Panpsychism and AI Consciousness

Marcus Arvan — University of Tampa

Corey Maley — University of Kansas

Abstract: This article argues that if panpsychism is true, then there are grounds for thinking that digitally-based artificial intelligence (AI) may be incapable of having coherent macrophenomenal conscious experiences. Section 1 briefly surveys research indicating that neural function and phenomenal consciousness may be both analog in nature. We show that physical and phenomenal magnitudes—such as rates of neural firing and the phenomenally experienced loudness of sounds—appear to covary monotonically with the physical stimuli they represent, forming the basis for an analog relationship between the three. Section 2 then argues that if this is true and micropsychism—the panpsychist view that phenomenal consciousness or its precursors exist at a microphysical level of reality—is also true, then human brains must somehow manipulate fundamental microphysical-phenomenal magnitudes in an analog manner that renders them phenomenally coherent at a macro level. However, Section 3 argues that because digital computation abstracts away from microphysical-phenomenal magnitudes—representing cognitive functions non-monotonically in terms of digits (such as ones and zeros)—digital computation may be inherently incapable of realizing coherent macroconscious experience. Thus, if panpsychism is true, digital AI may be incapable of achieving phenomenal coherence. Finally, Section 4 briefly examines our argument's implications for Tononi's Integrated Information Theory (IIT) theory of consciousness, which we contend may need to be supplanted by a theory of macroconsciousness as *analog* microphysical-phenomenal information integration.

Keywords: analog, artificial intelligence, computation, consciousness, digital, panpsychism

There is a long-running debate over whether artificially intelligent machines (AIs) have minds, particularly whether they are phenomenally conscious. On the one hand, functionalists have long argued that mental states are functional states, and that insofar as AI could instantiate the same functional states as human brains (assuming that mental states are realized in brain states), AIs

could be phenomenally conscious no less than we are. However, others have sought to cast doubt on whether AIs could have minds in this sense. Searle (1980), for example, famously argues that AIs merely have syntactic capacities, necessarily lacking genuine semantic capacities required for consciousness. It is not our aim to enter these particular debates. Instead, we argue that the most popular version of an increasingly influential theory of the nature and distribution of consciousness—panpsychism—has underappreciated implications for consciousness in AI. Specifically, if *constitutive micropsychism*—the view that 'all facts are grounded in/realized by/constituted of consciousness-involving facts at the micro level' (Goff *et al.* 2020: Section 2.2.)—is true, then AI implemented by digital means may not be phenomenally conscious in any meaningful way, except purely—and only momentarily—by chance.

We proceed as follows. Section 1 briefly surveys evidence that human neural processing and phenomenal consciousness may be analog in nature, involving physical and phenomenal magnitudes that relate to each other (and to physical stimuli) in monotonically covarying relationships. Section 2 then argues that, if these analog relationships hold, and if and micropsychism is also true, then neural processing must somehow convert fundamental physical-phenomenal magnitudes into a *coherent* manifold of phenomenally conscious experience (viz. the kind of coherent mix of visual, auditory, kinesthetic, cognitive, and emotional phenomenal experiences that define our first-personal mental experience). Then, in Section 3, we argue that, because digital computation abstracts away from fundamental microphysical magnitudes—and hence, from monotonic (analog) relationships between various microphysical-phenomenal qualities—it follows that if our claims in Sections 1 and 2 are true, then digital AI are phenomenally scrambled, with no coherent first-personal visual, auditory, kinesthetic, cognitive, or emotional phenomenology (except potentially—and only momentarily—by remote chance).

That is, if micropsychism is true *and* the brain and phenomenal properties are both analog, then digital AI are incapable of coherent macrophenomenal experiences. However, because it is an open question whether micropsychism is true and we do not purport to prove our claims in Sections 1 and 2—but instead recognize ongoing debates about the analog-digital distinction and the extent to which brains are analog devices—our final conclusion is conditional: *if* panpsychism is true, then digital AI *may* be incapable of coherent macrophenomenal consciousness in any meaningful way. Finally, Section 4 briefly suggests that this result raises important new questions for an influential information-based approach to consciousness: Guilio Tononi's (2011, 2012) *Integrated Information Theory* (IIT). If we are correct, then IIT may not be true: (coherent) macroconsciousness may not be information integration *simpliciter*, but rather the right kind of *analog* microphysical-phenomenal information integration.[1]

[1] Our argument differs substantially from another recent line of argument that (some) digital computers may be incapable of phenomenal consciousness (Tononi and Koch 2015). Based on Tononi's Integrated Information Theory (IIT) of consciousness, which holds that phenomenal consciousness is identical to maximally integrated information, Tononi and Koch argue that two functionally identical systems can have the same input-output function, while only one of those systems integrates information whereas the other does not. On IIT, the latter system—even if it otherwise functioned like a human brain—would be a 'zombie' system with no conscious experience (see also Oizumi *et al.* 2014: pp. 19-22). Further, as Tononi and Koch (2015) and Koch (2019) elaborate, because current digital computers cannot integrate information in anything like the fine-grained way that human brains do (Koch 2019, pp. 142-4), if IIT is true, then it may take *neuromorphic electronic hardware* 'built according to the brain's design principles' for AI to 'amass sufficient intrinsic cause-effect power to feel like something' (ibid., p. 150; see also Tononi and Koch 2015, p. 16, fn. 15). Our argument is more radical than this in at least two respects. First, our argument implies that even a 'neuromorphic' digital machine could fail to realize coherent macrophenomenal consciousness—for such a machine might still fail to manipulate *fundamental microphysics* in a way necessary for combining fundamental phenomenal qualities into a coherent macrophenomenal manifold. Second, as we explain in Section 4, our argument entails that IIT itself may be false. If we are correct, then the only way for AI to have coherent macrophenomenal consciousness may be for them to be analog machines that integrate fundamental *microphysical-phenomenal* magnitudes in the right way.

1. Are Brains and Phenomenal Consciousness Analog?

There are a number of competing analyses of how digital and analog representation differ (see Goodman 1968; Katz 2008; Kulvicki 2015; and Maley 2011). While we will not enter into this debate, we will base our argument upon the conception of the analog-digital distinction in the literature that most directly characterizes analog computation (Maley 2011, forthcoming). Other conceptions of analog and digital computation might be able to accommodate our argument, but we will set that aside for now.

On the so-called Lewis-Maley conception of the analog/digital distinction (as labeled by Adams 2019), digital computation involves manipulating complex representations of numbers; those complex representations are composed of sequences of individual numerals, such as '4', '3', or '6'. (Lewis 1971; Maley 2011, forthcoming). This is the usual way we represent numbers. In decimal (i.e., base-10), the sequence of numerals '436' represents the number four hundred thirty-six. Roughly speaking, we can think of this as a three-digit name of the number; digital representation is just the representation of this kind of name.[2] Digital computers typically use binary (i.e., base-2), where four hundred thirty-six would be represented by the sequence '110110100'—but the principle is the same.

When it comes to computation, the individual numerals of a digital representation are implemented in circuit elements that are in one of two voltages: five volts implements a '1', and zero volts (or sometimes negative-five volts) represents a '0'. Thus, the entire representation of

[2] There are other ways of naming numbers, such as using words in a natural language (e.g. 'four'), using different kinds of numerical conventions (e.g. Roman numerals), or using purely conventional single symbols (e.g. π). Digital representations (and other numerical conventions) are a special kind of name that specify the value of the number named as a function of the individual numerals. Chrisomalis (2020) discusses these systems—and many others—in fascinating detail.

four hundred thirty-six is physically implemented as a series of circuit elements, the first at five volts, the second at five volts, the third at zero volts, etc., where each individual circuit element is to be interpreted as a numeral in a particular place (i.e., a digit) in the digital representation. In decimal representations, we have the ones place, the tens place, the hundreds place, and so on; in binary, we have the ones, twos, fours, and so on. The manipulation of digital representations, then, involves the manipulation of the individual digits. Again, this is something we are all quite familiar with: elementary methods of adding and multiplying numbers—the kind of thing many of us learned in primary school—involve manipulating the individual numerals of digital representations (but using pencil and paper, rather than electronic circuitry).

