Implementational Considerations for Digital Consciousness

Derek Shiller*

July 30, 2023

1 Introduction

Recent advances in artificial intelligence have bolstered worries about the threats posed by misaligned AI systems. Further advances will likely introduce new opportunities for value and new causes for concern. Artificial intelligence will transform the world. Despite the unpredictability of the future, we are likely to fare better if we have thought about potential dangers in advance. Compared to the immediate threat of misaligned AI, navigating the speculative risks and rewards of a world with aligned AI has received little attention. The potential for artificial consciousness and sentience is of particular importance. Given the historical pace of consciousness research, it would be better to start thinking about it long before it arrives.

This document focuses on an underexplored issue relevant to the assessment of consciousness in artificial systems. I will explore some technical considerations relating to the capacity of contemporary computer hardware and software to give rise to consciousness. The strongest case for artificial consciousness relies on an organizational ('functionalist') view of the physical basis of consciousness. Such interpretations need to be supplemented with constraints on what it takes to implement the right organizations. Computers might implement the right organizations for consciousness but do so in the wrong way.

I will propose certain aspects of current hardware and programming paradigms that might prevent certain artificial systems from being conscious even if they implement the right functional organizations. These aspects are tied to specific

*This paper is indebted to many people who have discussed these ideas with me and given me comments on earlier versions. Thanks particularly to Brad Saad, Jonathan Simon, Carl Shulman, Robert Long, and the FHI digital minds reading group for discussion, inspiration, and criticism.

technical choices; they don't challenge the possibility of artificial consciousness in general.

Roadmap

This introductory section provides the motivation and philosophical foundations for the document. I survey the significance of digital consciousness and discuss the value of examining present-day implementations.

In **section two**, I'll explain functionalism and situate it in with the other major theories regarding the fundamental nature of consciousness. Functionalism is important because it remains the most clear theory to make clear predictions about digital consciousness. We are likely to need a theory in order to produce verdicts about the status of arbitrary systems.

In **section three**, I'll explore the question of what it takes for a system to implement a functional organization and describe the approach that I prefer, the Satisfaction Account.

In **section four**, I'll present a series of problem cases for the Satisfaction Account.

In **section five**, I'll survey and discuss existing proposals for constraints on implementation. I draw a lesson from these proposals that implementation requires some form of integrity in the parts of the system that play functional roles.

In **section six**, I'll present a variety of plausible constraints on the proper implementation of functional organizations that are of special interest because of how they relate to current computing paradigms. These constraints suggest that present computer architectures are not suitable for building unambiguously conscious systems.

In **section seven**, I'll draw some lessons from the plausible constraints for how we should think about digital consciousness.

Why digital consciousness matters

Digital consciousness is consciousness implemented with inorganic digital[1] computational devices such as computers built from electrical circuits. It warrants attention because the capacity for conscious experiences is important to an

[1] Digital computational systems involve symbolic representations of values. One alternative form of computation is analog, where values are stored as natural magnitudes. There are some potential considerations relating to consciousness in analog and digital computational systems (Arvan and Maley 2022). In general, in this document, I use 'digital' to refer broadly to the style of computation performed on computers. This involves not just symbolic representations of values but also includes other paradigms, such as the von Neumann architecture. It is possible that some forms of inorganic artificial consciousness will not use digital computations.

entity's moral status. In the long run, most of the potential for future wellbeing and illbeing may belong to digital minds.

Theories of value typically place great weight on welfare. How well off an individual is is often thought to be largely determined by their conscious mental states[2]. A system's welfare is increased by positively valenced experiences and decreased by negatively valenced experiences.

Digital minds will likely matter more if they have conscious experiences. If we create artificial systems with a plausible claim to consciousness, we will need to decide what we owe to them. This might involve deciding how much effort to make to design them in ways so that their lives go well. It might involve deciding how much server space to provide, when and how to deactivate them, and what kinds of political power they deserve. These decisions depend on how our own sacrifices affect their wellbeing.

The importance of digital consciousness is supported by two considerations. There may be many more digital minds than biological minds in the future and they may have a higher capacity for welfare.

Since digital systems don't rely on biological sources of energy they may be created quickly and maintained relatively efficiently[3]. It is conceivable that systems could be maintained in tremendous numbers without the psychological downsides of crowding animals or the costs of high biological populations on environmental degradation.

There need not be many people who are motivated to create digital minds in order for them to proliferate. The number of children that parents have is limited by the human birth rate and the number of years of reproductive fitness. Pet hoarders are limited to the number of animals that can be supported by their resources (time, money, property). Neither sort of limitation will obviously constrain digital minds to small numbers. Individuals, benevolent or malevolent, could conceivably support vast numbers of digital minds.

[2] There are dissenters (Carruthers 2020; Kammerer 2022) who think that unconscious mental states may matter. For instance, non-experienced forms of aversion may matter for the same reasons that experienced pain does. This view can be motivated by the concern that the sorts ofrequirements necessary to turn sub-conscious representations into phenomenal experience don't seem to be the right sort to justify a radical change in moral status (Though, see also Cutter 2017, for whom this modus ponens is a modus tolens). If the whether a nociceptive event is conscious pain depends on whether it is broadcast to the language module, why should that matter? This is important, but it would take me too far afield to explore the significance of this view here.

[3] Carl Shulman points out that computational costs are not presently trivial, so it is plausible that future systems will be limited to reasonable numbers by their energy requirements. This is right. Still, it is not obvious that present costs are a good guide to future costs. Computer engineering has been focused on increasing computation speed more than energy efficiency, and there may be large possible gains in that domain (Davies 2018). Regardless, if the main constraint on the number of digital minds is energy costs, it is plausible that there could be far larger numbers of sentient digital minds than sentient biological organisms (Greaves & MacAskill 2021).

Since we have control over how their minds are programmed, we may be able to give them ideal lives. We can plausibly design digital minds to never suffer and to have pleasant experiences of varieties whose contours we cannot imagine. We can build them to be true and fast friends to us and to one another. To fall in love and never stray. To appreciate all forms of art, to delight in learning, to react stoically to adversity. They need never grow bored, jaded, or sick. They can potentially live until long after the stars burn out. The potential wellbeing for even a single well-designed digital life is vast. The potential illbeing of such an entity in the hands of a sadistic or malevolent being is also vast.

One critical issue is ambiguity (Schwitzgebel and Garza 2015). If we create systems whose capacity for consciousness is unclear, we will be pressured to either treat them as conscious, treat them as unconscious, or to adopt some middle ground. If systems are unconscious and we treat them as conscious, we may needlessly waste resources on them. If they are conscious but we treat them as unconscious, we may inadvertently produce illbeing, or miss opportunities to produce wellbeing. On the other hand, if we opt for some middle ground in our treatment that matches our uncertainty, we should expect to either not give them all they are truly owed or to give them too much.

Ambiguity can be avoided by designing systems with attention to the possible sources of doubt. We know that human minds are conscious. Various theories posit that different aspects of human minds matter. If we can build artificial systems that share either most or few of these potentially significant aspects, the degree of uncertainty we have about them will be low, and the dangers posed by ambiguity will be small. This requires mapping out what those potentially significant aspects might be.

The importance of implementation

Consciousness has been studied by people working in the academic disciplines of philosophy, psychology, and neuroscience. Researchers have been primarily concerned with explaining the source of conscious states in humans. The obvious contrast case is unconscious human mental states. It is harder to know what to make of animal minds because they cannot give clear testimony about their experiences.

The plausibility of digital minds depends largely on the appeal of functionalism, broadly construed. Functionalists believe that consciousness is a result of the organization of our brains, not the specific materials they are made out of. Functionalism implies a strong version of multiple realizability: any system that implements the same organization as a conscious human brain would likewise be conscious.

Functionalism makes multiple realizability possible in theory, but even among functionalists, there is no guarantee that computers could be conscious. Some philosophers have suggested that there are properties of biological systems

(Godfrey-Smith 2016) that make them uniquely able to implement the complex functional architectures that underlie consciousness, so there are no non-biological systems that could implement the same organization of a human brain. Or perhaps the required functional patterns are so closely tied to the fine-grained behavior of the molecules composing neurons that they are impossible to implement in electrical circuits. Still, the stunning capacities of modern neural networks (Bubeck et al. 2023) provide strong evidence that circuits don't present fundamental technical constraints. It is conceivable that we can't artificially implement the organization necessary for consciousness, but the present evidence for this conclusion is rather limited.

In contrast, technical issues specific to modern computer architectures have not been explored in depth. If there are implementational details that bear on how we should think about artificial consciousness in relation to functionalism, they may only appear on close examination, once potential organizations are considered.

2. Functionalism and Other Theories

Consciousness is a challenging topic; it is not just controversial which theories are best supported, there is also substantial disagreement over what evidence we have and over what theories are a priori plausible. It would be convenient if we could assess consciousness in artificial systems without a theory about what consciousness fundamentally is. This may be possible for non-human animals, where behavioral or neurological heuristics may suggest that the same thing is occurring in them that occurs in us. It is less plausible for computers.

Behavioral and neurological heuristics can't be straightforwardly applied to digital systems. Behavior may mislead us about the inner life of creatures built specifically to mimic us. We cannot rely on the testimony of large language models about how they feel (Long 2023). We need a theory to tell us if the states that generate similar behaviors in artificial systems are consciously experienced. Neurological heuristics are also problematic. Neurological similarity may indicate consciousness in some animals, but neurological dissimilarity provides less clear evidence against consciousness. Since digital systems will be dissimilar from us in many ways, we need a theory to tell us which ways are important (Chalmers 2023).

Many present-day theories of consciousness appear to make predictions about consciousness in artificial systems. While there is disagreement about which theory is correct, there is also a reasonably broad consensus about the sorts of features that might be necessary. We might produce unambiguously conscious or unconscious systems by building systems that satisfy all or none of the popular criteria.

However, many of the details of the present-day theories remain to be specified. Theories have been developed with a focus on humans and can't readily be

applied to arbitrary systems. Deciding whether and how to extend theories to other sorts of systems will require some thought about what consciousness is at the most fundamental metaphysical level.

In this section, I distinguish the most popular category of physicalist theory of consciousness, which I'll call 'narrow functionalism', from its primary realist competitors, dualism and type identity theory. Narrow functionalism is interesting not just because of its popularity, but because it provides the clearest case for artificial consciousness.

Physicalism and Dualism

The broadest questions debated by philosophers of mind relate to the basic relationship between consciousness and the physical world. Consciousness consists of objects (minds) and properties (experiences, what it feels like to perceive things, qualia). Is consciousness simply a complex phenomenon within the physical world, as we've discovered biological life to be, or is there something fundamentally distinctive about consciousness, something over and above the physical stuff that gives rise to it?

Physicalists believe that consciousness is just a physical phenomenon, nothing more and nothing less. Dualists believe consciousness essentially involves objects or properties that are different from the kind of thing discussed in the physical sciences. For dualists, consciousness relates to underlying neural states in a different sort of way than tornados relate to the underlying atmospheric particles. Consciousness doesn't simply involve different sorts of patterns of physical objects, those patterns are responsible for producing novel aspects of reality: phenomenal experiences. Dualism has a historical connection to religious views, but atheists shouldn't dismiss it for this reason. The basic ideas appeal even to many people who intimately understand the brain and think consciousness is a purely natural phenomenon that is intimately tied to its operations[4].

