Consciousness and the Laws of Physics

Sean M. Carroll
California Institute of Technology and Santa Fe Institute
seancarroll@gmail.com

Abstract

We have a much better understanding of physics than we do of consciousness. I consider ways in which intrinsically mental aspects of fundamental ontology might induce modifications of the known laws of physics, or whether they could be relevant to accounting for consciousness if no such modifications exist. I suggest that our current knowledge of physics should make us skeptical of hypothetical modifications of the known rules, and that without such modifications it's hard to imagine how intrinsically mental aspects could play a useful explanatory role.

Draft version of a paper submitted to *Journal of Consciousness Studies*, special issue responding to Philip Goff's *Galileo's Error: Foundations for a New Science of Consciousness*.

Introduction

We don't fully understand consciousness. That's hardly surprising. The human brain, which is at least somewhat involved in consciousness, contains roughly 100 billion neurons and 700 trillion synaptic connections. It is arguably the most complex structure in the known universe. Even as neuroscience makes impressive advances in understanding the brain, it seems prudent to anticipate that we have a number of conceptual and technical breakthroughs yet to come that could bear in important ways on the question of consciousness.

We do, on the other hand, understand the basic laws of physics governing the stuff of which brains are made. They take the form of an effective quantum field theory describing a particular collection of matter particles interacting via force fields. There is certainly much of physics remaining to be discovered, but in the specific regime covering the particles and forces that make up human beings and their environments, we have good reason to think that all of the ingredients and their dynamics are understood to extremely high precision (Carroll 2021a). Modern physics, in other words, provides evidence for what philosophers call "causal closure of the physical": physical events have purely physical causes (Loewer 1995, Papineau 1995), at least in the regime relevant to human life. Without dramatically upending our understanding of quantum field theory, there is no room for any new influences that could bear on the problem of consciousness.

Given this situation, it might seem surprising to a disinterested observer to learn that anyone would argue that the best route toward understanding consciousness involves augmenting or altering the ontology suggested by fundamental physics. To start with the least-well-understood aspects of reality and draw sweeping conclusions about the best-understood aspects is arguably the tail wagging the dog. When we can't remember where we put our car keys, we don't typically respond by going out and buying a new car.

Nevertheless, a prominent strain in the philosophy of consciousness proposes to do just that (Chalmers 1996, Goff 2017, 2019). The justification for such a radical move is that there will be something *qualitatively* missing in any account of consciousness based purely on physical ontology as we currently understand it. This perspective arises from a conviction that physics can explain behavior, but not the first-person experiences characteristic of human consciousness; that physics may account for the dynamics of the stuff in the universe, but it doesn't illuminate the intrinsic nature of that stuff.

In this paper I support the idea that physics is in such good shape that the most promising strategy for trying to understand consciousness is as a (weakly) emergent phenomenon that leaves physical ontology untouched, rather than trying to extend or elaborate that ontology with specifically mental aspects (cf. Moran 2021; for a contrary view see Smolin and Verde 2021). After reviewing the Core Theory and our reasons for being confident in its accuracy, I will discuss what it would mean to modify it, either directly in the dynamics or by adding additional ontological features. I further argue that any approach in which mental aspects leave physical behavior unchanged are self-undermining and fall short of accounting for consciousness. It is always possible that contemporary physics is inadequate and in need of modification, but a close examination highlights the difficulty of doing so in a rigorous and convincing way.

The Physics Underlying Everyday Life

The history of physics is rife with premature claims that we are close to understanding everything. These unfortunate episodes should not lead us to forget that we do understand some things.

Science often employs multiple vocabularies or theories for describing the same physical situation, often at different degrees of focus or coarse-graining. These are often called "levels," although strictly speaking they need not be arranged hierarchically. Within any level, we can specify the domain of circumstances in which a particular theory is applicable. The claim here is that there is one level of description – that of effective quantum field theory – and a well-defined regime – interaction energies below certain thresholds, broad enough to include every situation encountered in ordinary human life – where we have very good reasons to believe we know precisely what is going on.