Analog computation, on the other hand, is much simpler. It also involves manipulating representations of numbers, but it does so in a fundamentally different manner. Instead of representing the number four hundred thirty-six via its digits, an analog representation represents this number via its magnitude. Examples here include how numbers are represented in analog thermometers or clocks. In the case of a thermometer, a number (viz. temperature) is represented by the height of a liquid; in the case of a clock, it is the angle of each individual hand. In short, analog representation is the 'representation of magnitudes, by magnitudes' (Peacocke 2019, p. 52). Importantly, unlike the received view in which 'analog' is synonymous with 'continuous,' and 'digital' is synonymous with 'discrete', the Lewis-Maley view does *not* count analog and digital as exhaustive of all representational types. There are others, such as purely conventional symbols (like pictographs, or the symbol 'π'). However, we focus on the Lewis-Maley characterization because each one forms the basis for a particular type of computation.

The numbers represented in both digital and analog devices can—and often do—represent other things, such as the color value of a pixel, a musical note, or a letter in a word

processing program.[3] At their fundamental computational level, however, digital computation involves manipulating strings of ones and zeros, which are implemented as individual sequences of alternating high and low voltage elements. Analog computation is different, reflecting the fundamentally different way that analog representation does its representing (see Beck 2015, 2019; and Maley 2011, forthcoming). In analog computation, the *physical quantities* that represent the magnitude of a number are manipulated, rather than strings of digits that represent the number. For an example of a very simple difference between the manipulation of the two kinds of representation, consider two different kinds of thermometer. A typical analog thermometer represents a temperature by the height of a column of liquid; representing a higher temperature requires a *literal* increase in the height of the liquid (e.g., the height of the column of liquid would increase from 72 mm to 73 mm, where 73 mm is *literally* taller than 72 mm). In contrast, a digital thermometer represents a temperature by a string of digits. Representing a higher temperature requires a different string of digits (e.g., the digit string '72' would change to '73'), where neither *digit string* is greater nor less than the other: just like on this very paper (or screen), the *size* of the characters is the same, even though the usual interpretations of those strings differ). Similarly, whereas digital computers represent variables by voltages that represent a string of 0s and 1s, which then represent the digits of binary numbers, analog computers represent the magnitude of variables by the magnitudes of voltages (Maley 2018a).

In short, digital representation requires the physical implementation of symbols (i.e., numerals), and those symbols (in the right sequence) in turn represent numbers (and the numbers, in turn, can represent time, temperature, musical notes, etc.). A digital representation

[3] Analog computers are virtually never used for these purposes now, although they were once the dominant computing paradigm before the 1970s.

is, in this way, a kind of second-order representation. Analog representation requires the physical implementation of *magnitudes*, and those magnitudes represent numbers (where, again, those numbers can represent time, temperature, musical notes, etc.). An analog representation is, in contrast, a kind of first-order representation.

Importantly, emerging empirical evidence suggests that brains appear to be analog computing devices, at least in many important ways (Maley 2018b, 2020). Although neurons were thought for many years to function in an all-or-nothing fashion, recent research has shown that the precise physical manifestation of action-potentials can have different computational effects (Ibid., Zblili & Debanne 2019), and that neurons can even communicate with each other through variable changes in the magnitude of electromagnetic fields surrounding their cell bodies (Chiang *et al*. 2018). These (and many other) features of neural functioning are analog because changes in particular magnitudes covary with changes in the magnitudes of what they represent. For example, certain neurons fire more frequently as the intensity of certain stimuli increase. Cochlear neurons in the inner ear increase their firing rate as loudness increases; neurons involved in vision increase their firing rate as light intensity increases. Again, these changes are literal: physical magnitudes like firing frequency, or the amplitude of the neural spike, increase in measurable ways. It is not like the change in a digital representation, where the change is only in the interpretation of the symbols manipulated.

This evidence does not prove that brains are wholly analog devices, as it is possible that brains use entirely different representational schemes that are neither analog nor digital.[4] Nevertheless, the above evidence suggests that brains are at least *partially* analog (since, again,

[4] 'Non-analog' is typically taken to mean 'digital', but there are other schemes that are neither (e.g. the symbol π, mentioned in footnote 9 above).

analog features *of* neural spikes, or groups of neural spikes, appear to affect what those spikes represent). Further, even if brains use some additional, non-analog representational scheme along with analog representations, then our argument still goes through—for, as we will now see, whatever physical-functional aspects of neural activity that phenomenal consciousness supervenes upon, *the nature of phenomenal consciousness itself* suggests that it would have to supervene in important ways upon analog features of the brain. Phenomenal consciousness appears to have analog features, and if constitutive micropsychism is true, then these analog features must supervene on *something* in fundamental microphysics—the most plausible candidate being analog changes in microphysical magnitudes (mass, charge, spin, etc.).

To see why, consider some respects in which *phenomenal consciousness* appears to be analog, at least insofar as it varies in graded magnitudes (Beck 2019). Phenomenal visual experience is one example. The phenomenal experience of redness (or any other color) is not merely 'on' or 'off' (like a one or zero). Rather, each phenomenal color has its own *hue*—in the case of red, its *redness*—as well as other magnitudes along other dimensions, such as *brightness* and *saturation*.[5] Notice, next, that these are plausibly representational differences. If, for example, one is looking at a vase of flowers, a particular hue in one's visual field (the particular redness) will represent some flowers (red roses) as being a different color than other flowers (blue irises) in another part of one's visual field. Similarly, if one looks at a lightbulb while varying the settings on a dimmer switch, different phenomenal experiences of *brightness* will represent changes in the light's physical brightness. Other types of phenomenal perceptual

[5] Although some readers may doubt whether hues are magnitudes—as hue does not come in different degrees—*particular* hues clearly do come in degrees: something can clearly be more or less red, as well as more or less of a combination of one hue with another (different orange hues, for example, are different graded combinations of red and yellow).

experiences—auditory, olfactory, kinesthetic, etc.—are similar. For example, the phenomenally experienced sound of a jet engine differs in magnitude along dimensions such as *tone*, *timbre*, *loudness*, etc. Other types of phenomenal experience (kinesthetic experience, emotional experience, etc.) have unique phenomenal magnitudes of their own. Like phenomenal redness, phenomenal sadness is not simply 'on' or 'off': it instead has distinctive phenomenal qualities that can vary in intensity along different dimensions, such as the sadness of grief, disappointment, etc.

Our basic point here is that phenomenal experiences appear to represent in an analog manner. There are different dimensions to phenomenal consciousness (and to individual conscious experiences), and those dimensions vary in a graded manner that is represented not via digits, but by magnitudes (whether they do so continuously or discontinuously is a point will return to this point later). For example, when we see and hear a purring grey cat, on a first pass we see some shade and intensity of a patch of grey on his fur and hear the frequency and intensity of his low purring. We can easily imagine what it would be like for the grey to be darker or lighter, duller or brighter. Similarly, we can easily imagine the purr being higher or lower in pitch, louder or quieter. All of these graded differences appear to be differences in magnitudes, in much the same way that physical quantities can differ in magnitude (voltage, length, temperature, or charge).