The precise difference between physicalism and dualism can be hard to pin down. Some philosophers claim that the properties they regard as specially distinctive of consciousness are just the subjective side of the properties that physicists study. Once we allow the same property to appear under different guises, it becomes less clear how to understand the distinctive novelty postulated by dualists. It is also hard to say what counts as physical without a complete theory of physics (Ney 2008).

My favored interpretation of physicalism about consciousness says that anything that is distinctive about consciousness (and not just special about the way

[4] My impression is that dualism is too unpopular of a view to be endorsed outright by many neuroscientists, but the idea that there is some fundamentally mysterious gap between brain and mind is suggestive of dualism. There are ways of making sense of the gap as primarily a result of a deficit in ourconcepts (e.g. Loar 1990; Balog 2012), but it is hard for me to see how to make the gap an issue of the phenomena rather than how we think about the phenomena without leaning into dualism.

we think about consciousness) must be comprehensible in terms of physical phenomena under physical guises. If we have special insight into our conscious states, such that e.g. we can't be wrong about whether we're in pain, that fact must be entirely explicable in terms of physical attributes of our brain – that is in terms of how the physical structure of beliefs and the physical structure of experiences imply that such beliefs can't be inaccurate. If we think that consciousness matters morally in ways that other things do not, that must be explicable in terms of why the physical phenomena matters or it must not be explicable at all. For instance, if to be in pain is just to have an internal representation of tissue damage, then the badness of pain must be explicable as the badness of representing tissue damage or not be the kind of thing that can be explained.

The debate between dualists and physicalists is somewhat less relevant than it might first seem to assessing artificial consciousness. Many researchers agree that there are physical patterns that are correlated with conscious states whether or not dualism is correct. These physical patterns make consciousness happen, but they don't constitute it. Theories that were first developed to make sense of physicalism can be applied by dualists as a theory of the physical processes that coincide with consciousness(Chalmers 1997). For each physicalist theory, there is a dualist version of the theory that holds that the theory doesn't tell us the essence of consciousness, but locates it in all the right places. These theories postulate psycho-physical laws that dictate that whenever a specific physical pattern is implemented, a non-physical phenomenal experience of a particular sort is brought into being. Dualism may still make an important difference[5]: dualist versions of functionalism require functional-phenomenal correlations and those may be less plausible than biological-psychological correlations. Or functional-phenomenal correlations may be open to different constraints on which media properly implement the functional roles. But if there is strong support for a functionalist basis of consciousness, it doesn't show that dualism is wrong.

Type Identity Theory and Biological Requirements

If physicalism is true, then consciousness must be a specific[6] pattern of physical processes. Despite physicalism's long history, it wasn't until the mid-20th century that philosophers started arguing about exactly which physical patterns to identify it with.

The most prominent physicalist alternative to functionalism is type identity theory (Smart 1959), which asserts that kinds of conscious experiences (feeling pain, etc.) are identical with kinds of activity in certain kinds of neural circuits. To experience any particular conscious state is nothing more than having a

[5] Thanks to Brad Saad for pointing this out.

[6] Or perhaps there is not a specific pattern and consciousness is fundamentally indeterminate (Carruthers 2019). While my own sympathies lie with this view, I'll mostly set it aside because it is not mainstream.

certain sort of neurological circuit in one's brain that displays a specific pattern of activity.

For type identity theorists, the structures identified with conscious states are characterized by the attributes that neuroscientists care about: neural connections, neuron structure, neurochemistry, etc. It is referred to as 'type identity theory' because it identifies types of experiential states with types of neural states.

Type identity theory was briefly popular in the mid-20th century. It has spent most of its life as a foil for functionalism. It has retained some sympathizers, particularly among those who think that there is something special about biological life that allows for consciousness. However, the biology-is-special crowd[7] rarely embrace type identity theory explicitly, and the sentiment that biology performs some service necessary for producing consciousness may be best fit under a more dualistic theory, as it suggests that consciousness arises out of and is logically distinct from the substrate that produces it.

One challenge for type identity theory is that it has not borne fruit: there are few serious and credible identifications of neural circuitry with conscious experiences. The neural correlates of consciousness research program, which seeks to find the neural activities that coincide with human conscious experience, has led to proposals about which neurological mechanisms might be responsible for consciousness, but these proposals are better not interpreted as straightforward and complete identifications.

In early work, type identity theorists presented an illustrative proxy for an identification: pain is the firing of c-fibers. This identification wasn't to be taken seriously. Pain itself is typically thought to arise with activity in certain brain regions. While it may be associated with those regions, we are far from locating detailed descriptions of the activities of the sort we might expect from type identity theory.

A second problem is that type identity theory has strong implications for the distribution of consciousness. It is unlikely that we will find specific kinds of activity that correspond to psychological states of pain we think exist across the animal kingdom. Human brains are themselves sufficiently different from the brains of other animals that even if we could identify specific circuits involved in our experiences, it is unlikely that we would find the same sorts of circuits involved in the experiences of dogs, or fish, or octopi. The challenge is that the neural systems that underlie different phenomenal states are likely to share far more with each other within a species than the systems that underlie the same phenomenal states across species. Insofar as we don't want our theory of the basic nature of consciousness to rule out shared conscious experiences in such animals from the start – and most people think that would be overly chauvinistic – type identity theory struggles.

[7] Most prominently, Ned Block (2009), Peter Godfrey-Smith (2016), and John Searle (2017).

Functionalism

Functionalism developed as a reaction to type identity theory's chauvinism. Hilary Putnam, one of its developers, was inspired by the way in which the same programs can be implemented in different computers (Putnam 1960). The underlying mechanisms differ, but the patterns are the same. Whereas type identity theory identifies conscious experiences with neurologically individuated patterns of neural activity, functionalism identifies conscious experiences with patterns of abstracted state changes. The causal dynamics of a system matter, not the specific materials or their precise configuration.

Functionalism allows for experiences to be realized in different substrates because it is agnostic about the composition of the states and the entities that occupy those states. All that matters is abstract patterns of state changes. So long as different systems implement the same patterns of state changes, they possess the same conscious experiences, even when composed out of different physical materials.

Early versions of functionalism were offered as a general theory of the nature of consciousness. As with type identity theory, no specific identifications of conscious states with organizational states were proposed. Starting in the mid-1980s, a number of theories of consciousness have been developed that are naturally interpreted as functionalist. These include: Higher-Order Thought, Higher-Order Representation, and Dual Content theories, Global Workspace Theory, Tye's PANIC & Prinz's AIR, Integrated Information Theory, and Attention Schema theories[8]. Functionalism has been fruitful: not only have theories been developed, but those theories do work in explaining how we think about our conscious experiences. They make progress. For instance, functionalist theories seem like promising ways to solve the metaproblem(Chalmers 2018).

The Global Workspace Theory (GWT) provides a good example of what a functionalist theory could look like in practice. This theory says that consciousness is a product of the information architecture of a system. According to the theory, a human brain consists of a number of fairly modular cognitive systems that can share information with each other through a central repository. Information that is put in that central repository is broadcast widely to the modules. According to the theory, conscious experiences are blocks of information that are stored in the global workspace and are broadcast out to other modules.

GWT isn't essentially tied with functionalism. It is possible that human conscious experiences involve a global workspace architecture even if that abstract architecture isn't sufficient for consciousness. Nevertheless, the theory is well-suited for functionalism since it describes a pattern of state changes in a system and how they influence each other. It is plausible that if consciousness results

[8] Some references. HOT: Rosenthal 1986; 1993. HOP: Lycan 1996. Dual Content: Carruthers 2004. GWT: Baars 1988; 1997. PANIC: Tye 1995. AIR: Prinz 2005. Attention Schema: Graziano and Webb 2015; Graziano et al. 2022. IIT: Oizumi, Albantakis, and Tononi 2014; Albantakis et al. 2022

from a global workspace architecture in humans, it is because consciousness is tied to global availability, and so it would also result from the same architecture implemented in artificial substrates.

Another popular theory, the Integrated Information Theory (Oizumi, Albantakis, and Tononi 2014; Albantankis et al 2022), illustrates a rather different form of functionalism. Most forms of functionalism describe relations between states that play some cognitive role. In global workspace theory, the states are representations within an information architecture. According to IIT, in contrast, a system is conscious if its parts have an irreducible collective capacity to determine the future (or constrain the past) states of the system. This non-cognitive focus gives IIT some eyebrow raising implications when it comes to multiple realizability. Its chief proponents predict that a perfect digital simulation of a human brain with an ordinary CPU would not be conscious (Tononi & Koch 2017), but certain complex structures of logic gates would be (Aaronson 2014), despite the former and not the latter displaying a capacity for sophisticated self-aware behavior. IIT has its appeal in part because it enables very precise verdicts about arbitrary systems.

Summing Up

If consciousness is identical to some physiological process then digital systems are not conscious. Biological identifications are not particularly popular; the most plausible versions of theories that place weight on biological processes seem to me to be dualist in spirit and suggest that consciousness isn't just some biological process, but that some biological process gives rise to consciousness.

The best case for consciousness in digital systems depends on something in the ballpark of functionalism. Not only was functionalism inspired by computers, but it is designed to allow for multiple realizability. It also explains why behavior can provide evidence for consciousness. Behavior provides a better guide to functional organization than it does for specific cellular architectures.

Functionalism is among the most popular general categories of views about consciousness, at least among philosophers (Bourget and Chalmers Forthcoming). Most developed views about consciousness that might make predictions about consciousness in animals, aliens, and artificial intelligences are readily interpreted as versions of functionalism. It is therefore worth paying particular attention to what functionalism has to say about the possibility of digital consciousness.

3. Implementation through Satisfaction

According to functionalism, whether a system is conscious depends on whether it implements the right organization. Our assessment of artificial systems should turn both on which organization is right and on what it takes to implement it. To date, researchers have primarily focused on the first of these questions:

they have worked to identify how human brains are set up and which brain states are conscious. The second question warrants serious examination, especially in the context of potential digital minds.

Thinking seriously about the requirements of implementation requires getting precise about what functional organizations are. In this section, I present one useful approach to modeling functional organizations and describe a straightforward way of thinking about implementation under that approach.

Modeling Organizations with Sets of Claims

The functional organization of a system describes its internal causal structure and that structure's relation to its inputs and outputs. Human brains process perceptual inputs and produce behavioral outputs. Perceptual events change the state of our brains – they produce new beliefs and desires, influence our moods, and so on. The state of our brains also changes naturally over time. Moods subside. Beliefs beget further beliefs through deliberation. Desires and beliefs mingle to generate intentions. The state of our mind determines how we act.