A quantum field theory is, unsurprisingly, a quantum theory of fields. The fundamental ontology of any quantum theory is specified by a "quantum state" or "wave function," expressed mathematically as a vector in an abstract Hilbert space (Carroll 2021b). In a quantum field theory, that state can be thought of as being constructed from possible configurations of fields that take on values at each point in spacetime.

Fortunately, the details of this formalism are not necessary for our present purposes. Once we quantize the fields, appropriate configurations – essentially, low-lying energy states – can be interpreted as collections of interacting particles. These circumstances are more than broad enough to encompass human beings and their environments. Thus, we can think of people and the objects around them as configurations of certain particles. In particular, human beings are made of atoms; those atoms are made of protons, neutrons, and electrons; the protons and neutrons are made of quarks and gluons. These particles interact through gravitation, electromagnetism, and the nuclear forces, and get mass from a background Higgs field.

The dynamics of these particles and forces are governed by an effective quantum field theory known as the "Core Theory," consisting of both the Standard Model of particle physics and the weak-field limit of general relativity (Wilczek 2015). This theory is not the ultimate theory of everything, nor is it intended to be. The world might not be described by a quantum field theory at the deepest level; that description might emerge from a more fundamental set of degrees of freedom and dynamical laws. And the Core Theory is certainly not supposed to cover every circumstance – dark matter and the Big Bang, to name some obvious examples, are not included. But we have excellent reasons to believe that the entirety of the "everyday life regime" supervenes on the ontology and dynamics of this theory (Carroll 2021a). If there is a more fundamental level, its properties are irrelevant to the autonomous dynamics of the Core Theory. And if there are additional particles and forces, they interact too weakly with the known fields to exert any influence on human behavior; otherwise they would have already been detected in experiments.

Our confidence in this picture derives from the fact that quantum field theories are the practically unique way to satisfy the general principles of quantum mechanics and relativity; from symmetries ensuring that any unobserved fields must be too weakly-interacting with ordinary matter to be relevant for everyday-life dynamics; and the property of effective field theories that the dynamics themselves are fully determined in terms of a very small number of parameters. We can't know for certain that the Core Theory suffices to correctly describe the behavior of the particles and fields making up human beings, no matter how good our arguments become, but any proposed modification of this theory should be held to a very high standard indeed. Just as with any hypothetical new physical model, it should be quantitative and precise, detailing exactly how the explicit dynamics of the Core Theory are meant to be modified, and how such modifications are consistent (or not) with features such as unitarity, locality, symmetries, and conservation laws, not to mention experiments.

Domains of Applicability

In the context of the relationship between consciousness and the laws of physics, it is worth being a bit more explicit about how we specify the "domain of applicability" of a theory (Carroll 2016). The general idea is that there is a set of physical situations in which the predictions of the theory are meant to be accurate, with no claims being made for situations outside that set. Newton's theory of gravity does not correctly describe the emission of gravitational waves by orbiting black holes, but is perfectly adequate for sending a rocket to the Moon. In this case the relevant domain of applicability consists of situations when the gravitational field is weak and all relevant objects are moving slowly compared to the speed of light. In other circumstances, the Newtonian limit doesn't apply, and we must use Einstein's theory of general relativity.

The empirical foundation of the Core Theory has been established through a line of experimental and observational results stretching back to Faraday, Rutherford, and many others. But the most precise constraints come from modern-day particle colliders, which typically measure the results of scattering individual particles off of each other. One might sensibly wonder whether results from such a paradigmatically reductionist setting can be straightforwardly extrapolated to something as complex as a human brain, which contains roughly 10^{27} particles. Perhaps brains are just not within the domain of applicability of the Core Theory.