Is it possible that phenomenal consciousness merely *appears* to have analog features without actually being analog? Anything is possible—and we recognize that introspection can be unreliable (see Schwitzgebel 2006). That being said, all of the following four things seem true:

1. Philosophers who work on phenomenal consciousness routinely appeal to introspection in making arguments about the nature of consciousness (see e.g.

Chalmers 1996; Goff 2017; and Strawson 2006, 2016).

2. It is unclear what other good options there are here, as third-personal investigations of consciousness (such as Dennett's (1993) 'heterophenomenology') appear ill-suited to dealing with the *first personal* nature of phenomenal consciousness (Nagel 1974).
3. Our basic phenomenal-experiential knowledge of what *redness, greenness, yellowness* (and various shades thereof) look like appears to be about as epistemically certain as anything (since we *see* these colors and their gradations); and
4. When combined with the aforementioned evidence that brains have analog features, the *apparent* analog features of phenomenal experiences are highly suggestive: they are evidence (albeit defeasible) that phenomenal consciousness may be analog too.

2. If Panpsychism, then Macrophenomenal Coherence May Require Analog Computation

According to standard versions of panpsychism, phenomenal consciousness—or, at least, some primordial form of it ('microconsciousness')—is a fundamental feature of the physical universe, much like mass or charge (see Chalmers 2013). So, for example, whereas mass, charge, and spin are treated in physics as fundamental physical properties, panpsychists believe there are compelling grounds to think that phenomenal properties may well be fundamental to the universe as well. To simplify considerably, the basic argument for panpsychism is that because (1) phenomenal consciousness clearly exists, and (2) it seems irreducible to anything structural or functional (which is what the entities of physical sciences deal with), it follows that (3) phenomenal consciousness must somehow be the 'categorical ground' of the physical world: a nonstructural, *qualitative* part of reality 'behind' the physical properties studied by the sciences, making phenomenal consciousness a 'fundamental building block' of nature (See Goff *et al*. 2020, §3 for a more nuanced overview. Cf. Chalmers 2013; Goff 2017; Russell 1921, 1927; and

Strawson 2006, 2016). Here, we focus on the most common form of panpsychism, known as constitutive micropsychism. On this view, what we will call 'macroconsciousness' (i.e., the consciousness found in minds like ours) is somehow constituted by phenomenal properties of smaller entities. Macroconsciousness is not fundamental; its microphenomenal constituents are. As we will now see, micropsychism generates a problem of macrophenomenal coherence for digital AI precisely because of how it differs from other influential theories of consciousness (such as functionalism, Cartesian dualism, mind/brain-identity theory, etc.).

In taking microphenomenal properties to inhere in microphysics, constitutive micropsychism faces a challenge known as the 'combination problem.' This is the problem of how macroconsciousness can be 'built out of' these microphenomenal features of reality (Chalmers 2016). Although there are some promising solutions to this problem (e.g., Mørch 2014), we cannot enter into this debate. Instead, whatever the solution to the combination problem may be (and assuming our arguments in Section 1), if constitutive micropsychism is true, then it must be the case that human brains (and the brains of any other conscious creatures) somehow generate coherent macrophenomenal experiences—the kind of seamless first-personal experiences of vision, hearing, thought and feeling we all have—by (A) *manipulating* microphysical-phenomenal magnitudes in an analog manner, and (B) *combining* those magnitudes in a way that is macrophenomenally coherent. While looking at an orange over a period of time, for example, we would expect that changes in lighting (for example) would coherently alter the brightness and hue that we phenomenally experience. A visual phenomenal experience of an orange does not 'flicker' between looking, say, square one second, then shiny, then purple, then foggy, then transparent. If constitutive micropsychism is the case, then the stable, coherent macrophenomenal image of an orange we have while looking at an orange must

somehow be achieved by manipulating microphysical magnitudes, *combining* those physical magnitudes and the microphenomenal experiences that inhere in them in a manner that puts together the microphenomenal experiences in a *macro*-coherent way.

It is important to note that, much like physical units, the fundamental units of phenomenal consciousness may be *discrete*, yet still combine in a way that forms gradations of the type needed for the analog view of phenomenal consciousness. On the Lewis-Maley view we adopt, to be analog does not require continuity, but only monotonic covariation in magnitude between the representation and what is represented. That increase or decrease can happen in steps, or it can happen smoothly (i.e., discretely or continuously). For example, consider the fuel gauges found in cars. Older cars often have a physical dial that more-or-less continuously moves from 'F' to 'E' as fuel is consumed; this dial is an analog representation of the amount of fuel in the car, because as fuel decreases, so does the literal angle of the dial. Newer cars often have a different way of displaying fuel. Instead of a physical dial, fuel is displayed as a bar graph on an LED or LCD display. Importantly, that bar graph is composed of discrete segments. Nevertheless, this is still an analog representation of the amount of fuel: as fuel decreases, the number of segments decreases. In contrast, we can imagine a fuel gauge that simply displays the number of gallons (or liters) of fuel in the tank (e.g., '6.5' for six and a half gallons). Here, as fuel decreases, the digits displayed do not increase or decrease (the way they would if we changed the font in a paper): they simply change. Again, the *interpretation* of those digits changes, of course: although the numeral '4' is neither greater nor less than '5', the *number* four that the numeral represents clearly *is* less than the number five. The important point here is that analog representations can be continuous or discrete, and that issue is completely separate from whether a representation is digital.

Now, we do not think—and certainly do not claim—that anyone currently has any idea of how brains accomplish the kind of analog processing necessary for combining *fundamental* microphysical-phenomenal magnitudes in a coherent way. Indeed, the relationship between neural processing and fundamental physics remains poorly understood. For example, although some have argued the human brain is not a quantum computer (Litt *et al.* 2016), recent studies have found quantum mechanics to be involved in basic elements of life such as photosynthesis (Thyrhaug *et al.* 2018), as well as in perceptual brain processing (Pereira & Furlan 2009). Further, recent work in physics has provided a possible mechanism for how the brain might exploit nuclear spins in phosphate ions to achieve quantum memory effects (Fisher 2015). However, beyond this and what is known about synapses and communication via electromagnetic fields, very little is known whether neural processing depends on the magnitudes of fundamental microphysical entities, such as the strong or weak nuclear force or the force of gravity. Finally, we also currently lack any real understanding (if micropsychism is true) of which phenomenal qualities (or their 'protophenomenal' precursors) inhere in which fundamental microphysical properties and magnitudes thereof.

However, let us set aside these issues for now. For the time being, we can summarize our view as follows. There are three separate things which appear to stand in a three-way analog relationship. First, there are physical stimuli that vary along certain dimensions (e.g., different wavelengths and amplitudes of light entering the eye). Second, there are neural representations of those stimuli, which also vary along certain dimensions (e.g., neural-spike shapes, firing rates, changes in EM fields around neural bodies, etc.). Third, there are phenomenal conscious experiences, which also vary along certain dimensions (e.g., hue, brightness, saturation, etc.). We reviewed evidence that the neural representations of physical stimuli are *analog* representations,

and then argued that phenomenal consciousness has a phenomenological structure that mirrors the structure of neural representations (which, in turn, has a structure that mirrors the physical stimuli). Whether it is correct to call all phenomenally conscious experiences *representations* (and if so, what to call them representations *of*) is contentious. For, as we saw earlier, macrophenomenal states often seem to represent things: my visual experience of a vase of flowers seems to represent those flowers to me. And, of course, there is a large literature arguing that phenomenal consciousness is *essentially* representational (see Lycan 2019 for an overview). However, for our purposes, these matters need not be settled. What is more important is the analog relationship between all three types of quantity involved: as a physical stimulus *literally* increases along some dimension (e.g., brightness in lumens), the relevant part of the neural representation of that stimulus *literally* increases (e.g., neural firing rate in some part of the visual cortex increases), and our phenomenal experience of that event *literally* increases along some dimension (e.g,. the experience of the light getting brighter). Finally, and crucially, micropsychism is unique among major theories of phenomenal consciousness in holding unambiguously that macrophenomenal coherence requires some kind of *micro*physical-phenomenal coherence. Allow us to explain.

