Oxford. http://www.fhi.ox.ac.uk/reports/2008-3.pdf

Schrödinger, E. (1951). Mein Leben, meine Weltansicht [My Life, My Worldview]. Cambridge: Cambridge University Press.

Schweizer, P. (2019). Triviality Arguments Reconsidered. Minds & Machines 29, 287–308. https://doi.org/10.1007/s11023-019-09501-x

Searle, J. R. (1983). Intentionality: An Essay in the Philosophy of Mind. New York: Cambridge University Press.

Searle, J. R. (1990). Is the brain a digital computer?. In Proceedings and addresses of the American Philosophical Association (Vol. 64, No. 3, pp. 21-37). American Philosophical Association.

Shagrir, O. (2001), Content, Computation and Externalism. Mind, 110, 369-400.

Shagrir, O. (2012). Can a Brain Possess Two Minds? Journal of Cognitive Science, 13, 145-165.

Shagrir, O. (2020). In defense of the semantic view of computation. Synthese 197, 4083–4108. https://doi.org/10.1007/s11229-018-01921-z

Simon, J. A. (2017). The Hard Problem Of The Many. Philosophical Perspectives, 31, 449–468.

Sprevak, M. (2010) Computation, Individuation, and the Received View on Representation. Studies in History and Philosophy of Science, 41, 260–270.

Sprevak, M. (2018). Triviality Arguments About Computational Implementation. In Sprevak, M., & Colombo, M. (Eds.). (2018). The Routledge Handbook of the Computational Mind (1st ed.). pp. 175-191. Abingdon: Routledge. https://doi.org/10.4324/9781315643670

Stallings, W. (2015). Computer Organization and Architecture (10th edition). London: Pearson.

Stoljar, D. (2024). Physicalism. In Edward N. Zalta & Uri Nodelman (eds.) The Stanford Encyclopedia of Philosophy (Spring 2024 Edition), https://plato.stanford.edu/archives/spr2024/entries/physicalism

Sulis, W. (2020). An information ontology for the process algebra model of non-relativistic quantum mechanics. Entropy. 10.3390/e22020136

Vardi, M. (2012). What is an algorithm? Communications of the ACM, 55(3): 5. 10.1145/2093548.2093549

Westphal, J. (2016). The Mind-Body Problem. Cambridge, MA: MIT Press

Wheeler, J. A., (1989). Information, Physics, Quantum: The Search for Links. Proceedings III International Symposium on Foundations of Quantum Mechanics, Tokyo, 354–358.

Woods, D., & Neary, T. (2007). The Complexity of Small Universal Turing Machines. In Cooper, S.B., Löwe, B., & Sorbi, A. (eds) Computation and Logic in the Real World. CiE Lecture Notes in Computer Science, vol 4497. Springer, Berlin, Heidelberg. https://doi.org/10.1007/978-3-540-73001-9_84

Appendix. Other ways to avoid the trilemma

This appendix considers the potential to avoid the trilemma via three broad strategies. First, whether we can tolerate observer-dependency, concluding that there is significant theoretical work required to translate the various methods into a CTM perspective and see if they can sustain the three desiderata at acceptable cost. Second, whether we can look outside the scope restrictions in this paper. Thirdly, whether we can relax its two metaphysical assumptions. In these latter two strategies, unfortunately, it seems that any success would come at the cost of introducing the mind-body interaction problem so challenging for dualism.

A.1. Tolerating observer dependency

The introduction discussed researchers favouring observer dependency, albeit primarily focused on cognition rather than consciousness. Observer dependency is sometimes also implicit in other accounts. For instance, only functional mechanisms are allowed to instantiate CTM algorithms under some mechanistic accounts, defined as "complex systems of components that are organized to perform functions" (see overview in Piccinini & Maley, 2021). But without an external observer, it is not obvious what counts as complex or as a function. Where along the sliding scale of evolutionary journeys would we place the limit without arbitrariness, given grey areas, edge cases, and the possible perspectives of very different entities? Human brains have a function to think? Viruses have a function to replicate? Rivers have a function to flow downhill?