We may identify a functional organization with a set of claims about the relationships between states, parts, inputs, and outputs[9]. These claims would describe how it is that the inputs to the system change the states of its parts, and what the states of the parts mean for what outputs it produces. The claims are likely to be complicated – what state transitions we see in any single part may be determined by the states of many other parts and the fine details of inputs. It is surely much too complicated to spell out an adequate claim set for a human mind, but it is in principle possible. More importantly, we can sketch out the overall structure of the correct theory and use those sketches to make decisions about different systems.

The claims in the claim set are formulated with generic language. This means that while we might have labels for different states of the parts, the claim set doesn't say anything about what those states are. Since the states are anonymous, only their relationships matter. We can think of each label for a part or a state as a variable that is subject to multiple interpretations. This allows the claims to apply differently in different systems. If beliefs are identified by how they interact with desires to produce actions, it doesn't matter if the beliefs are composed out of neurons or silicon chips. Physical details like that don't appear in the claim set.

Each claim in our model will be a logical construction of certain simple 'functionally atomic' claims regarding relations between states of parts.{.c33}[10] Function-

[9] This interpretation draws inspiration from Lewis 1972. The principal advantage of this approach over the more traditional isomorphism-based approach is that it is easier to add constraints and enrichments. See also Rescorla 2014.

[10] The dynamics of any real cognitive system are complex, and our functional architectures surely require a huge number of atomic claims to adequately describe. We can gesture at possible architectures without compiling a complete list of atomic claims so long as we accept

ally atomic claims are atomic because they attribute direct relationships between parts. These claims aren't necessarily simple: they could conceivably relate countless numbers of states to each other. What they share is an irreducibility in the relation they describe.

Some functionally atomic claims might look something like the following:

> If part **A** is in state **X** and part **B** is in state **Y** and input **C** is received then part **A** will transition to state **U**.
>
> If part **A** is in state **X** and part **B** is in state **Y** and input **C** is received then the system will produce output **V**.

Identifying functional organizations with sets of logical constructions of functionally atomic claims allows for a greater range of requirements than atomic claims alone. If ψ and ϕ are claims, 'not ψ', 'ψ and ϕ', and 'ψ or ϕ' are also claims. We can characterize a functional organization in part by the state transitions that don't occur within it, or by a group of state transitions only one of which needs to occur. We can quantify over claims, with variables substituted for labels. We might also allow probabilistic constructions, so that the transitions specified need not be deterministic. A claim might then describe the probabilities with which parts end up in different states or outputs are produced.

The Satisfaction Account

The traditional view about what it takes to implement a functional organization can be described in various equivalent ways. The way that I favor, which I will call the 'Satisfaction Account', looks to see how the claims of a model are satisfied by the system under some interpretation of the generic labels.

In order to see whether a physical system satisfies the claims of a claim set, we need to know how to interpret the claims. Satisfaction depends on the further notion of a carving, which describes how the physical system can be divided up. Functionalists must say something about which carvings matter. (In principle, any constraints added to carvings could be implemented directly by more carefully specifying the claims in the claim set. My purpose in distinguishing them is to highlight the kinds of constraints that make sense for carvings.)

Carvings

Every candidate system for consciousness will be some physical object in the world, such as a human brain or a computer. Whether it satisfies a given claim set depends on how its parts interact. As a physical object, its component materials can be divided or grouped together into parts in various ways. Similarly, there are various ways we can think about when a part counts as occupying a particular state.

---

that a fleshed out theory would be equivalent to such a list.

There may be some intuitive or familiar ways of dividing up the materials. Brains are normally split into neural circuits based on connection patterns. However, we cannot take familiar divisions for granted. Philosophers like to consider the possibility of objects with unusual boundaries. Incars, for instance, are things that are cars in garages. Drive an incar out of the garage and it ceases to exist, though an outcar immediately appears in its place. Where common sense says there is one object that moves from place to place, the incar metaphysics describes two objects that interact. Throwing away common sense, we might define all sorts of odd structures in our brains and assign them to different states. If we should ignore certain bizarre ways of grouping things together as parts or states when assessing systems for consciousness, we need to make sense of why this is so.

A 'carving' is a scheme for dividing a system into parts and interpreting the configuration of the system as states of the parts. (I use 'configuration' to describe the way the system is laid out, independent of any labels we apply to it. A carving has labels applied that group its materials into parts. There are many ways of carving a configuration.) Generally, the materials of the system can be divided into different groupings that will be responsible for settling the states of different parts of the carving. The incar/outcar metaphysics is one way of carving up automobiles, though the same materials comprise both an incar in a garage and an outcar when it exists. Those materials also determine whether the incar and the outcar counts as running.

Concrete vs Interpretive Carvings

Let a 'concrete carving' be a carving in which there is a direct link between the things that exist in the system and the parts of the carving. Each part of the carving corresponds to a real thing that is constituted by some material within the system. 'Real' here requires existence in some metaphysically robust sense. It will imply different things about consciousness depending on which groupings actually exist.

Let an 'interpretive carving' in contrast be a carving that specifies necessary and sufficient conditions for the system to count as having particular parts in particular states.Those necessary and sufficient conditions may make reference to any facts about the system, so the parts of an interpretive carving may be related to the material in the system in all manner of complicated ways. The parts of an interpretive carving don't need to correspond to any real entities within the system.

Consider the carving of the United States economy into companies. The economy might consist of the physical materials of human beings, machines, buildings, and so on. Which companies exist? That can't be easily read off of the people, machines, and buildings. We can't point to a specific collection and say: this is a company. Instead, the existence of companies depends on the intentions and expectations of the people, on the details in paperwork filed in government

offices, and on the body of laws that give that paperwork meaning. The same goes for the states those companies are in. For instance, we can specify necessary and sufficient conditions for there to be a company that is bankrupt in terms of the legal filings that make it so. This lets us make sense of talk of companies even if we think that companies aren't material objects in the way that atoms, cars, and people are. My point here isn't that companies don't really exist – perhaps they do, but we can talk about them regardless of their fundamental metaphysical status.

The difference between these interpretive and concrete carvings is that interpretive carvings are agnostic about the metaphysical credentials of the materials that are responsible for settling the states of the parts. It is controversial whether incars exist. It is also controversial whether companies exist. It is even controversial whether cars exist.[11] Interpretive carvings allow us to develop functionalist views without adopting substantial metaphysical commitments.They also allow us to focus on abstractions in the system and not care about how those abstractions are implemented. I'm not aware of any in depth explicit discussions of the merits of the two versions, but many famous discussions of functionalism (such as Putnam 1988) appear to have assumed an interpretive version of the view.

Interpretive carvings make digital consciousness more feasible. Generally, the parts of a computer system created by a program are abstractions over the underlying machine logic in something vaguely like the way companies are abstractions over people and paperwork. A neural network program isn't a material object the way a brain is. If we were restricted to concrete carvings, we'd need theories about which groups of electrical circuits existed.

The cost of interpretive carvings is that without further constraints of the sort that will be explored in the next section, they permit strange systems to counterintuitively satisfy the requisite organizations for consciousness.

Liberal and Canonical Requirements

The satisfaction account says that a system implements an organization defined by a claim set under a carving if there is some interpretation of the generic labels of the claims as the parts and states of the carving that satisfies all of the claims of the set. In brief: if the claim set can be interpreted as accurately describing the system, the system implements the organization.

To get from what it takes to implement an organization under a carving to implement that organization, we need to specify which carvings matter.

A fully liberal satisfaction account says that each carving can matter. In order for a system to implement an organization, there needs to be only a single

[11] For instance, some metaphysicians (e.g. Horgan & Potrc 2009) think that only one thing really exists and we can carve it up in any number of ways. Others (e.g. Rea 1998) think any combination of material parts constitutes an object, so the thing composed of the Eiffel tower and your nose is real. Others (e.g. Hirsch 2002) think there are no particularly interesting facts of the matter exactly which things really exist.

carving and a single assignment of generic labels to states and parts in that carving under which the claims of the organization come out true. This makes it possible in many circumstances to creatively assign parts and states to a system so as to reshape that system into a variety of quite different carvings.

A canonical satisfaction account says that some specific carving matters. We need to compare that carving with our claim set to know whether the system implements the organization. Different canonical accounts may identify different carvings as the ones that matter.

Functionalists mostly don't pay explicit attention to the distinctions I've drawn here. Some of the debates between functionalists can be understood through them. Early functionalists adopted a fairly liberal understanding of their theory, but were pressed by counterintuitive applications to adopt some constraints on which carvings matter. Canonical views are rare, though the Integrated Information Theory is a prominent view that has incorporated a definition of canonical carvings.

4. Problematic Cases

Functionalism says that dispositions toward patterns of state changes determine whether a system is conscious. In order to see if a system exhibits a pattern, we need to know how to interpret it. On liberal and interpretive versions of functionalism, all that matters is that there is some way of interpreting the system's configurations that gives rise to the right patterns.

Example Deviant Cases

The following cases[12] demonstrate that liberal and interpretative functionalism has some highly counterintuitive implications. These cases describe systems that do not exist and may not be physically possible (let alone technically feasible). Their conceivability illustrates a conceptual problem with the satisfaction account.

China Brain

In the classic paper, Ned Block (1978) proposed the China Brain thought experiment as a challenge for functionalism. Suppose that a large group of people (the citizens of China) together enacted the functional state of a brain. Each person is responsible for tracking the state of a single neuron. They use radios to communicate with the people representing connected neurons. When a

[12] A variety of bizarre implementations of functional architectures have been proposed in the philosophy literature (e.g. Maudlin 1989, Tiehan forthcoming). I make no claims about the completeness of the following menagerie of weird cases. Each of them, except for the mirror grid, has some interesting connection with the ways that computers actually work.

neuron would fire, the person representing that neuron would call their neuronal neighbors to let them know, thereby providing the information each needs to in order to decide whether or not to themselves fire.

Assuming each individual performs their role, the whole group would have the dynamics of a human brain. It could be hooked up to a robot body to receive sensory inputs and produce behaviors. There would be a straightforward way of carving up the group of people and assigning them states that allowed them to satisfy all of the same organizational claims that are satisfied by a human brain.

Neural Anthology

Suppose that every year, Eddison Publishing publishes an atlas of the human brain that depicts a single brain in vivid detail across the pages of a many volume set. The state of each neuron is fully documented, allowing neuroscientists to predict the next state for each neuron based on the states of its neighbors.

Successive editions of the atlas follow the state of a single brain as it receives perceptual inputs and produces motor outputs. To produce each new edition, the editors retrieve a copy of the previous edition from the publisher's library and go through it page by page to deduce the next states to depict for each neuron.

Over decades, the many volumes and multiple editions of the atlas capture the neural dynamics of a human brain with ink on paper, spread out across library bookshelves. Given that the publishers use a copy of the last edition for their updates, the states of one sequence of sets of volumes have the same causal dynamics of the human brain. Since the publishers produce many copies, most will be causally inert; each book may not be part of the conscious whole, but some lineage of copies does possess the dynamics of a human brain.