First, Cartesian dualists obviously deny that microphysics has anything essential to do with phenomenal consciousness itself—as, for the Cartesian, consciousness is a *non*-physical substance capable of coherent macrophenomenal experiences as a disembodied soul.

Second, functionalists allow that *any* physical mechanism that carries out the right functions will be equally phenomenally conscious, irrespective of what is going on at a microphysical level. Now of course, it is possible, at least in principle, that the kinds of analog features of phenomenal consciousness we have discussed—such as subtle variations of

phenomenal color hues, saturation, etc.—must for some yet-to-be-understood reason be functionally realized in microphysics, such that *no* digital A.I. could carry out the right functions for generating coherent macroconsciousness—in which case our argument would extend to functionalism. Yet, although this is a possible form of functionalism, two points are crucial: (1) functionalists typically understand phenomenally conscious mental functions as being realized by macrophysical states (e.g. neurons or neural networks); and (2) there are no clear reasons (independent of panpsychism) to think that a digital computer couldn't simulate in a *macro*physical neural network whatever relevant analog functions human brains instantiate for generating coherent macroconsciousness. After all, whatever human brains do at a microphysical level, it seems possible (at least in principle) for a digital A.I. to simulate *that*—as (for example) videogame engines might do in simulating the behavior of subatomic particles, including their spin, charge, etc. So, even if human brains realize coherent macrophenomenal consciousness in virtue of carrying out relevant analog functions at a microphysical level, it is possible in principle—according to functionalism—for a digital computer to simulate (and hence, by the functionalist's lights, realize) the *same* functions (and hence, coherent macroconsciousness) in a macrophysical system (such as a neural network). It is only *micropsychism* which holds that consciousness fundamentally exists at the microlevel and must be built up (or combined) into a coherent macrophenomenal manifold. Consequently, functionalism provides no unambiguous grounds for thinking that macrophenomenal coherence requires any form of *micro*physical-phenomenal coherence—and hence, no reason to think that digital A.I. couldn't realize the kinds of analog physical-functional relationships necessary for macrophenomenal coherence.

Third, similar considerations apply to the mind/brain-identity theory of consciousness, the view that phenomenal consciousness is strictly identical to physical entities or processes. The

identity theory of consciousness can in principle take two forms: token-identity theory (according to which individual instances of phenomenal experience are identical to physical states) or type-identity theory (according to which *types* of phenomenal experiences are identical to *types* of physical states). Token-identity theory, however, is consistent with functionalism—for, as Smart (2017, §6) notes, functionalists can accept token identity claims about consciousness, allowing (e.g.) that a particular pain is identical to a particular physical process. What functionalists deny is that the *type* of physical process—e.g., whether a physical process is carbon- or silicon-based—has any bearing on consciousness. As such, only the *type*-identity theory is a genuine alternative to functionalism, because on this theory the *type* of physical constitution does makes a difference to consciousness irrespective of a system's functional characteristics (Smart 2017; see also Block 1978 and Searle 1984). But now, when it comes to type-identity theory, there are two possibilities: (A) that phenomenally conscious states are identical to types of *micro*physical states (such as the complete microphysical states of neurons), or alternatively, (B) phenomenally conscious states are identical to certain types of *macro*physical states (such as, to take a common example, 'C-fibers firing'). And indeed, Searle (1998, p. 1935) explicitly defends this latter form of identity theory, holding that although 'consciousness and indeed all mental phenomena are *caused by* lower-level neurobiological processes in the brain', his view is that 'consciousness and other mental phenomena are *higher level features of the brain*' (our italics). Notice, next, that it is entirely possible on this kind of *macrophysical*-type identity theory that digital A.I. might have relevant types of macrophysical states (e.g. a simulated 'C-fiber firing' in a neural network) that—even if the underlying processing is digital—are *identical* (at a macrolevel) not to our type human consciousness, but (perhaps) to some form of machine consciousness. So, for example, a digital computer can

clearly simulate, at a macrolevel, an analog clock, where the 'hands' are patterns of pixels on a screen, implemented by an underlying microphysical digital process. Now, there is a sense in which this is identical to a *type* of analog representation at a macro level of abstraction. Similarly, if phenomenally conscious states are identical to *types* of macrophysical states, and a digital A.I. can simulate *types* of macrostates similar to human neural states—including simulating their analog features at some level of abstraction (which as we have just seen digital machines can do)—then it is possible on identity theory that digital A.I. might simulate *types* of (macro)physical states (and analog relationships between them) identical not to human consciousness, but rather some distinct form of *machine consciousness*. The point then is this: the mind-brain identity theory provides no *unambiguous* reasons to think that microphysical states are the relevant type of state identical to phenomenal consciousness, as opposed to macrophysical state-types (such as a type of neuron firing, etc.).

Finally, as for cosmopsychists, they deny outright that, 'the structure of the macro level brain is derived from its structure at the micro level', as they take phenomenal consciousness to be macrophysical features of the Universe as a whole (Goff *et al.* 2020: Section 4.4.2. See also Nagasawa & Wager 2016).

Thus, while there are complex and interesting questions regarding whether digital AI could have coherent macrophenomenal experiences according to other theories of consciousness, none of these other theories *unambiguously* entail what we will now argue is true of micropsychism: namely, that if micropsychism is true, then digital AI may not be capable of realizing coherent macroconsciousness in any meaningful way because of how digital computation abstracts away from microphysical-phenomenal magnitudes.

3. If Micropsychism, Then Digital AI May Lack Phenomenal Coherence

We have seen that, in many cases, phenomenal consciousness involves *varying magnitudes* of different phenomenal qualities or dimensions (hue, brightness, loudness, and so on). We have also seen that the brain appears to be, at least in large part, an analog computer, carrying out its functions by manipulating different physical magnitudes (different levels of voltage in neural spikes, as well as electromagnetic fields) that represent the magnitudes of physical stimuli. We called the relationship among these three kinds of magnitudes—phenomenal qualities, physical stimuli, and neural representation—a kind of three-way analog representation.

If we are correct about this relationship, then for an intelligent agent to have a *coherent* set of phenomenal experiences—that is, a coherently-structured phenomenal visual field like our own (with coherent visual images rather than incoherent 'static'), coherent auditory phenomenal experiences (such as hearing the sound of a jet engine rather than auditory 'static'), coherent emotional experiences, conscious thoughts, and so on—then that agent's mind must *combine* microphenomenal magnitudes in a phenomenally coherent way, in much the way that a painter places color blotches on a canvas to paint a coherent picture of flowers. Again, we do not claim to know *how* this may happen in brains. We simply note that there is strong evidence to support this three-way analog relationship between stimulus, neural representation, and phenomenal experience.

However, if micropsychism is true, then the relevant phenomenal magnitudes exist as 'microphenomenal properties' at the level of *fundamental microphysics*. Consequently, if micropsychism is true, then the only way to achieve macrophenomenal coherence of the sort we experience—such as, again, the kind of coherent visual phenomenal experiences we have of seeing flowers in a vase, rather than experiencing incoherent visual 'static'—is for the

phenomenal features of the mind (i.e., the mechanisms that neurons comprise) to be realized in an analog mechanism (the human brain) that manipulates microphysical quantities (mass, charge, etc.) in a way that in turn brings together relevant phenomenal qualities (colors, etc.) in the *right way* (viz. realizing a coherent visual experience of flowers rather than 'static'). But this is simply not what digital computers do or can do, given the nature of digital representation.