If we accept the basic framework of effective quantum field theory, this concern is unfounded; everything that happens inside biological organisms here on Earth is unambiguously within the purview of the Core Theory. Its domain of applicability is bounded by two criteria. The first is that gravity must be weak, so that we can treat the gravitational field as an ordinary quantum field, sidestepping subtleties of horizons and Hawking radiation. "Weak" is a relative term, and in this case means "the gravitational potential is much smaller than one." In practice, this means "we are nowhere near a black hole." This criterion is easily met by everything we know of in the Solar System, human brains included.

The other criterion comes from effective field theory. The modifier "effective" indicates that the domain of applicability of the theory is specified in terms of energies – in particular, the amount of energy transferred between particles when they interact. An effective field theory is meant to be accurate when energy transfers remain lower than some explicit cutoff. In the case of the Core Theory, experiments have established its accuracy at energy transfers of up to 10^{11} electron volts. Electro-chemical reactions inside biological organisms, meanwhile, happen at less than 10^{2} eV. Shrinking the domain of applicability of the Core Theory while remaining within the framework of effective quantum field theory requires a mistake in our current understanding by a factor of over a billion, which seems implausible.

The effective field theory paradigm also features very specific properties of the field dynamics: they are local (interacting only with other fields at the same spacetime point), and governed by a simple and inflexible set of equations. (In the technical jargon, by "relevant" and "marginal" operators, with other "irrelevant" operators living up to their name.) So below, when I refer to "within the effective-field-theory paradigm," this is what is meant: a theory of quantum fields, evolving under the appropriate simple dynamical equations, applicable in circumstances where gravity is weak and interactions feature energy transfers below the cutoff.

Within its domain of applicability, the Core Theory is what we might label *causally comprehensive*. If we give a complete specification of the quantum state of the Core Theory fields within that regime, there is a specific equation that unambiguously predicts how it will evolve over time. This equation is sufficient to describe everything human beings generally do, unless they jump into a black hole or stick their hand inside the beam of a high-energy particle accelerator. There are no ambiguities or loose ends. The fact that brains are big, complex things is irrelevant. The Core Theory makes specific predictions for how any particular brain will behave; our choice is to either accept that prediction, or modify the theory in some way. There is no third alternative (Aristotle 2002).

Ontology and Dynamics

Despite the extraordinary empirical success of the Core Theory, and the fact that human beings and their brains are made out of particles interacting within its domain of applicability, there is a lingering worry that no physicalist picture is up to the task of accounting for consciousness, even as some higher-level weakly-emergent phenomenon. There are various ways of expressing this concern: conscious experiences are inherently first-person and subjective; merely physical objects cannot feel what it is to be like something; describing the behavior and functions of objects does not explain their intrinsic nature; and others. See Goff (2019) for an overview.

One common reaction to these concerns is to contemplate modifications of the underlying ontology suggested by modern physics: to suggest that a quantum state built upon interacting fields obeying strict equations of motion is incapable in principle of accounting for consciousness, and that we instead need to add specifically mental aspects to our description of reality. We may contemplate ontological modifications as dramatic as substance dualism, in which an immaterial mind is distinct from the physical body but interacts with it, or idealism, in which the physical world is a kind of projection of a fundamentally mental reality. In this paper I will focus on more subtle approaches, in which mental aspects or properties are related to, but augment, the basic physical reality. Approaches under this umbrella include property dualism, which posits distinct mental properties in addition to physical properties (Chalmers 2003b); Russellian monism, which posits both physical and mental aspects belonging to a single underlying set of properties (Russell 1927, Chalmers 1996, Strawson 2006, Goff 2017); and other forms of panpsychism, epiphenomenalism and related approaches (Papineau 2020). For

convenience I will refer to any new ontological features as "mental aspects," which is meant to include potentially autonomous properties as well as intrinsic qualities that might supervene on the physical situation.