We identify three possible ways to tolerate observer dependency: explicitly adopt a relational/relativistic approach on consciousness, embrace infinite regress, and a consciousness pluralism. It is hard to see how any of these ways gels well with CTM, but CTM theorists may wish to develop an account leveraging these or related principles.

Both relational perspectives (e.g. Barad, 2022) and relativist perspectives (e.g. Lahav & Neemeh, 2022) on consciousness adopt a fundamentally different view on what it means to be an entity that experiences the world from a "first person perspective". A relational perspective might argue that it is the relationships between algorithms that are conscious, rather than algorithms themselves. In the context of this paper, we might argue that there is therefore always an external observer and we can be external observers for each other. Unfortunately, algorithms are relational internally at every step – each output step is the result of a transformational relationship applied to a set of inputs, leading to panpsychism. How can an inside and an outside of an algorithm be defined without re-introducing many of the problems in this paper?

Perhaps something special happens in the infinite regress situations of one system being the observer for a second system whose observations resolve arbitrariness for the first. Perhaps this special something causes a first person perspective to manifest. Some theories do explore related notions, potentially providing a novel way out of these challenges, albeit only by weakening the substrate neutrality or traditional causality of CTM, either by leaning on a strongly emergent property or by invoking a new physical modality. For instance, Hofstadter (2007) describes the concept of strange loops in which complex chains of self-reference in

the brain might lead to the emergence of a conscious self. Through this lens, logical self-referential paradoxes become the cause of consciousness, rather than a paradox to avoid at all costs, and cause-and-effect relationships can invert. Barnden (2020) introduces the idea of meta-dynamism as a novel physical phenomenon, imaginable as a self-referential, temporally non-local causal whirl and a possible explanation for consciousness.

Perhaps evolution can "bootstrap" an observer from nowhere via some very large causal or relational regression. CTM theorists may wish to explore these principles, but simply gesturing to magical emergence in the face of an infinite regress feels scientifically unsatisfying and hard to defend epistemologically. Perhaps evolution itself could be seen as intentional (in the dictionary sense), such that evolution serves as the external observer for all life. But saving CTM consciousness by reifying a traditionally non-teleological natural process and imbuing it with consciousness seems a steep price to pay, particularly for the physicalists traditionally attracted to CTM.

Finally, perhaps there are multiple types of mechanisms that produce phenomenal consciousness. CTM might produce first person perspectives subject to an external observer whose first person perspective is generated by some non-computational phenomenon, such as property dualism (overview in Robinson, 2023) or Goffian panpsychism (Goff, 2019). However, saving one account of phenomenal consciousness by introducing a second seems peculiar. Perhaps multiple types of consciousness exist, some computational and some not, but theorists would need to invoke a specific epistemology by which this surfaces as the best choice solution.

A.2. Relaxing scope restrictions

Two scope restrictions specified in §2 were that the algorithms be finite and Turing-simulable.

If we first consider a state space of algorithms that require infinite descriptions or infinite physical space, then we can avoid near-panpsychism, since the entities in consideration have a finite physical structure on which to implement algorithms (e.g. rocks, streams, cities). However, the human brain is also physically finite, so CTM would be unable to account for the explanandum that launched the entire exercise. New options may open up if we allow the finite brain to be a portal or antenna to some infinite non-physical space, but it is hard to see how the surviving account would be called CTM, as opposed to dualism, platonism, or something - and there remains the familiar dualist challenge of specifying that portal/antenna and explaining how it interacts with the physical realm.