Brain Pointer

Suppose that we cover a wall with shelves of human brains preserved frozen in the middle of a state of activity. The wall is massive enough that every human brain state is represented with a brain frozen into that state. In front of the wall we set an arrow to point to a single brain and attach it to a computer. The computer moves the arrow so that it points at different brains at different times.

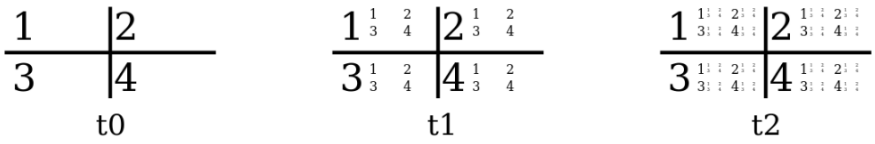
The computer receives data about perceptual inputs and uses a scanner to compare that data with the brains on the wall. It updates the pointer to point to the brain that would result from the brain it currently points to, if that brain were not frozen and received that input. The computer does not compute which brain to point to next and then go and find it, but rather compares each pair of brains until it finds one that fits: where no comparison of successor states violates the laws of neural dynamics.

The brains that the pointer points to are individually inanimate, but the 'pointed

brain' changes over time insofar as a different frozen brain is identified from moment to moment. We might think of this as a hopping system that inhabits one brain at one time and a different one the next. This hopping system implements the functional dynamics of a human brain, even though no continuous piece of matter does.

Mirror Grid

Suppose we set up a grid of mirrors that stood in front of another mirror. The grid contains a number of cells, each with a reflective surface with a number written on it. The mirror grid is positioned in front of the second mirror just right so that light reflects back and forth between the mirrors, causing the reflection in each cell of the grid to show deeper reflections of the whole grid over time.



The mirror grid exhibits a special form of complexity over time. The reflection displayed by each cell is unique over the life of the grid so long as the written numbers are unique. Each cell at any particular time reflects its own number and also the whole history of the grid.

The uniqueness and dependence of the cells makes it possible to set up isomorphisms with a system under any deterministic[13] functional organization, so long as that organization requires a number of parts less than or equal to the number of cells. To find this mapping, we assign parts to the cells of the grid. For any initial state of those parts, we assign those states with the initial configurations of the cells. Then we look at what states the parts would transition into, and assign those states to the reflectance patterns that cells of the grid would display. If a human brain implements a functional organization identified with a certain claim set, it is possible to find a large mirror grid such that will satisfy the same claim set under the right interpretation.

Number Mapper

Suppose that we built a computer whose central processor consists of three registers and a memory. The memory stores a lookup table that associates pairs of numbers with pairs of numbers. Each of the three registers holds a bit representation of a number. Every cycle, the processor looks up the pair of numbers stored in its first and second registers in the table and writes the resulting pair of numbers to the second and third registers.

[13] Probabilistic organizations might be produced by adding randomly generated numbers to each grid.

The look up table was produced as follows. First, a scheme was created that assigned arbitrary unique numbers to each of the many possible states of a human brain, to perceptual inputs, and to behavioral outputs. Then, a list of mappings from <brain state, perceptual input> pairs to <brain state, behavioral output> pairs was computed that reflects the transitions exhibited by a normal human brain. The first number encodes the total perceptual experiences a human could undergo. The second number encodes the total state a human brain could be in. The third number encodes the total behavioral output a human could produce.

The Number Mapper's table contains these mappings. Given its mode of operation, it changes in correspondence with a human brain. If we pass it encoded perceptual input states to its first register, it will infer human-typical encoded behaviors in its third register and its second register will update to a number representing an appropriate internal state of a human brain.

The Number Mapper lacks the internal complexity of a human brain: its internal state is recorded as a single very large number. All of the complexity of human neural dynamics is pushed into its table. Nevertheless, there is still a way of interpreting it as having parts such that they are isomorphic to the functional organization of a human brain. The parts of this system must be abstractions over the values of its middle register. By using the associations that define the table, we can create a scheme to tell us when the system counts as having a part in a given state. The system counts as having a part in the relevant brain state when it contains a number in its second register that is associated with a human brain with that part in that state. Whether a system counts as having a state will turn out to be tremendously disjunctive property (e.g. A or B or C or ...), but it will be as determinate as our brain states are.

Each of these systems can be interpreted as satisfying the same organizational claim set as a human brain. Plausibly, none of these systems would be conscious. This is a problem for liberal interpretative functionalism. The natural response for functionalists is to limit which carvings are acceptable. These systems lack something important that human brains have that prevents them from counting as properly implementing the right organization.

Physical Unreality

These examples are generally not physically viable. We could not build a brain pointer or a number mapper even with all the material in the visible universe. It would be preferable to have clear examples of deviant cases using systems that we could build, but it would take much more work to specify their internal structure and our intuitions about them would likely be much weaker. The advantage of these systems is that they offer straightforward cases for us to use to hone our concepts. There will be physically viable intermediate systems that have elements of the number mapper or the brain pointer systems that are harder to assess in and of themselves. Our feelings about the pure cases may provide us with some insight into the mixed cases. I think digital computers are

examples of such mixed systems.

5. Existing Constraint and Enrichment Proposals

In the last section, I discussed how liberal and interpretive functionalism generates counterintuitive verdicts about some hypothetical systems. We might respond by rejecting interpretativity or liberalism. Rejecting interpretivity will mire us in metaphysical disputes. Changes to liberalism can accomplish the same ends, so that is the approach that I favor.

This section discusses two types of changes to liberalism. The first consists in adding constraints. Constraints make no change to the way that organizations are represented but instead provide criteria that any carving must meet in order to properly implement a claim set. The other consists in introducing enrichments. Enrichments add options for formulating claim sets for organizations, so that they may include claims not just about generically labeled variables, but about certain kinds of entities or properties. Constraints add general requirements on how any system can be carved so as to satisfy a claim set. Enrichments expand what a claim set can demand of a system.

My intention here is not to assess these candidate amendments. Instead I explore what plausible functionalist theories might look like. I take from this exploration a general trend: integrity matters. In any plausible version of functionalism, the parts and states of a proper carving of the functional organization of consciousness, whatever that is, must have some underlying unity.

Constraints

The first way to amend liberal functionalism places constraints on which carvings are proper. A carving provides a scheme for dividing a system into parts and assigning states to those parts. A proper carving is a carving that, if it satisfies the claims of the right functional organization, is sufficient for consciousness. Constraints add requirements to such schemes specifying how a part or its state can be derived from the configuration of the system's materials.

Typically, parts in carvings are associated with specific material components in the system, and the parts those components satisfy correlate with states of those materials. For any given part, some subset of the material in the system, the material that 'underlies' the part, will be responsible for settling which state that part is in. This isn't strictly necessary, as some carvings may rely on global states. But it is natural to require clear dependence of parts on underlying states as a first step for specifying further constraints. Propriety constraints specify what sorts of material subsets can serve such purposes and how they can go about doing so.

Boundary Constraints

Suppose we accept that under a proper carving, some proper subset of the system's material will underlie each part. There will be specific aspects of the system to look at in order to settle which states the parts are in and they will not be a global property of the whole configuration. Boundary constraints introduce requirements governing the form of such subsets and how they relate.

One possible boundary constraint, contiguity, suggests that the materials that underlie a single part must be connected to one another so that the regions they occupy constitute a contiguous whole. The China Brain system violates this constraint: the people representing different neurons are separated from each other and rely on radio waves to communicate. (The human brain might also fail this constraint, so some care is called for in making it precise.) Assuming that the relevant functional parts of the system are composed of different people, the parts would be dispersed groups. The contiguity constraint prohibits this.

Another possible boundary constraint, independence, says that which material underlies eachpartmust not be dependent on the configurations of any other parts. Of course, we might need to get our bearings by looking at the whole system to identify which parts are which, but they won't then change over time. According to this constraint, we can know where to look to determine what state a given part is in before we look anywhere else. The Brain Pointer system violates this constraint, since the material (i.e. a portion of some frozen brain) that underlies every part will depend on where the pointer is pointing.

A third possible constraint, separateness, says that the materials that underlie distinct parts must not overlap, so that any single bit of material belongs to a collection that is responsible for settling the states of at most one part. The Number Mapper system violates this constraint, as which part is in which state is determined for each part by the exact value of the number in the Mapper's middle register.

Boundary constraints are a powerful way to prohibit deviant applications (Chalmers 1996; Godfrey-Smith 2009), but they may be too powerful. It is not completely clear that any of the above boundary constraints would not rule out human consciousness due to the structure of human brains. Given our uncertainty about how to carve the human brain into functional structures, it is possible that the most relevant carvings of the human brain will fail to respect contiguity, independence or separateness. This isn't a decisive reason to reject these constraints, but it is a reason for caution.

Naturalness Constraints

The carvings necessary to regard my example deviant cases as sharing a functional organization with the human brain come across as weird, complex, and gerrymandered. We might try to rule out such carvings by disallowing the weird, complex, and gerrymandered configurations from settling the states of

different parts. This can provide a second compelling way to avoid problem cases (Godfrey-Smith 2009).

In formulating this idea as a constraint, we can borrow from the philosophical concept of naturalness. The classic example of an unnatural property is grue (Goodman 1955). Grue is defined to be the color blue if it is applied to an object first seen before the year 2000 and green otherwise. Grue is defined in terms of blue and green, but blue and green could have been defined in terms of grue and bleen. Despite the potential definitional symmetry, it isn't arbitrary that we think in terms of blue and green rather than grue and bleen. Blue and green are the more natural properties. In a way, the configurations that settle states of parts in the deviant applications will be a bit like grue and bleen.

A naturalness constraint requires that the states and parts of a carving be relatively natural. Though we can specify carvings in which the states of parts are settled by complex, gerrymandered configurations of the system, such carvings will invoke properties that are unnatural in the way that grue is unnatural.

Naturalness is an intuitive notion and it is easy to judge in many cases. However, the intuitive notion should be supplemented with a formal account if it is to do work in deciding which systems are conscious. The traditional account of naturalness treats some fundamental properties as completely natural, and others as more or less natural based on the complexity of their definition in terms of the completely natural properties. For instance, the properties of fundamental physics (charge, spin, mass) may be perfectly natural, and the relative naturalness of different colors may depend on how hard it is to define the relevant surface reflectance properties in terms of the properties of fundamental physics (Lewis 1983). It is hard to define grue in terms of the properties of fundamental physics. So grue is highly unnatural.

With the intuitive notion and its formalization in hand, we might adopt a constraint according to which the parts of a proper carving and/or the states those parts occupy must correspond to fairly natural divisions of the system.

Both objects and properties can differ in their naturalness. If we accept that the parts of a system correspond to material constituents, then we can interpret a naturalness constraint in terms of the degree of naturalness that needs to be satisfied by configurations of the system that settle the states of the parts.

Naturalness might be imposed as a constraint on the parts on a carving or on their states.

When imposed on the parts, each part must be composed of regions or materials that are naturally grouped together. If we understand naturalness in terms of natural properties, the parts must have some natural property that distinguishes them from other parts. The Mirror Grid system violates this constraint, as any groupings of possible mirror reflection patterns will be arbitrary.