To see why, consider the representations a digital computer processes. Suppose we were to program a digital computer to simulate a neural array with as much precision as one wants—as is often done in artificial intelligence with 'neural net' architectures. Although such a system might perfectly replicate the *functional* behavior of actual neurons at a software level, it would only do so through representations that are physically realized as a set of 'on' and 'off' electrical signals that do not vary in physical magnitude (beyond simply being 'on' or 'off'). In other words, the physical implementation of that neural network will only involve a large series of circuit elements that are in one of two physical states: five volts (standing for 'on' or '1') or zero volts (standing for 'off' or '0'). This is simply because the digital representations in which digital computers traffic abstract away from any *particular* physical implementation. In the case of a typical digital computer, it will be implemented via voltages; if implemented in some more exotic kind of digital computer, it will be implemented via water pressure, or even beer cans and string. In every case, the magnitudes of those physical values only represent individual numerals of a digital representation; magnitudes *themselves* are not physically represented at all. Once again, an analog representation of six would be a *physical magnitude* of six units (e.g., six volts, or six centimeters, or six neural firings per second), whereas a digital representation of six would be (in binary) a sequence of four voltages at zero, five, zero, and five volts (to represent 0101). The relationship between the magnitude represented and the physical magnitude(s) of the

representation are relatively arbitrary in the digital case, but quite constrained in the analog case.

Importantly, on the version of panpsychism we reference above, phenomenal qualities accrue to *magnitudes* of microphysical quantities. This means that a digital computer carrying out vastly many different tasks—simulating human vision, hearing, cognition, and so on—could only instantiate, *at most,* one variation along different dimensions of phenomenal properties, flashing, for example, 'on' and 'off' in a binary fashion, for every dimension of every sense modality simulated. Thus, in the case of a digital implementation, the aforementioned analog relationship simply does not hold. As physical stimuli change, representations change, but they do not do so in a physically, monotonically covarying way. If, as we suggest here, the physical magnitudes that do the representing in an analog representation are responsible for the changes in phenomenal magnitudes, then digital representations can only represent experiential magnitudes in a dramatically stunted manner.

Consequently, if panpsychism is true, then digital computers simply do not bear the right kind of relationship to fundamental physical-phenomenal magnitudes necessary for generating coherent macrophenomenal experiences. For example, even if they were able to digitally represent 'yellow bananas' in their visual field, the volume of a 'jet engine' in auditory inputs, and so on, digital computers would not—if panpsychism is true—actually have a phenomenal experience *as of* a yellow banana[6], the phenomenal auditory sound *as of* a jet engine, and so on.

[6] We want to distinguish here between having a phenomenal experience *of* a banana and having an experience *as of* a banana, where the difference is as follows. As Fisher (2009) contends, perhaps all that a physical, functional, or phenomenal state must do in order to represent (or be *of*) a banana is to be causally or functionally related to banana(s) in the external environment in some way. That may well be the case. Still, whether a phenomenal experience *resembles* an actual banana in a coherent fashion (viz. first-personally *looking like* or seeming *as of* a small yellow fruit) also seems relevant to representation: namely, for qualitatively representing the banana in consciousness *as it really is* (Summers & Arvan forthcoming, §3). Our point is that

Instead, they would represent such things, across all cognitive, emotional, and sense modalities (sight, hearing, touch, thought, etc.) using a single physical-phenomenal magnitude (voltage X instantiating phenomenal redness) alternating 'on' and 'off' (or one of two levels). As such, if panpsychism is true, then although digital AI might behave just like you or I—claiming to see visual objects, avoid and manipulate objects in their environment, cry out in pain, and so on—their first-personal phenomenal experience would be comprised by single physical-phenomenal magnitudes alternating on and off, having little or no correlation with what they represent (or claim to represent). What would it 'be like' to be a digital AI, then? Simple: it would be like a scrambled jumble of simple phenomenal states alternating on and off like static on an old television set (Biggs 2009 describes a similar case in a different context).

Now, to be sure, we reiterate that we do not know how the combination problem can be solved. However, given our other premises about the analog nature of neural representation (and the uncontroversial assumption that for micropsychists, what happens in our brains physically is implicated in what we experience phenomenally), it seems that consciousness like ours is such that an increase in some dimension of phenomenal experience (the loudness of a sound) correlates with a *literal* increase in the neural representation (i.e. an analog representation) of the physical loudness of that sound. As physical loudness increases, neurons fire faster (or generate larger spikes, or create strong fields), and we experience the sound getting louder. We do not know why *those* particular neurons correlate with the experience of sound intensity, rather than the brightness of a color. Perhaps future brain science will tell us what these correlations are—by, for example, correlating particular macrophenomenal experiences (e.g., experiences of

even if digital AI could represent bananas in Fisher's externalist sense—visually 'tracking' bananas in their environment—digital AI cannot do so in a manner that produces a coherent first-personal phenomenal experience *as of* yellow bananas.

redness) with particular microphysical magnitudes in neural processing (e.g., action-potential of shape X), etc. This might produce some picture of how brains combine fundamental microphysical-phenomenal magnitudes into coherent first-personal phenomenal experiences. Or, perhaps neuroscience will *never* be in a position to determine how fundamental phenomenal qualities inhere in fundamental physics due to the so-called 'explanatory gap' (McGinn 1989). We cannot resolve these issues, as they are open philosophical and empirical questions. Our point is simply that, if constitutive micropsychism is true, and given that we have good reason to suppose that the analog relationships we have articulated hold, then *somehow* our brains combine fundamental microphysical-phenomenal magnitudes in a way that generates coherent macrophenomenal states.

Before moving on, we should address two potential objections. We mentioned earlier that digital computers typically use binary representations of numbers. According to our argument, the physical implementation of binary representations implies that digital AI would have—at best—a kind of flickering of phenomenal consciousness. However, one might wonder, couldn't we address this by having AI systems that use different digital representations, so that they could have different levels of variation? A digital computer that used decimal representation would have ten different voltages in its physical implementation, for example, and that would seem to allow for ten variations for each dimension of phenomenal experience. If we used still higher representational bases, wouldn't the problem be solved?

While it is true that using larger bases would allow for more phenomenal 'levels,' the phenomenal experience would still not be correlated with what is being represented. Again, the digital representation and computation abstracts away from its physical implementation, using multiple numerals in different places to represent different parts of quantities. To illustrate,

suppose a digital computer uses a decimal base, representing numerals by voltages. To represent eight hundred twenty-eight, it represents the numerals '828' by a circuit element at eight volts, another at two volts, and another at eight volts. Importantly, the two instances of eight volts represent very different things at the level of the digital representation: one represents eight *hundreds*, and the other represents eight *ones*. However, given what we have argued, the two instances of eight volts would not combine in the right way, because they are not part of the same physical magnitude. Whatever is represented by eight hundred twenty-eight at the level of the digital program and whatever those three voltages give rise to phenomenally simply do not correlate in the right way.