Any such approach must specify whether, and how, it modifies the dynamics of the theory as well as the ontology. In our conventional understanding, consciousness exerts an important influence on behavior: I can have a conscious experience and talk about it. (Admittedly, highly trained philosophers are reported to be able to imagine removing consciousness from a being without affecting its behavior in any way.) Observed human behavior can be traced to electrical and chemical signals in our brains and nervous systems. Explicitly mental aspects of ontology could affect this behavior by, for example, influencing the rates of chemical reactions, or the strength of electromagnetic forces, or the probability of certain quantum outcomes. For this paper we are imagining that such effects do not arise from simply pushing around the fields (as in substance dualism), but from potentially altering the properties of the fields themselves.

In what follows we will examine different kinds of relationship that a theory of consciousness might have to the physical dynamics of the Core Theory, as well as the possibility that there is no relationship at all. Our goal is not to comprehensively catalogue the possibilities, but just to highlight some of the challenges faced by any approach that aspires to explain consciousness by adding mental aspects to the fundamental ontology of the world.

Consciousness and Quantum Mechanics

Quantum field theory is a subset of quantum mechanics. Like any quantum theory (and in contrast with classical theories), the dynamics of the Core Theory come in two parts. There is a law of evolution that describes how an undisturbed quantum state evolves deterministically over time, referred to as the *unitary* dynamics. The law can be cast as a version of the Schrödinger equation or in a number of equivalent formulations. The other part takes the form of a probabilistic algorithm expressing how the wave function responds to being measured. Operationally, a measured wave function "collapses" onto a state with a definite value of the quantity being measured, so we label this the *collapse* dynamics. Collapse introduces a stochastic element, with the probability of different outcomes being related to the original wave function by the Born Rule. There are therefore two broad strategies one could contemplate for modifying the dynamics of the Core Theory: altering its unitary dynamics, or its collapse dynamics.

The questions of what precisely constitutes a quantum measurement, what happens when one is performed, and what is the correct ontology describing quantum systems, have remained controversial. Contrary to some claims in the popular literature, however, the quantum measurement problem does not by itself provide evidence against physicalism. There are perfectly physical models that account for quantum phenomena without introducing any specifically mental aspects or preferred roles for observers,

including Many-Worlds, pilot-wave theories, and objective-collapse models. Other models do put "agents" front and center, casting the quantum wave function as a representation of the agent's knowledge of a system, while still remaining resolutely physicalist. We don't need to distinguish between these competing theories for our present purposes; see overviews by Norsen (2017) and Maudlin (2019), and cf. Rovelli (2021).

According to textbook quantum mechanics, when one measures a quantum observable such as position or momentum or spin, only certain specified outcomes ("eigenvalues") can be obtained. To each possible outcome, the quantum state assigns a complex number, the amplitude. The Born Rule states that the probability of obtaining that outcome is the modulus-squared of the corresponding amplitude. Importantly, there is no hidden structure within this rule; once we know the amplitudes, experimental outcomes are truly randomly chosen from the appropriate probability distribution.

The Born Rule has thus far passed experimental tests (Lin et al. 2017), but the fact that both consciousness and quantum measurement remain mysterious makes it tempting to imagine that there is a connection. What we are interested in here is not the prospect that consciousness causes wave function collapse (Wigner 1961; Stapp 2001; Chalmers and McQueen 2014), but that somehow wave functions collapse in just the right way to account for consciousness. Penrose and Hameroff have developed an approach in which wave functions collapse when certain physical criteria are met, which they argue can explain aspects of human cognition (Penrose 1989, Penrose and Hameroff 2011, Penrose 2014). However, although this program is often described as an approach to "consciousness," it does not attempt to answer the qualitative questions of first-person experience any differently than any other purely physical account. Similarly, quantum entanglement may play a role in cognition (Fischer 2015), but this is a matter of information processing, without any special connection to qualitative experience.