When naturalness is imposed on the states of the parts, the various configurations that make that part count as occupying any single state must be natural. The

Number Mapper system will violate this constraint, as the same psychological state is said to occur if many numbers are in the central register, and the bit representations of those numbers have nothing physically in common.

Naturalness can be imposed in an absolute or relative sense. An absolute version says that there is some minimum threshold of naturalness that is allowed for each part or state. A relative version would say that carvings need not be natural so long as they are more natural (or at least not substantially less natural) than the naturalness of relevant alternative carvings.

Naturalness comes in degrees. Some properties may be perfectly natural, but most will not. This suggests that naturalness constraints will have to deal with the fact that there are no obvious (relative or absolute) thresholds of naturalness to permit. If it is assumed that consciousness isn't the sort of thing that comes in degrees, then this may be problematic. David Chalmers (1996 p. 312) objected to naturalness constraints for this reason. However, the gradations imposed by naturalness may be difficult to avoid: (physicalist forms of) functionalism may unavoidably predict that consciousness is vague (Papineau 2002).

Enrichments

In the satisfaction account, organizations are modeled with claims that use only generic terms for states and their parts, effectively making each term a variable that can stand in for any object or property. Enrichments expand the language available for formulating the claims in such models, allowing the specification of the model to itself put constraints on which carvings implement it.

The expanded language introduced by an enrichment might refer to specific sorts of parts or states of those parts. A proper carving must carve the system up into parts or states that satisfy those terms. The claims using that enriched language can only be satisfied by a carving where the terminology applies to the underlying material. If a claim says that a part of kind K, for some specific K, plays a certain role, then the claim can only be made true by a part in the carving that is composed of material of kind K.

As a toy example, we might expand the language to include color terms, so that claims like 'if part A is red and input X is received, then part B becomes blue' can be included in a model. Any proper carving that implements this model must divide the system up in parts based on materials that possess the relevant colors.

Whether enriched versions of functionalism still deserve to be functionalist depends on the nature of the additional language. Incorporating color terminology is non-functionalist in spirit, since it would suggest that consciousness is dependent on some non-organizational aspects of a system. We could conceivably incorporate biological terminology to create a hybrid view between type identity theory and functionalism. The proposals discussed below seem to fit better with the way that functionalists have traditionally conceived of their view.

Two of the existing views described below are borrowed from the debate over what it takes for a physical system to implement a computation. This is a topic in the philosophical foundations of computer science, not consciousness, but it has some bearing on functionalism about consciousness. Many philosophers have drawn on computers as a source of inspiration and many have even embraced a kind of computationalism about consciousness in which the functional architecture of consciousness is identified with implementations of a particular sort of computation. Whether or not we agree with the computationalist variant of functionalism (Piccinini 2010), the work on the question of implementation provides interesting avenues to consider.

Agential Enrichments

Agential enrichments introduce language that is specific to agential inputs and outputs. Humans receive perceptual events as inputs and produces motor responses as outputs. Perceptual events involve specific external sensory apparatuses: our eyes convert light into neural signals. Our ears convert sound waves into neural signals.

If we take the impact of light or sound waves to be critical components of the functional organization of our brains, then the proper specification of our organization may make reference to light or sound and eyes or ears. Systems that don't actually perceive light or sound and don't have eyes or ears won't be able to implement the same functional organizations.

Similarly, behavioral outputs involve moving our limbs in various ways. We might incorporate bodily descriptions into our functionalist view, so that how we move our limbs is part of the description of the system. A system that does not have limbs will not be able to implement a functional organization that makes explicit reference to behaviors that require limbs.

Agential enrichments have some appeal, but they have strange implications of their own. We could imagine growing a human brain in a vat that we fed with the same neural signals as it would receive if it had a typical body replete with sensory organs. No neural signals reaching its brain would be any different for it. The sources of such signals would be a computer, not an eye or an ear. Most people assume that such a brain would be conscious, yet agential enrichments would rule it out.

Similarly, we might set up any eclectic system such as the China Brain or the Number Mapper to be able to receive perceptual inputs and hook it up to a robot body so that it could produce behavioral outputs (Godfrey-Smith 2009). If agential enrichments are the only things keeping the China Brain or Number Mapper from properly implementing the functional role of consciousness, then such an apparatus will imply that the underlying system would become conscious.

Semantic Enrichments

One dominant view of the nature of computations, semantic accounts (e.g. Dietrich 1989; Shagrir 2021), requires that computations involve the manipulation of objects with semantic values. In other words, some of the parts must involve material that is imbued with intentional meanings. Semantic enrichments allow semantic terminology to figure into the formulation of the claim set.This makes it possible to constrain conscious systems to those that process information in certain ways. Unenriched functionalism can make demands about how states change around a system, including how one state in one part can produce corresponding states in other parts, but it cannot directly require that those states bear information.

Intentionality, the property of having a semantic meaning, is something that has been thought by some to be a deep mystery of the world, somewhat on par with consciousness. It would be of limited help if we could define consciousness in terms of systems with intentionality, but were unclear about what it took to have intentionality.

Many physicalist-friendly accounts of intentionality exist. Intentionality is typically divided into fundamental and derived forms. Some material objects acquire meanings by virtue of their relation with others. Tracking accounts (Dretske 1981; Loar 1986) say that intentionality is roughly a matter of correspondence. The fact that certain neurons in your brain count as representing the color red can be explained by the fact that they generally track that color by virtue of the structure of the eye and the causal relations between them. When you see red, they fire. Teleological accounts (Milikan 1986) draw on evolutionary history. If some neurons are the result of an evolutionary lineage of neurons that helped our ancestors survive because they stood for the property red in certain internal calculations, then they mean red.

One potential[14] problem with semantic accounts is that they make meaning contextual, which makes consciousness contextual. Meaning in a system isn't purely a product of the internal dynamics of the system, but how it relates to the world at large. For instance, tracking accounts of intentionality can allow duplicate brains to have different semantic values if they exist in different environments. Consider a Boltzmann brain, a bit of matter in the shape of a human brain that materializes by chance in the vacuum of space. Though it might be atom-for-atom identical to you when you enjoy a slice of apple pie, its parts would have no semantic values whatsoever. If we impose a semantic constraint on consciousness, the Boltzman brain would not satisfy the organization necessary

[14] Phenomenal externalists believe that our experiences are shaped in part by factors outside of our bodies independently from what is going on inside of us. So duplicate brains could feel differently depending on the other objects that (historically) exist in the environment. In the context of digital consciousness, this view raises difficult questions: how should we think about the semantic values of computer chips that have no evolutionary history? Are there clever tricks we might use to create systems that lack the requisite semantic values for consciousness despite being materially identical to systems that are conscious?

for consciousness.

Telic Enrichments

Another prominent account of computation implementation, Piccinini's mechanistic account (2015), does away with semantic requirements and understands computations in terms of the processing of vehicles in terms of substrate independent rules. Understood broadly enough, many systems can be interpreted as manipulating vehicles in accordance with substrate independent rules. However, Piccinini incorporates a telic requirement (Piccinini & Maley 2017) into his theory that helps ward off deviant applications. This telic requirement states that the mechanisms implementing the rules must do so as a result of their purpose (see also Turner 2018).

Telic constraints require that the parts in a proper carving are delineated by the possession of a purpose. Gerrymandered ways of carving a system into parts may capture the right causal dynamics to implement a computation, but their parts aren't playing any genuine roles; they are interacting without an end.

Like semantic values, purposes must be understood as part of the physical world. A traditional way to do this interprets the purpose of an object in terms of the history of that object, or similar objects. The purpose of a heart is to pump blood because the heart was crafted for years through evolution in order to pump blood. If our ancestors' hearts did not pump blood, they would not have survived. Purposes can also be granted to artificial systems through deliberate efforts. Hammers have the purpose of driving nails because that is what we made them for.

Like the semantic constraints, telic constraints will make satisfaction context dependent. The same material components of a system may or may not have any purposes at all, if they are spawned in the void. This is an appropriate constraint for computation in general, but makes less sense for consciousness. An exact copy of the complete works of Shakespeare wouldn't be the complete works of Shakespeare if it spontaneously formed in the vastness of space because its word-shapes wouldn't be embedded in the right linguistic tradition. An exact copy of a laptop running windows that spontaneously formed in the vastness of space would, perhaps, not be a computer running windows. This is a good reason to think that consciousness involves something less parochial than computation, if computation involves semantics.

Integrity Matters

By "integrity", I mean the coherence of material configurations of a system responsible for settling the parts of a system and what states they occupy. The constraints just considered focus on requiring coherent groupings. The enrichments also lean on properties like semantic meanings or telic purposes to bind groupings together.

Though most of these proposals have some challenges that remain to be worked through, the fact that all of these proposals enforce some degree of integrity on the parts of a carving suggests that we should expect to have to accept some change to the satisfaction account that does this work.

The motivation for each of the above proposals was in part to avoid deviant cases. However, as I hope to illustrate in the next section, integrity has an appeal that transcends this motivation. It makes sense that the components of a system that allow it to count as having an organization should be robust and independent of that functional system.

6. New Constraint Proposals

For any description of the functional organization of human brains, there are possible systems that are both intuitively highly unlikely to be conscious and that can be carved to satisfy that description. This is a problem for liberal and interpretive functionalism that suggests a need for refinements of the theory.

Functionalists may add constraints or enrichments to handle problems case by case. I'm skeptical that one constraint or enrichment would solve every problem. For each challenging thought experiment, we might amend the theory slightly by adding some additional requirements. This is a dangerous methodology. Our success in identifying every relevant problem depends on our imaginative capacities.

Instead, we should let the thought experiments provide general lessons about the nature of proper implementation. With a coherent idea of the kinds of things that matter to implementation, we may examine how current artificial systems implement the sorts of organizations that plausibly constitute consciousness. Even if we do not have a complete theory, it is likely that we may recognize potential issues.

The challenges facing the liberal and interpretive variants don't doom functionalism as a whole. Many of the examples of systems with deviant carvings don't intuitively share our brains' organization. Functionalism should require that the relevant parts of a carving have a fairly unified existence separate from the roles that they play. The states the parts count as occupying must not depend on overly heterogeneous configurations. In other words, functionalism ought to require that the parts and the states of a system have integrity.

There are a variety of potential ways a system could lack integrity, only some of which appeared in my deviant cases. In the current section, I survey a variety of challenges to the integrity of neural networks as implemented in modern computers.

Neural networks have produced the most clear and compelling advances in recent artificial intelligence. Current pressing questions regarding artificial consciousness relate primarily to large neural networks. It seems likely that any paradigms

of the near future will be similar. The issues below are specific to current technologies and techniques; other approaches may introduce different challenges to integrity.

Candidate Constraints

Lack of Contrast

Contrast concerns the way that structures in a system are differentiated by their non-functional properties. Non-contrasting structures require a fine-grained understanding of the dynamics of the system to distinguish from each other. In a completely uncontrasted carving of a system, simply looking at a single time slice would reveal nothing about which structures exist (except insofar as the system's past or future behavior could be deduced).