It might be objected here, in turn, that a digital system can represent not only via base-2 or base-10 computation, but rather in a far more fine-grained way through *strings* of such representations—such as the string '10101011111' representing some specific hue of blue, with some amount of brightness, at some particular point in space. Why not think that digital computation *could* represent the apparent 'analog' features of phenomenal consciousness (different hues, brightness, etc.), though digital strings? This objection, however, is predicated upon a misunderstanding that we can reveal via the following dilemma. In a digital system, any string of representations (e.g. '10101011111') must occur either (1) *all at once* at a given time, but spread out over space (i.e. one node or processor representing '1', another representing '0', and so on, *all at the same time*); or (2) in a *sequence across time* (i.e. this one node representing '1' at time t, '0' at t_{+1}, and so on). The problem then is this: neither way of digitally representing a string resolves the problem we are presenting for coherent digital AI consciousness. On the first possibility—different distributed nodes representing '1' or '0' all at the same time—then all that a digital AI would represent at any given point in time would be whatever phenomenal

magnitude inheres in *those* physical quantities at different nodes at that particular instant. To see what we mean, suppose for the sake of argument that at a microphysical level, five volts instantiates *phenomenal redness* and zero volts instantiates *phenomenal greenness.* As we have seen, if constitutive micropsychism is true, then something like this may well be the case—as micropsychism takes phenomenal qualities to inhere in fundamental physics. Next, suppose that we program a set of digital nodes to represent the digital string '10101011111' all at once (i.e., at *t*). Here is the problem: if panpsychism is true, then *regardless* of what '10101011111' represents digitally, the analog physical features that realize this digital string at *t* (node 1 firing at five volts, node 2 at zero volts, node 3 firing at five volts, etc.) will at most realize an incoherent synchronic cluster of *red and green* phenomenal experiences at *t*. On the other hand, when we consider the second possibility—a digital processor representing each binary digit in the string *across* time ('1' then '0', then '1', etc.)—then once again all we would have is *phenomenal redness and greenness* flashing on and off, a kind of 'diachronic color static', again not presenting the digital AI with any coherent first-personal visual experience. The point here is this: neither form of digital processing—neither representing a complex digital string all at once, nor diachronically over time—can do what we are arguing must occur for phenomenal coherence to occur: namely, some kind of *microphysically* appropriate way of combining fundamental, analog physical magnitudes (i.e. magnitudes of voltage, spin, charge, etc.) in a manner that would render their inherent *phenomenal* features coherent at a macro level (viz. having a coherent first-personal phenomenal visual field, coherent auditory experiences, etc.).

This brings us to another potential objection, which is that it is a mistake to argue from premises about *human* consciousness—namely, that human consciousness is analog and depends

on analog features of physical brains—to conclusions about digital AI consciousness.[7] After all, our argument depends on the fact that digital AI are fundamentally different than human minds: specifically, in having a digital rather than analog architecture. However, if panpsychism is true, how can we be so sure that *digitally-based* minds are not conscious in at least *some* way similar to our own consciousness? For example, in the *Terminator* film series, the visual field of the eponymous robot is depicted as being grayscale (with a red overlay). For simplicity, suppose this system uses a base-4 digital scheme, where there are four gradations for each 'pixel.' Thus, each pixel of its visual experience needs just a single digit,[8] because each digit can take one of four values: one corresponding to white, one to bright red, one to dark red, and one to black. Why not think that this robot—this digital system—has visual phenomenal consciousness of the kind depicted in the films (which is similar enough to our own to at least be recognizable as *some* type of consciousness), despite not being an analog system?

The answer lies in the fact that digital representation abstracts from its microphysical implementation (as mentioned above), and as such, the same digital representations can be implemented in different kinds of (micro)physical media, and in two different ways. Consider again our base-4 example: we need four different voltage levels to represent four different digits. But which four voltages? Any four will work. They could be zero, one, two, and three; they could be negative-two, four, one, and seven; they could be six and a half, seven, negative-thirty, and thirty. It does not matter. All that matters is that there are four *different* voltage levels to represent the different digits. This may seem strange, but it is no different from the fact that the Arabic numerals we use to represent, say, the numbers one, two, three, and four have no

[7] Acknowledgement redacted for review.

[8] Whereas in a binary scheme, two digits would be needed to represent four different values.

systematic relationship among them: they are simply different from one another. In any case, from the perspective of the digital functioning of the system, all that matters is that four digits are represented: 0, 1, 2, and 3. Thus, qua *digital* system, *any* of these voltage-to-digit mappings would be possible (as well as infinitely many more). But if micropsychism is true, then any *particular* microphysical instantiation that happened to correspond to a coherent phenomenal experience would be completely a matter of chance.

What's worse is the second way in which the same digital representations can be implemented in different kinds of physical media. In our base-4 example, there is really nothing to prevent the digit of each pixel from being physically represented by different types of physical media. One pixel could be represented by different voltage levels, another by different temperatures, another by a gear being in four different positions. This would be a very complicated mechanism, but from the perspective of a digital system, these would all be equivalent. Although not quite this elaborate, real digital computers do use different physical quantities to represent bits: voltage levels in cache memory, for example; magnetic fields in hard drives; and physical divots in CDs and DVDs. However, once again, assuming micropsychism, only some values of some of these properties would be of the right type and value to be the physical basis for coherent macrophenomenal consciousness.

On top of both of these considerations, diachronic increases and decreases in a phenomenal quality would only correspond to a simultaneous increase or decrease in the microphysical medium implementing the representation of that quantity in an accidental manner. If the Terminator 'sees' a light that is literally getting brighter, it will have to digitally represent an increasing number corresponding to that brightness. In our example where only four values are possible, that would go from the digit 0, to 1, to 2, and to 3. But as we saw, those digits could

be represented by, say, four volts, negative three volts, one volt, and two hundred volts, respectively. Another functionally-identical digital Terminator could use a different set of voltages. But by hypothesis, it is the analog relationship between the magnitude of physical stimuli, the magnitude of their microphysical (neural) representation, and the magnitude of their phenomenal qualities that holds. If the magnitude of the microphysical representation does not have the right monotonic increase or decrease with the physical stimuli, the analog relationship no longer holds, and all bets are off. The view that each digitally-implemented Terminator has the same kind of phenomenal consciousness—or a consciousness anything like our own—is possible, but not compatible with the theses we advance here.

Taken all together, we believe these three points establish that although it could be *metaphysically* possible to program digital AI to have phenomenally conscious states (though it is hard to know, given our epistemic situation), our current lack of understanding of the fundamental relationships between microphysical and phenomenal states means that, at least for now and for the foreseeable future, any attempt to program digital AI with coherent phenomenal consciousness might succeed only by *mere, completely random chance*. In fact, it seems that we would have to restructure a digitally implemented AI so that it is actually analog in nature. Interestingly, at the level of the external, observable behavior and function of such an AI, no difference would be detectable if were changed from being digital to analog. But in terms of the physical and representational architecture, the substrate upon which such a system was implemented would go from being unable to support coherent, phenomenal consciousness to one that does. We thus submit that if constitutive micropsychism is true, then digital AI—at least now and for the foreseeable future—may lack coherent phenomenally conscious experiences.

4. New Questions for the IIT Theory of Consciousness

Although it has its share of critics, one influential theory of consciousness that has emerged in recent years is Tononi's (2011, 2012) Integrated Information Theory (IIT) theory—which holds that the 'level of consciousness' a system has can be best understood in terms of the degree by which that system *integrates information*. Tononi has formalized the level of informational integration a system has as a measure 'ϕ', arguing that his account explains an array of empirical findings about consciousness.

We need not concern ourselves with many details of Tononi's account here. The relevant point for our purposes is simply that our argument raises new *prima facie* questions about the motivating idea behind IIT: namely, the claim that consciousness is simply a matter of information integration. Tononi bases IIT on five axioms:

1. *Intrinsic experience*: consciousness exists and has an 'intrinsic perspective.'
2. *Composition*: consciousness is structured ('within one experience, I may distinguish a book, a blue color, a blue book', etc.).
3. *Information*: consciousness is composed of 'phenomenal distinctions' that distinguish conscious experiences from one another (e.g., a conscious experience of a *bedroom* has different phenomenal informational content than a conscious experience of *a bird,* etc.).
4. *Integration*: consciousness is unified ('I see a whole visual scene, not just the left side of the visual field independent of the right side', etc.)
5. *Exclusion*: 'Consciousness is *definite,* in content and spatio-temporal grain.'