What about "Turing-simulable"? This restriction excludes some types of computation, such as working directly with real numbers or real randomness (Copeland, 2002), certain interactive perspectives on computation (Goldin & Wegner, 2008), or accounts where computational run-time or complexity matters alongside function (e.g. Eliasmith, 2002). Advocates typically explain that such features can be simulated with arbitrary accuracy on a Turing Machine (see also objections to hypercomputation in Miłkowski, 2014 and §4.3 of Piccinini & Maley, 2021). However, if we do require that an actual quantum phenomenon or

an exact real-valued physical input is part of the algorithm, then we have weakened substrate-neutrality.

Finally, we could look outside the scope definition of "algorithm" itself, although this may feel a sacrifice too far for CTM theorists. Rather than a specific set of state transformations, we could look to other properties of causal networks. Perhaps it is the "volume" of information in a causal network, in terms of some measure of its possible states (e.g. Shannon Entropy), that conveys conscious once executed on a relevant physical system. Or perhaps it lies in its Kolmogorov Complexity or Assembly Index (Marshall et al., 2022), or in the extent of information transformations or transmissions relative to a given scope.

Modern accounts of information may provide other avenues for non-algorithmic definitions, subject to further study. For instance, constructor theory (Deutch & Marletto, 2015) resolves concerns around defining information and distinguishability but treats a physical system under analysis as given. In principle, non-algorithmic definitions could be made safe from observer-dependency – properties can be found that are inherent in physical objects – but the environment delineation challenge still remains.

More generally, unless we can invoke the kinds of phase transitions discussed in §4.1, we are left with near-panpsychism, as almost any physical feature or object component is capable of being in two or more states or having the minimal version of other features discussed above. This route may be possible, but significant theoretical work is needed to define such a property and test its robustness against the arguments in this paper.

A.3. Relaxing metaphysical assumptions

The primary metaphysical assumptions were that CTM algorithms only generate consciousness once executed in a physical substrate.

If the concept of an algorithm, outside of a physical implementation of it, is necessary for consciousness, then it is hard to avoid the conclusion that such concepts are either ubiquitous (i.e. we have not avoided panpsychism) or nowhere (i.e. humans are not conscious), similar to how humans "discover" rather than "invent" mathematics from the perspective of mathematical platonism (e.g. Montero, 2022). Or, if humans have some special portal or antenna to a mathematical realm, we have the same issue as with non-physical space discussed above.

If a physical description or blueprint of an algorithm is adequate for consciousness (i.e. prior to its execution, and despite arguments from Hemmo & Shenker, 2019), then the meaning assignment arbitrariness challenge is returned in full, as any such description requires a language in which it can be written/read and infinite such languages exist. All possible algorithms would therefore be written in any sufficiently information-dense structure.

What about non-physicalist substrates for executing the algorithm? An immediate challenge is the issue of physical/non-physical interaction mentioned above. A second more subtle challenge is that even assuming a substrate beneath that which we experience as physical, it is hard to see how that saves CTM. If we assume a conscious agent substrate (Hoffman et al.,

2023) or informational substrate (Sulis, 2020; Wheeler, 1989) on which seemingly physical objects supervene, we are still faced with the fact that algorithms are defined in a substrate-neutral manner. If we wish to maintain substrate-neutrality, we already accepted that their execution supervenes on causal mechanisms in seemingly physical structures. The fact those physical structures supervene in turn on something else does not change anything from the algorithm's perspective. Unless a bedrock of consciousness means all things that supervene on them are conscious – but this merely brings back in panpsychism, such that CTM algorithms are trivially conscious because everything is.

A more creative metaphysical solution might be to abandon the counterfactuality of algorithmic power and embrace a block universe or, relatedly, to rely on a fixed past to produce CTM consciousness while allowing counterfactual futures to remain open in principle. If there is only ever one path taken through a causal mechanism, this dramatically collapses the space of possible algorithms enacted in a physical structure, which may make it easier to meet the investigative avenues of research described in the main paper. But what does causality really mean in such a setting?

There is no shortage of creative alternatives to the options considered in the main paper, but it is hard to see how any surviving theory could still be called CTM.