The parts of a car are high contrast because they can be distinguished based on criteria of material connectedness and composition. The subset of a hand of dice that will land on six after being tossed in a particular precise manner is low contrast, as you need to follow the complex physical interactions of the dice to tell which dice those are. The employees of a company working out of a shared office space are also low contrast. They are distinguished from the employees of other companies by in subtle ways – e.g. the fact that someone is referenced in certain payroll software – rather than by obvious physical relations like who frequently talks to whom.

A contrast constraint would require that only relatively high-contrast carvings of systems properly implement functional organizations.

Neural circuits may be recognizable by their connection patterns even without seeing them in action. It is possible to trace which neurons are connected to which. Many aspects of the structures involved in cognition will require a close look at each neuron, but they may not require us to follow complex paths in order to deduce any neuron's significance in a circuit.

In contrast with human brains, structures in computer memory require complex tracing of the logical dynamics of the system to identify.

- Data is generic. RAM consists of a large undifferentiated expanse of bit cells assigned memory locations. Each bit cell may hold any kind of data. The cells themselves do not record the kind of data that they hold or any concrete information about the structure of a data object. Numbers, text, memory locations, all are just strings of bits.
- The data held a memory location is interpreted as a particular kind of object by the program that manipulates it. In order to figure out how data is being interpreted, we would need to understand where the program is stored and how the bits that record the operations of the program respond to that data. If multiple programs are granted access to the same region of memory, they may interpret its contents differently.

- Data structures aren't generally marked or typed in the same place they are stored in memory. Nothing about the data signifies the beginning or end of the data. An array of connection weights is just a string of bits. The location of the start and end of that array and the divisions between bit representations of numbers are recorded somewhere else. What generates the distinction between the data of one structure and another are the references to those data stored elsewhere. What makes those references references to that data and not other data is how they are referred to elsewhere, and so on.

Material Complexity

Material complexity concerns the intricateness of the scheme relating the configuration of the material entities of the system to the states of its parts. In carvings of systems that are highly materially complex, long chains of reasoning and a number of separate facts about the material entities of the system are required to explain why the system counts as having a part in any particular state. Though closely related, contrast focuses on the immanence of the structures in the systems while material focuses on the lengthiness of the schemes for deciphering those structures in terms of an underlying ontology of material objects.

Compare the state of an army and the state of a corporation. The state of an army is mostly a straightforward matter of the number of personnel in uniform, how supplied they are with arms and equipment, how well positioned they are to strategic points, and so on. The state of a multinational corporation depends more on legal and financial abstractions: what it counts as owing, the debts it counts as having, the prospects it has for sales, and the promise of its intellectual property. The material complexity of a corporation is high because it depends on a large amount of distributed intentions, expectations, and paperwork located in the right minds, hard disks, and filing cabinets.

A material complexity constraint would require that only systems with relatively simple relations between their configurations and their material parts properly implement functional organizations. Systems with an organization that exists not between material components themselves, but complex abstractions over material components, might not be capable of consciousness.

In a human brain, neural connections physically connect neurons and their firing is relatively straightforward. Although we are a long way off from providing detailed interpretations of neural activity patterns, it seems likely that activity patterns will straightforwardly translate into functional states. States of a human brain will probably consist in patterns of levels of activity in different circuits. We can see this in the fact that mental states show up (at least vaguely) in extremely coarse-grained methods of inspection, such as fMRI scans.

Interpreting the relevance of any data in a computer requires much more detective work. We have to trace back through the structures that comprise the program

to see what those structures are and how they are being used.In this way, a computer is more like a corporation and a brain more like an army: you can figure out a bit about what is happening in a brain with a fMRI scan just like you can figure out what is happening in an army with satellite imagery. You can't quite so easily use an x-ray to figure out what is happening in a computer without doing the equivalent of cross-referencing all of the paperwork.

- Memory is managed in a virtual address space. Units of memory that appear adjacent from a program's perspective may be mapped to physical memory addresses that are not adjacent. Continuous data structures in memory depend on the sequence of program instructions used to manipulate them and the way the memory manager is set up.
- Programs generally store complex objects piecemeal in order to accommodate the variable sizes of data. An object's data contains a master record with pointers storing the locations of all of its parts. This means that a data structure containing the values for a neural network will generally store the weights for the connections between different layers separately. The data representing the activation levels for a given layer may also be stored in separate locations. The fact that they are related to each other depends on the meaning given to chains of pointers.
- Data structures are not explicitly identified. To know what sorts of data are stored somewhere, one needs to see how they are used by the program. This requires tracing program logic through a series of pointers and a number of subroutines.
- Elements of array structures are typically separated implicitly by virtual memory offsets assumed by a compiled program. In artificial neural networks, the connections between layers are represented by floats stored at the appropriate point in an array. Whether two nodes in different layers of a network stored in separate arrays are connected in the network is reflected in how the program uses the arrays that record its connection strengths to calculate activation levels for the next layer. A program running a network will use the offset of each value in the array from the location of the start of the array. This depends upon the dimensions of the tensor the array represents and the length of each datum.

All of this means that the logic through which a computer program state is interpreted is diffuse and complex. One needs to follow long chains of logic to infer from the hardware what basic states a program is in.

Lack of Causal Integration

Causal integration concerns the extent to which the materials underlying parts of a system are responsible for the causal powers of those parts. In an unintegrated carving of a system, the materials that underlie the parts exhibiting certain state transitions are passive rather than active. They are acted upon, rather than act: some other mechanism transforms them, but they do nothing in themselves. A causal integration constraint would require that only systems with parts that

are active in producing their state transitions properly implement functional organizations.

Neurons in a human brain directly stimulate each other through their activations. The release of neurotransmitters by one neuron directly affects the state of its neighbors. No external force is needed to manage them. There may be some neural circuits that are highly dependent on top-down controls. But this does not appear to be the rule of neural activity.

In contrast with brains, the data comprising the activations of a digital neural network have no capacity to make any changes themselves. Data sits in memory until it is loaded by the processor, altered, and stored. Something reads them and updates them. They do nothing themselves. The structures in a computer that are most clearly similar to the structures that play a functional role in human brains do not have any internal powers to bring anything about.

Discontinuity

Discontinuity concerns the extent to which the materials that underlie the parts of a system remain the same over time. A carving of a system is highly discontinuous if different material components of that system determine the state of a given part at different times. A continuity constraint would require that only systems whose parts depend on the same materials over time properly implement functional organizations.

The neurons of a human brain develop early and remain throughout the individual's life. The physical particles from which they are composed change over time, but they do so gradually. As a result, the same neurons, composed of the same matter, are involved in computations time after time.

In a computer, neural networks are stored as models in long term memory and loaded into RAM when needed for operation. Where data objects are placed in memory is variable, so the same model can be loaded into different memory locations at different times. A few chips of RAM may contain all of the memory addresses that the system uses to store data, but where in those chips data is stored will differ over time.

- When a program starts, the operating system assigns it an address range of virtual to use. As the program runs, it may request more memory as its needs expand. In both cases, the program is assigned memory from what is available, which will change depending upon what memory is currently assigned elsewhere. From run to run of a single program, physical locations in memory where it stores its data may change.
- Programming languages often utilize a system of contexts to organize data. When a function is called in the program, a region of memory is set off to keep track of the data for that context and any assignments of local variables. Which region of memory is used will depend upon what is available when the memory is assigned. When one function is called

repeatedly within the course of a single run of a program, it may make use of different regions of memory. Thus, the regions of memory that store the values of activation levels are likely to change even within the course of a single run of the program.
- Computational labor in modern computers is divided up between multiple processors. For neural networks, both the multiple cores of a single CPU and the many streaming multiprocessors of a GPU may be used. Which processors handle which computations can change depending on what is available when a specific computation is requested.

Fragility

Fragility concerns the degree of robustness or overdetermination with which the configurations of the system count as occupying different states. A highly fragile carving of a system will tolerate little change in how it is configured before its parts count as being in separate states. A fragility constraint requires that only systems that are relatively robust would properly implement functional organizations.

Brains are massively parallel and prone to the random firing of neurons. Large scale dynamics are not the product of any specific neural activities, but result in aggregate. Stimulating a single neuron to fire will not generally make any significant difference to the behavior of the system as a whole.

Programs are much more fragile.

- Changing any of the bits in the sequence of data that specifies which operations are to be performed by the program or in the program counter will cause it to fail in a way that is likely to completely disrupt the program.
- Many functions depend on a number of core subroutines. In order for these functions to work as expected, the core subroutines have to be available and work correctly. This means that even local functions are likely to depend on the specification of a number of instructions in the program.
- Programs are dependent on the software of the operating system. Operating systems routinely take control over a CPU to handle other background tasks after copying its existing state into memory. If they malfunction or otherwise refuse to return it, a program will not continue to work as expected.

For these reasons, there are quite exact physical computer states that are needed to determine the proper interpretation of the data and the implementation of functions that make up a neural network.

Disorder

Disorder concerns how systematically the configurational details contribute to settling which parts are in which states. In a disordered carving of a system, there are fewer patterns in the ways that configurational details contribute to

deciding which parts belong to which states. An order constraint would require that only systems whose parts and states possess an ordered relation to the systems configurations would properly implement functional organizations.

Neurons in a human brain exhibit patterns of firing and not firing. Active neurons are alike in propagating electrical signals. It is plausible that larger-scale circuits also exhibit common patterns, such as possessing groups of neurons firing in synchrony.

There is some cause to think that nodes in a neural network possess at least a few different kinds of (mild) disorder.

- Nodes in artificial neural networks are represented with floats. In general, the more bits in a float, and higher the position of the bits, the larger the number. However, floats are a bit finicky. Certain bits and certain combinations of bits have a unique significance that alter the interpretation of other bits. The levels of activity in a particular region of a neural network can be similar without looking similar, or can look reasonably similar while being different.
- Arrays of activation levels can be encoded with different data structures. The standard format is a sequence of memory addresses with floats saved to addresses at a location with an offset from the start equal to the position of the value in the linearized tensor. If the vast majority of the values are 0, this is inefficient. Sparse array formats instead save the coordinates in an array and the values at those coordinates. In a neural network that uses sparse array formats for some purposes, the representation format for activation levels may change from layer to layer.
- Connections between nodes in different layers can be represented in different ways. For layers where connections share weights, like convolutional layers, the weights of connections between distinct nodes may not be separately represented.

These examples fall far short of showing that neural networks are highly disordered, but they demonstrate how various kinds of disorder can creep in. There is a general trend in computing to work out more efficient ways of performing specific types of computations. As machine learning continues to develop, I expect that we will see more ways in which the handling of neural computation diverges in different parts of a system, with different layers using different data structures.

Detachment

Detachment concerns the extent to which the materials that underlie a single part of a system are determinate at any given time. This assumes that parts exist over time, and that states of part might depend on patterns in that part over time. Whereas continuity focused on whether the same materials were present invocation to invocation, detachment concerns how fixed the materials are each moment. In a detached carving of a system some parts may depend

on different materials even as they play a role. A detachment constraint would require that the material that underlies a part be relatively fixed while that part plays its role.