(Tononi 2015, 'Axioms: Essential properties of experience')

Tononi then argues from the exclusion axiom to a coarse-graining principle as a further *postulate* of IIT:

> [T]he exclusion postulate also applies to spatio-temporal grains . . . This means that, if cause-effect power is more irreducible than at a finer grain then, from the intrinsic perspective of the system, the coarser grain excludes the finer one.
>
> (Ibid., 'Postulates: Properties required of the physical substrate of experience

This is crucial for the following reason. On the one hand, Tononi and Koch (2015) write that although 'IIT was not developed with panpsychism in mind (sic)', it is nevertheless the case that, 'in line with the central intuitions of panpsychism, IIT treats consciousness as an intrinsic, fundamental property of reality' (p. 11). However, on the other hand, Tononi is also clear that the coarse-grained principle entails that microphysics is in principle *entirely irrelevant* to macroconsciousness:

> If such a local maximum of integrated information is indeed identical with consciousness, as claimed by IIT, it follows that a set of mechanisms in a state capable of generating consciousness also constitutes a *local maximum of causal power* . . . Any lesser cause (one that is less irreducible), including microlevel causes (the molecules in my brain), or proximal causes (the muscles in the finger) are excluded . . . the macro level *supersedes* the micro level . . .
>
> (Tononi 2012, p. 309)

Tononi adds:

> According to IIT, the neural states that are important for consciousness are only those that have maximum cause–effect power on the system itself. For example, assume that, from the intrinsic perspective of the system, maximum cause–effect power was achieved when coarse-graining firing states . . . In this case, IIT predicts that finer grained neural states, despite their demonstrable neurophysiological effects, *make no difference* to the

content of experience.

(Tononi *et al.* 2016: p. 453; italics added. See also Hoel *et al.* 2013)

As Tononi and Koch (2015, p. 16, fn. 15) note, these core assumptions of IIT—the coarse-grained principle, which again is derived from the exclusion axiom—suggests that digital AI with 'neuromorphic hardware', or hardware capable of integrating information at a macrolevel like human brains, could very well have phenomenal consciousness like ours. However, if our argument is sound, then these core assumptions of IIT may be misguided. If micropsychism is true, then—contrary to IIT's coarse-grained principle—coherent macroconsciousness requires integrating *micro*physical-phenomenal information in the *right kind of analog manner*—manipulating fundamental microphysical magnitudes (mass, charge, spin, etc.) in ways that combine microphenomenal properties into a coherent macrophenomenal whole. Given that, as noted above, Tononi holds that 'in line with the central intuitions of panpsychism, IIT treats consciousness as an intrinsic, fundamental property of reality', our argument thus reveals two important things about IIT's relationship to panpsychism: namely, (i) if *micro*psychism is true, then IIT is false (since micropsychism denies the coarse-grained postulate of the exclusion axiom), and (ii) IIT is at most consistent with versions of panpsychism which deny that microphysical phenomena constitute macroconsciousness. Now, such a version of panpsychism does exist: namely, cosmopsychism—the view that consciousness is fundamentally grounded in the Universe as a whole, and the conscious experiences of creatures such as us are derived not from microphysics but from the macro-Universe (Goff 2017). However, while cosmopsychism has some proponents, many take micropsychism to be the most plausible form of panpsychism.

Notice, finally, that even if this is right—that is, even if micropsychism undermines IIT's coarse-grained principle (and hence, Tononi's exclusion axiom)—there may still be a broader

sense in which macrophenomenal consciousness is constituted by a kind of information integration: namely, by *macro*coherent integration of *analog, microphysical-phenomenal information*. However, if this is correct, then it may be that the only way to integrate microphysical-phenomenal information in a macrocoherent way is via *precisely* the physical stuff and processes utilized by carbon-based human and animal brains. That is, it could be that attempting to realize coherent macrophenomenal consciousness in any other physical materials or processes would result in an incoherent jumble of multimodal phenomenal states (i.e. a 'senseless mishmash' of auditory, visual, kinesthetic, olfactory, and other phenomenal states). We believe these to be questions to which the science and philosophy of consciousness currently have no good answers. To settle these questions—that is, to know whether, if panpsychism is true, diverse physical mechanisms (such as silicon processors) may be capable of generating coherent macroconsciousness—we would plausibly need a fine-grained understanding of how fundamental microphysical and phenomenal properties relate (i.e. which microphysical phenomena realize redness, greenness, etc.). For, without this kind of knowledge, all we would know is that carbon-based human brains *somehow* integrate fundamental microphysical-phenomenal information to generate coherent macroconsciousness. To know whether silicon-based systems could do so as well, we would need to know how fundamental microphysical and microphenomenal properties relate. Is it possible that a future science of consciousness might achieve such knowledge—mapping specific microphenomenal states (e.g. phenomenal redness) to specific microphysical states (i.e. 5-volts)—such that scientists might discern precisely how brains integrate *analog* microphysical-phenomenal information to generate macrophenomenal coherence? Might some such knowledge in turn lead us to ascertain whether coherent macrophenomenology is multiply realizable in different, physically realistic mechanisms—for

example, in *analog* (rather than digital) silicon-based processors? Again, we cannot resolve these issues. Our argument, however, raises them as crucial questions for future research in the science of consciousness.

5. Conclusion

This paper has shown that there are reasons to believe that human brains may function in an analog manner, and that phenomenal consciousness may be analog in nature, as well. We then argued that, because micropsychism holds that microphenomenal properties inhere in the universe at a fundamental level—somehow 'binding' to the world at the level of fundamental microphysics—then, if micropsychism is true, beings like you and I have *coherent* phenomenal experiences (coherent phenomenal visual fields, coherent auditory and tactile experiences, etc.) in virtue of our brains *manipulating* microphysical properties in some kind of *analog* fashion. Finally, we argued that insofar as digital computation does *not* manipulate microphysical properties in an analog fashion, it follows (if the rest of our assumptions are right) that digital AI are incapable of instantiating coherent macrophenomenal experiences. Thus, our argument shows that if panpsychism is true, then digital AI *may* very well be phenomenally scrambled. While this is not a proof that digital AI *are* phenomenally scrambled—as there are outstanding questions here, ranging from whether micropsychism is the best form of panpsychism to debates about the digital-analog distinction—our findings are important. Are human brains truly analog? Is phenomenal consciousness? If so, how exactly *do* brains integrate analog microphysical-phenomenal properties to achieve macrophenomenal coherence? We cannot definitively settle these questions. However, we have defended some provisional answers, and believe that we have shown that these are vital questions that future work in the science and philosophy of consciousness should investigate further.

References

Adams, Zed (2019). The history and philosophical significance of the analog/digital distinction. *Pitt Center for Philosophy of Science*. Pitt Center for Philosophy of Science Colloquium, available at https://sopha2018.sciencesconf.org/189501/document, retrieved 27 July, 2021.

Beck, Jacob (2019). Perception is Analog: The Argument from Weber's Law. *The Journal of Philosophy* 116(6): 319-49.

Beck, Jacob (2015). Analogue Magnitude Representations: A Philosophical Introduction 66: 829-55.

Biggs, Stephen (2009). The scrambler: An argument against representationalism. *Canadian Journal of Philosophy* 39(2): 215-36.

Block, Ned (1978). Troubles with functionalism. *Minnesota Studies in the Philosophy of Science* 9: 261-325.

Chalmers, David J. (2016). The Combination Problem for Panpsychism. In Godehard Brüntrup & Ludwig Jaskolla (eds.), *Panpsychism*. New York: Oxford University Press, 19-47.