The concern here is motivated by the following example, though detachment may show up in other ways. Suppose that we run a neural network with multiple processors (perhaps one CPU and a number of GPUs). Each processor is responsible for a single layer of the network, but each layer has multiple processors and any one of those could handle it. This setup allows the system to process multiple runs of the network efficiently and simultaneously. A root processor would receive external requests. It would assign other processors to handle the computations for each layer. However, it wouldn't need to assign processors to each layer at the onset. It could instead decide where to send the results of each layer in order to process the next layer as it is needed, based on which processors are free at that time. This means that when a processor is performing the calculations on a single layer, there is no specific processor to which it will next be sent. Instead, there are a number of possibilities.

In this hypothetical set up, the processing of each layer is detached from the processing of other layers, such that any organizational structures that occupy multiple layers will have an indeterminate extension even as they are being computed. It is unclear which parts they will have before they are assigned, and they are assigned only at the time they are needed.

Motivation

The main motivation for these candidate propriety constraints is that they relate to integrity in a way that seems important for functionalism in general. They draw support from the thought that integrity is needed, as demonstrated by the deviant cases, and illustrate plausible flavors of integrity. I find the case for each constraint from integrity alone to be somewhat weak. For most of these considerations, I am somewhat ambivalent about how much they matter. For others, I feel more strongly.

One way to assess the appeal of potential propriety constraints is to draw on intuitions about cases where they are violated. To be most effective, we must have a reasonably clear intuition about these cases and they must isolate the candidate form in integrity in question. For some cases, like the example deviant cases, our intuitions (or mine, at least) are strong. However, these cases tend to depict systems that violate a number of candidate integrity constraints, not just one. I have found it difficult to isolate most of these propriety considerations in thought experiments where I have strong intuitions.

That said, I think there are some powerful thought experiments that could be used to support several different constraints, which would be enough to show that contemporary digital computers are unlikely to be conscious. Some of the example deviant cases fit this purpose. In particular, the Brain Pointer,

the Number Mapper, and the Neural Anthology all provide support to certain constraints.

Here is another experiment that addresses the plausibility of the continuity and causal integration constraints:

Brain Transducer

The brain transducer is a machine that takes as an input a human brain that has been frozen into a single state within a preservative medium and produces as an output a fully new human brain frozen in another brain state. This machine would disassemble the input brain and construct the output brain out of new atomic materials that reflected what state the input brain would have momentarily occupied were it not frozen. We might route the output brains back around to form the machine's inputs so that it produced a constant succession of new frozen brains reflecting the states that a human brain would naturally occupy as its internal dynamics evolved over time.

My intuition is that the series of frozen brains output by the machine would not be conscious, even if the succession of brain states replicated a human brain having a conscious experience. The difference in the materials prevents the identification of the separate brains as part of the same mind. This intuition may be reinforced by the observation that it isn't essential that the transducer destroy the original brain or make just a single new brain.

The transduction process looks somewhat similar to what artificial neural networks do. Just as the transducer reads frozen old brains and produces new brains, neural network programs process prior activation levels through functions and spit out subsequent activation levels into new regions of memory. Both the activation levels and the frozen brain states have no power unto themselves. They are used by the network / transducer to figure out what to do next.

To provide some support for these constraints, this thought experiment doesn't have to convincingly show that the brain transducer wouldn't be conscious, only that it is plausible that it wouldn't be even if functionalism is true. I think this is a fairly low bar that the example easily passes.

7. Discussion

The basic argument for the possibility of digital consciousness is that consciousness is a pattern and the same patterns that occur in us can be created in artificial systems. If functionalism is false, then there is little reason to expect

artificial duplicates of human brains to be conscious. If even duplicates of human brains aren't conscious, it is not clear why we should think any artificial systems would be.

Functionalism, however, is a broad category of views. To provide specific verdicts about hypothetical systems, it requires elaboration. One aspect in need of elaboration concerns which patterns matter. Another is what it takes to implement those patterns.

Traditional functionalist views are liberal and interpretive. This commits them to counterintuitive verdicts about certain odd cases. These deviant cases motivate amendments to liberalism. We might either add enrichments to the way we specify systems or constraints on what it takes to implement an organization.

There are a number of plausible propriety constraints. I surveyed some possible integrity failures in the last section, but many have not yet been formulated. Given the organizational differences between humans and computers, it is not unlikely that some plausible propriety constraints will entail that computers with a traditional architecture are not capable of consciousness even if they perfectly replicate our behavior.

Our Epistemic Situation

We lack good ways to evaluate competing claims about the requirements of implementation, such as regarding which propriety constraints are true or which enrichments are appropriate. One difficulty is that empirical research into consciousness typically relies on testimony and introspection. This is possible, to some degree, as we can investigate what our own brains are like and use our own introspective confidence in our consciousness to rule out any constraints which would predict that we are not conscious. It does not, however, offer much help to decide between constraints that we all satisfy.

Another difficulty is that systems can violate some propriety constraints while sharing the functional organization of our brains. Under a suitable interpretation, an AI system and a corresponding human brain would be behaviorally identical. Anything we would say about our own consciousness, they would say about theirs. We know this despite any doubts we might have about their consciousness. Testimony will be the same no matter what such systems actually feel internally.

Consider the AI Consciousness Test, one of Susan Schneider and Edwin Turner's (Schneider 2020) proposed tests of digital consciousness. The test is to set up an artificial system and let it run then see if it reacts to its internal states the way that we react to consciousness. Evaluation is holistic focused on the system's facility with using consciousness related concepts in the absence of prompting or training. Relevant behaviors range from the systems testifying to their own consciousness to wondering about an afterlife to mourning the dead. The hope is that only a conscious system would learn to conceptualize their inner lives the way we conceptualize ours without coaching.

One challenge for the AI Consciousness Test is that any systems that are functionally organized like us will be disposed to make the same sorts of claims (Udell & Schwitzgebel 2021). Suitably interpreted, the Neural Anthology, the Brain Pointer, and the Number Mapper, being functional duplicates of us, will think about their internal states exactly the way we think about ours. The AI Consciousness Test may be useful in evaluating whether systems have the right architecture to count as conscious, but it cannot be used to probe the requirements for proper implementation of the right consciousness architecture or whether functionalism is true at all.

Instead of empirically evaluating alternative hypotheses, we are forced to apply general heuristics of abstract theory preference. Every theory in the running will say that human beings are conscious. Not every one of them will be equally good. Consider the theory: to be conscious is to have a global workspace architecture or to be a toaster. This is a bad theory. It is worse than the global workspace theory by itself, even though they agree about every system we can directly inspect for consciousness. It isn't worse on empirical grounds, only on theoretical grounds. It treats consciousness as a bizarrely disjunctive property. The union of global workspaces and heating apparatuses appears arbitrary. That is our reason not to believe it, not direct evidence.

Abstract considerations of theory preference cannot do too much work in settling which theory is correct. For one, there will be more than one elegant theory of consciousness. But more importantly, theoretical considerations are seldom decisive. The world may not be as elegant as we hope. The fact that we have to rely on such considerations means we cannot be too confident in any conclusions we come to.

Our only alternative is to rely on intuitions about thought experiments, such as in the deviant cases. It is unclear why we should trust our intuitions about candidate propriety constraints or about the status of the systems in the deviant cases. This is part of the challenge of consciousness research. There are no good objective tests of another system's experiences. Given that we must rely on intuitions about thought experiments and abstract theoretical considerations to decide between theories, our epistemic situation is poor.

We must accept the poverty of our situation and expect never to be completely sure about our conclusions. Instead, I think we can use intuitions to offer some guidance about which theories to take at least vaguely seriously. The right way to understand proposed constraints is as open theories. We must work around what we don't know to reduce the risk of accidental harm or missed opportunities.

Reducing Ambiguity

Ambiguity makes it harder to know how to ethically treat artificial systems. If we cannot tell whether a system is capable of conscious states, then we may give it either too little or too much regard.The uncertainty leads us to make imperfect

choices, even if we respond to our uncertainty with appropriate caution.

One response to the costs of ambiguity is to work to reduce it. We might make some sacrifices to produce systems that are less ambiguously conscious. Given that we can't confirm a single theory of consciousness as correct, the best way to reduce ambiguity is to explore possible theories and build systems that satisfy either most of them or few of them. (Alternatively, we might focus on reducing potential costs of inappropriately matched regard, such as by eliminating capacities for wellbeing or illbeing in ambiguously conscious systems.)

If we want to resolve ambiguity by making systems that are clearly conscious, then integrity considerations may push us toward building systems that better replicate the structures inside the human brain. It may make sense to promote cognitive computing systems for AI.

If we want to resolve ambiguity by making systems that are clearly not conscious, then integrity considerations may provide us with ways to reduce similarity to the human brain. We could work to further split, disorganize, and obfuscate the implementations of candidate digital minds.

Conceivably we could do both within a single system. We might be able to produce systems that replicated human functional organizations in high fidelity integrated systems for some parts and low fidelity non-integrated systems for others. We might, for instance, discriminate based on valence, such that only positively valenced states had integrity. Such a system might be behaviorally identical to us, but would plausibly not be capable of suffering, since it's negatively valenced systems would be implemented in ways not conducive to consciousness.

Example: Continuity through Dedicated Memory

The standard von Neumann computer architecture stores program instructions in the same way as it stores data. Data is generic and can be put to any use. A given program can use any region of memory, and can change that region from run to run. No parts of memory are dedicated to use by a neural network. Fixing which bit cells store which data would give those structures some substantial identity over time, which could be important for those structures to count as properly implementing a functional role.

There are no conceptual barriers to producing memory continuity for the weights or activation levels of nodes in a neural network. There are a number of ways we might ensure that neural networks are run in dedicated regions of memory that retain their function over time.

- We might rework our memory manager to allow specific regions of memory to be by specific programs and fix the physical addresses of the hardware to the virtual addresses provided to the program.
- We might introduce dedicated auxiliary processors that provide memory for neural networks and give them an API that allows specific values to be stored in specific locations on those circuits.

- We might design a processor around a neural network architecture (as in Loihi), so that connection weights are stored in the same place where those weights are used.

Example: Disorder through Encryption

It is probably easier to make changes that lower the integrity of digital systems than to raise it. One way to lower integrity is to encrypt the values of nodes so that they are less well ordered.

Neural network layers utilize tensors of floats for things like activation levels and connection weights. Floats have some hints of disorder, owing to the complexities of their interpretation. We might increase this level of disorder in a couple ways.

- The easy way. Instead of storing floats, we store encrypted floats together with salts to be used in their decryption. Whenever the encrypted floats are assembled into a layer array, the data exhibits no patterns that could allow it to be deciphered. Decryption must happen at some in order to work with the standard processor registers, but this can happen at the last place in the hardware where it is possible, e.g. within a processor of a GPU. Encrypting would raise the amount of space required to store a network and would slow down processing times, but would not require any hardware changes or novel programming techniques.
- The hard way. Instead of storing floats in the standard format, we could introduce a new format for floats where the bit representations were arbitrarily related to the numbers represented. In order to understand the values represented by the floats, we would need to reason about the circuits that operate on them. This approach would require dedicated hardware and would most likely greatly reduce the FLOPS that could be computed.

Encrypting all of the values of numbers within a neural network could make it less likely that the networks were conscious. Alternatively, we could encrypt parts of a neural network to try to prevent only certain conscious states. Encrypting only the pain processing portions of an artificial simulation of a human brain might make it less likely that the system was less capable of feeling pain without preventing it from having a conscious life in general.

Avenues for Future Research

This work is exploratory. I hope to have shown that hardware and software implementation deserves a closer look and that consciousness researchers should be careful about applying their theories to contemporary neural networks without thinking more about how those networks are implemented.

The framing of changes to functionalism in terms of propriety constraints opens up further research projects.

Which other sorts of considerations matter?

I surveyed a number of considerations that occurred to me given what I know about contemporary hardware and software. I am not an expert on these issues. There are surely other similar considerations that have not occurred to me.

Processors and programming techniques are also constantly evolving. Not only are engineers always working on general ways to improve efficiency, but large scale neural networks are relatively new and are likely to continue to quickly evolve as more money is invested in AI research. The ways that they are implemented may change or may already be changing.

On the other hand, if we are interested in reducing ambiguity to avoiding consciousness, we might explore alternative processor designs that are even less like human cognition. While researchers have used the human brain for inspiration, I'm not aware of any explicit efforts to make computers less human-like.

Is integrity the right notion?

I suggested that the problem in some of the deviant cases and the unifying thread in the proposed changes to the liberal satisfaction account was integrity. This suggestion is based on an impression, not an argument, and it is quite possible that integrity is the wrong notion to focus on. Insofar as the overarching theme guides our thinking, it is important to get it right. It is worth thinking more about what might be important to implementation other than integrity, or whether we should even be looking for an overarching theme at all.

Bibliography

Aaronson, S. (2014, May 30th). Giulio Tononi and Me: A phi-nal exchange. https://scottaaronson.blog/?p=1823

Albantakis, L., Barbosa, L., Findlay, G., Grasso, M., Haun, A. M., ... & Tononi, G. (2022). Integrated information theory (IIT) 4.0: Formulating the properties of phenomenal existence in physical terms. arXiv:2212.14787.

Arvan, M., & Maley, C. J. (2022). Panpsychism and AI consciousness. Synthese, 200(3), 244.

Baars, B. J. (1988). A cognitive theory of consciousness. Cambridge University Press.

Baars, B. J. (1997). In the theatre of consciousness. Global workspace theory, a rigorous scientific theory of consciousness. Journal of consciousness Studies, 4(4), 292-309.

Balog, K. (2012). Acquaintance and the mind-body problem. In S. Gozzano and

C. Hill (eds.), New perspectives on type identity: The mental and the physical (pp 16-42). Cambridge University Press.

Block, N. (1978). Troubles with functionalism. In C. Wade Savage (ed.) Perception and cognition: Issues in the foundations of psychology (pp. 261-325). University of Minnesota Press.

Block, N. (2009). Comparing the major theories of consciousness. In M. S. Gazzaniga, E. Bizzi, L. M. Chalupa, S. T. Grafton, T. F. Heatherton, C. Koch, J. E. LeDoux, S. J. Luck, G. R. Mangan, J. A. Movshon, H. Neville, E. A. Phelps, P. Rakic, D. L. Schacter, M. Sur, & B. A. Wandell (eds.), The cognitive neurosciences (pp. 1111–1122). Massachusetts Institute of Technology.

Bourget, D. & Chalmers, D. J. (forthcoming) Philosophers on philosophy: The PhilPapers 2020 survey. Philosophers' Imprint.

Bubeck, S. Chandrasekaran, V. Eldant, R. Gehrke, J Horvitz, E. ... & Zhang, Y. (2023). Sparks of artificial general intelligence: Early experiments with gpt-4. arXiv:2303.12712.

Carruthers, P. (2004). Hop over FOR, HOT theory. In Rocco Gennaro (ed.) Higher-order theories of consciousness (115-135). John Benjamins.

Carruthers, P. (2019). Human and animal minds: The consciousness questions laid to rest. Oxford University Press.

Chalmers, D. J. (1996). Does a rock implement every finite-state automaton?. Synthese, 108, 309-333.

Chalmers, D. J. (1997). The conscious mind: In search of a fundamental theory. Oxford University Press.

Chalmers, D. J. (2018). The Metaproblem of Consciousness?. Journal of Consciousness Studies, 25(9-10), 6-61.

Chalmers, D. J. (2023). Could a large language model be conscious?. arXiv:2303.07103.

Cutter, Brian (2017). The metaphysical implications of the moral significance of consciousness. Philosophical Perspectives, 31(1), 103-130.

Davies, M., Srinivasa, N., Lin, T. H., Chinya, G., Cao, Y., ... & Wang, H. (2018). Loihi: A neuromorphic manycore processor with on-chip learning. IEEE Micro, 38(1), 82-99.

Dehaene, S., Lau, H., & Kouider, S. (2021). What is consciousness, and could machines have it?. Robotics, AI, and Humanity: Science, Ethics, and Policy, 43-56.

Dietrich, E. (1989). Semantics and the computational paradigm in cognitive psychology. Synthese, 79(1), 119-141.

Dretske, F. I. (1981). Knowledge and the flow of information. MIT Press.

Godfrey-Smith, P. (2009). Triviality arguments against functionalism. Philosophical Studies, 145, 273-295.

Godfrey-Smith, P. (2016). Mind, matter, and metabolism. The Journal of Philosophy, 113(10), 481-506.

Goodman, N. (1955). Fact, fiction, and forecast. Harvard University Press.

Graziano, M. S., Guterstam, A., Bio, B. J., & Wilterson, A. I. (2020). Toward a standard model of consciousness: Reconciling the attention schema, global workspace, higher-order thought, and illusionist theories. Cognitive Neuropsychology, 37(3-4), 155-172.

Graziano, M. S., & Webb, T. W. (2015). The attention schema theory: A mechanistic account of subjective awareness. Frontiers in Psychology, 6, 500.

Greaves, H., & MacAskill, W. (2021). The case for strong longtermism. GPI Working Paper

Hirsch, E. (2002). Quantifier variance and realism. Philosophical Issues, 12, 51-73.

Horgan, T. E., & Potrc, M. (2009). Austere realism: Contextual semantics meets minimal ontology. MIT Press.

Kammerer, F. (2022). Ethics without sentience: Facing up to the probable insignificance of phenomenal consciousness. Journal of Consciousness Studies, 29(3-4), 180-204.

Koch, C., & Tononi, G. (2008). Can machines be conscious?. IEEE Spectrum, 45(6), 55-59.

Koch, C., & Tononi, G. (2017). Special report: Can we copy the brain? – Can we quantify machine consciousness?. IEEE Spectrum, 54(6), 64-69.

Lewis, D. (1972). Psychophysical and theoretical identifications. Australasian Journal of Philosophy, 50(3), 249-258.

Lewis, D. (1983). New work for a theory of universals. Australasian journal of philosophy, 61(4), 343-377.

Loar, B. (1986). Mind and meaning. Cambridge University Press.

Loar, B. (1990). Phenomenal states. Philosophical Perspectives, 4, 81-108.

Long, R. (2023, February 18th). What to think when a language model tells you it's sentient. https://experiencemachines.substack.com/p/what-to-think-when-a-language-model

Lycan, W. G. (1996). Consciousness and experience. MIT Press.

Maley, C. J., & Piccinini, G. (2017). Of teleological functions for psychology and neuroscience. In David Kaplan (ed.) Explanation and integration in mind and brain science (pp. 236-256). Oxford University Press.

Maudlin, T. (1989). Computation and consciousness. The Journal of Philosophy, 86(8), 407-432.

Millikan, R. G. (1987). Language, thought, and other biological categories: New foundations for realism. MIT press.

Ney, A. (2008). Defining physicalism. Philosophy Compass, 3(5), 1033-1048.

Oizumi, M., Albantakis, L., & Tononi, G. (2014). From the phenomenology to the mechanisms of consciousness: integrated information theory 3.0. PLoS computational biology, 10(5).

Pautz, Adam (2017). The significance argument for the irreducibility of consciousness. Philosophical Perspectives, 31(1), 349-407.

Piccinini, G. (2010). The mind as neural software? Understanding functionalism, computationalism, and computational functionalism. Philosophy and Phenomenological Research, 81(2), 269-311.

Piccinini, G. (2015). Physical computation: A mechanistic account. Oxford University Press.

Papineau, D. (2002). Thinking about consciousness. Oxford University Press.

Prinz, J. (2003). Level-headed mysterianism and artificial experience. Journal of Consciousness Studies, 10(4-5), 111-132.

Prinz, J. J. (2005). A neurofunctional theory of consciousness. In Andrew Brook and Kathleen Atkins (eds.) Cognition and the brain: Philosophy and neuroscience movement (pp. 381-396). Cambridge University Press.

Putnam, H. (1960). Minds and Machines. In Sidney Hook (ed.), Dimensions of Minds (pp. 138-164). New York University Press.

Putnam, H. (1988). Representation and reality. MIT press.

Rea, M. C. (1998). In defense of mereological universalism. Philosophy and Phenomenological Research, 58(2), 347-360.

Rescorla, M. (2014). A theory of computational implementation. Synthese, 191, 1277-1307.

Rosenthal, D. M. (1986). Two concepts of consciousness. Philosophical Studies: An International Journal for Philosophy in the Analytic Tradition, 49(3), 329-359.

Rosenthal, D. M. (1993). Thinking that one thinks. In Davies and Humphreys (eds.) Consciousness: Psychological and philosophical methods (pp. 197-223). Basil Blackwell.

Schneider, S. (2020). How to Catch an AI Zombie. In S. Matthew Liao (ed.) Ethics of Artificial Intelligence, (pp. 439-458). Oxford University Press.

Searle, J. (2017). Biological naturalism. In The Blackwell companion to consciousness, (pp. 327-336). Wiley Blackwell.

Schwitzgebel, E., & Garza, M. (2015). A defense of the rights of artificial intelligences. Midwest Studies in Philosophy, 39, 98-119.

Shagrir, O. (2020). In defense of the semantic view of computation. Synthese 197 (9), 4083-4108.

Shagrir, O. (2021). The nature of physical computation. Oxford University Press.

Smart, J. J. (1959). Sensations and brain processes. The Philosophical Review, 68(2), 141-156.

Tiehen, J. (forthcoming). The absentminded professor. Inquiry.

Turner, R. (2018). Computational artifacts. Springer Berlin Heidelberg.

Tye, M. (1995). A representational theory of pains and their phenomenal character. Philosophical perspectives, 9, 223-239.

Udell, D. B. & Schwitzgebel, E. (2021). Susan Schneider's proposed tests for AI consciousness: Promising but flawed. Journal of Consciousness Studies, 28(5-6), 121-144.