Chalmers, David J. (2013). Panpsychism and Panprotopsychism. *Amherst Lecture in Philosophy*.

Chalmers, David J. (1996). *The conscious mind: In search of a fundamental theory*. Oxford university press.

Chiang, Chia-Chu; Shivacharan, Rajat S.; Wei, Xile; Gonzales-Reyes, Luis E.; Durand, Dominique M. (2018). Slow Periodic Activity in the Longitudinal Hippocampal Slice Can Self-Propagate Non-Synaptically by a Mechanism Consistent with Ephatic Coupling. *The Journal of Physiology* 597(1).

Chrisomalis, S. (2020). *Reckonings: Numerals, Cognition, and History*. Cambridge, MA: MIT Press.

Dennett, Daniel C. (1996). *Kind of Minds*: *Towards and Understanding of Consciousness*. New York: Basic Books.

Dennett, Daniel C. (1993). *Consciousness Explained*. Boston: Back Bay Books.

Fisher, Justin C. (2007). Why nothing mental is just in the head. *Noûs* 41(2): 318-34.

Fisher, Matthew P.A. (2015). Quantum cognition: the possibility of processing with nuclear spins in the brain. *Annals of Physics*, 362: 593-602.

Goff, Philip (2017). *Consciousness and Fundamental Reality*. New York, Oxford University Press.

Goff, Philip; Seager, William; and Allen-Hermanson, Sean (2020). Panpsychism. *The Stanford Encyclopedia of Philosophy* (Summer 2020 Edition), E.N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/sum2020/entries/panpsychism/>.

Goodman, Nelson (1968). *Languages of Art: An Approach to a Theory of Symbols*. Indianapolis: Hackett Publishing Company, Inc.

Hoel, Erik P.; Albantakis, Larissa; and Tononi, Giulio (2013). Quantifying causal emergence shows that macro can beat micro. *Proceedings of the National Academy of Sciences*: 110(49), 19790-19795.

Katz, Matthew (2008). Analog and digital representation. *Minds and Machines* 18(3): 403-8.

Kim, Jaegwon (1992). Multiple realization and the metaphysics of reduction. *Philosophy and phenomenological research*, 52(1): 1-26.

Koch, Christof (2019). *The Feeling of Life Itself: Why Consciousness is Widespread but can't be Computed.* Cambridge, MA: The MIT Press.

Kulvicki, John (2015). Analog Representation and the Parts Principle. *Review of Philosophy and Psychology* 6(1): 165-80.

Lewis, David K. (1971). Analog and digital. *Noûs* 5(3): 321–27.

Maley, Corey J. (forthcoming). Analog Computation and Representation. *British Journal for the Philosophy of Science*. https://www.journals.uchicago.edu/doi/10.1086/715031.

Maley, Corey J. (2020). Continuous Neural Spikes and Information Theory. *Review of Philosophy and Psychology* 11: 647-67.

Maley, Corey J. (2018a). Brains as Analog Computers, *Medium*, https://medium.com/the-spike/brains-as-analog-computers-fa297021f935, accessed on March 12, 2019.

Maley, Corey J. (2018b). Toward Analog Neural Computation. *Minds and Machines* 28 (1): 77-91.

Maley, Corey J. (2011). Analog and Digital, Continuous and Discrete. *Philosophical Studies* 155(1): 117-31.

McGinn, Colin (1989). Can We Solve the Mind-Body Problem? *Mind* 98/391: 349–66.

Mørch, Hedda Hassel (2014). *Panpsychism and Causation: A New Argument and a Solution to the Combination Problem*. Dissertation, Oslo.

Nagasawa, Yujin and Wager, Khai (2016). Panpsychism and Priority Cosmopsychism. In Godehard Brüntrup & Ludwig Jaskolla (eds.), *Panpsychism: Contemporary Perspectives*. New York: Oxford University Press: 113-29.

Nagel, Thomas (1974). What is it like to be a bat?. *The Philosophical Review*, 83(4): 435-450.

Oizumi, Masafumi; Albantakis, Larissa; and Tononi, Giulio (2014). From the phenomenology to the mechanisms of consciousness: integrated information theory 3.0. *PLoS computational biology*, 10(5), e1003588: 1-25.

Peacocke, Christopher (2019). *The Primacy of Metaphysics*. Oxford: Oxford University Press.

Pereira Jr., Alfredo & Furlan, Fábio Augusto (2009). On the role of synchrony for neuron–

astrocyte interactions and perceptual conscious processing. *Journal of biological physics*, *35*(4): 465-480.

Russell, Bertrand (1921). *The Analysis of Mind*. London: George Allen and Unwin.

Russell, Bertrand (1927). *The Analysis of Matter*. London: George Allen and Unwin.

Schwitzgebel, Eric (2006). The unreliability of naive introspection. *Philosophical Review*, 117(2):245-273.

Searle, John (2010). Why Dualism (and Materialism) Fail to Account for Consciousness' in Lee, R.E. (ed.) *Questioning Nineteenth Century Assumptions about Knowledge, III: Dualism.* NY: SUNY Press, 5-30.

Searle, John R. (1998). How to study consciousness scientifically. *Philosophical Transactions of the Royal Society of London. Series B: Biological Sciences*, 353(1377): 1935-1942.

Searle, John R. (1984). *Minds, Brains and Science*. Cambridge, MA: Harvard University Press.

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

Smart, J. J. C. (2017). The Mind/Brain Identity Theory. *The Stanford Encyclopedia of Philosophy* (Spring 2017 Edition), Edward N. Zalta (ed.), URL = <https://plato.stanford.edu/archives/spr2017/entries/mind-identity/>.

Strawson, Galen (2016). Mind and Being: The Primacy of Panpsychism. In Godehard Brüntrup & Ludwig Jaskolla (eds.), *Panpsychism: Contemporary Perspectives*. New York: Oxford University Press, 75-112.

Strawson, Galen (2006). Realistic monism: Why physicalism entails panpsychism. *Journal of Consciousness Studies*, 13(10-11): 3-31.

Summers, Micah & Arvan, Marcus (forthcoming). Two New Doubts about Simulation Hypotheses. *Australasian Journal of Philosophy*.

https://doi.org/10.1080/00048402.2021.1913621.

Thyrhaug, Erling; Tempelaar, Roel; Alcocer, Marcelo J.P.; Žídek, Karel; Bína, David … & Zigmantas, Donatas (2018). Identification and characterization of diverse coherences in the Fenna-Matthews-Olson complex. *Nature chemistry* 10(7): 780-86.

Tononi, Giulio (2015). Integrated information theory. *Scholarpedia: the peer-reviewed open-access encyclopedia*, http://www.scholarpedia.org/article/Integrated_information_theory.

Tononi, Giulio (2012). The integrated information theory of consciousness: an updated account. *Archives italiennes de biologie*, 150(2/3): 56-90.

Tononi, Giulio (2011). The Integrated Information Theory of Consciousness: An Updated Account, *Archives italiennes de biologie* 105(2/3): 56-90.

Tononi, Giulio; Boly, Melanie; Massimini, Marcello and Koch, Christof (2016). Integrated information theory: from consciousness to its physical substrate. *Nature Reviews Neuroscience*, 17(7), 450-461.

Tononi, Giulio and Koch, Christof (2015). Consciousness: here, there and everywhere?. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 370(1668), 20140167: 1-18.

Turing, Alan (1950). Computing Machinery and Intelligence, *Mind* LIX(236): 433-60.

Zbili, Mickael & Debanne, Dominique (2019). Past and Future of Analog-Digital Modulation of Synaptic Transmission. *Frontiers in Cellular Neuroscience*, *13*, 1290. https://doi.org/10.3389/fncel.2019.00160.