If one were interested in allowing mental aspects to affect the probability of quantum measurement outcomes, presumably that could be done. The Born Rule states that the probability of obtaining an outcome a is given by $p(a) = |\psi_a|^2$, where ψ_a is the component of the wave function corresponding to that outcome. We could imagine a new rule

$$p(a) = f(\psi_a, M_a), \tag{1}$$

where M_a represents some novel mental aspect of the situation. This modified Born Rule might affect the rate of certain chemical reactions inside a human brain, thereby allowing mental aspects of consciousness to influence our physical behavior, without showing up in experiments performed with non-conscious equipment.

Of course, such a rule for wave-function collapse represents a wild modification of conventional physics, not merely a loophole within it. A respectable theory along these lines would include a specification of what the mental aspects M_a are, an understanding

of their independent dynamics, and an explicit form of the new rule (1). All of these are possible to contemplate, but they remind us of the high standards to which any modified laws of fundamental physics should be held.

Furthermore, if one were convinced that purely physical ontologies are incapable in principle of accounting for the qualitative features of consciousness, the process of wave function collapse does not offer any unique opportunities. Regardless of when and how wave functions collapse, at the end of the day they are still wave functions. One could conceivably take a dualist approach, positing that a separate mental realm interacted with ordinary matter by triggering (or delaying) collapse. In that case, wave function collapse would play the role of a modern version of Descartes's pineal gland, mediating the interaction of mental and physical realms. If instead we think in terms of novel mental properties affecting chemical reaction rates, there would be no relevant difference between modifying the collapse dynamics and the unitary dynamics.

Consciousness and Quantum Field Theory

We turn next to the unitary dynamics. As discussed above, if we stay entirely within the effective-field-theory paradigm, both for the Core Theory fields and potentially new dynamical elements, there is no room for modifying the dynamics in ways that would be relevant for human behavior while remaining compatible with experimental constraints. Any new fields that would be relevant for what goes on in the human brain would have been discovered long ago.

We can nevertheless imagine that new mental aspects influence the quantum fields of the Core Theory, without themselves obeying the rules of quantum field theory. To see how that might work, it is useful to look at one part of the Core Theory: quantum electrodynamics, the theory of charged particles (including electrons, protons, and even atomic nuclei) with electromagnetic fields. To the extent that gravity, nuclear reactions, and radioactive decays can be ignored (all of which are presumably irrelevant for questions of consciousness), this is enough to include all of the physics relevant for human biology. The unitary dynamics can be summarized in a one-line equation:

$$A = \int_{k<\Lambda} [DA][D\psi] \exp\left\{ i \int d^4x \left[\sum_n \bar{\psi}_n (i\gamma^\mu \partial_\mu + q_n \gamma^\mu A_\mu - m_n)\psi_n - \frac{1}{4} F_{\mu\nu} F^{\mu\nu} \right] \right\} \quad (2)$$

We don't need to dive into this equation in detail, but a few points are worth highlighting. The expression tells us how a quantum state (describing, for example, a human brain) consisting of charged particles ψ_n and electromagnetic fields $F_{\mu\nu}$ evolves from a given initial state to a final one. It is entirely deterministic and causally comprehensive; indeterminism only comes from the non-unitary collapse dynamics. The expression in square brackets is the *Lagrangian*, which encodes the properties of different kinds of particles. The notation $\int d^4 x$ indicates that the Lagrangian is integrated over spacetime. This reflects the locality of the unitary dynamics: fields only interact with other fields

(and with themselves) at the same point in spacetime. We multiply the integral of the Lagrange density by $i = \sqrt{-1}$, exponentiate it, and integrate the result over all possible configurations of the fields to obtain the transition amplitude. The notation $k<\Lambda$ indicates that this is meant to be an effective theory, applicable below the cutoff energy Λ. See (Carroll 2016) for further elaboration. Any given approach to consciousness will either modify this equation, or it won't.

The properties specified by the Lagrangian include the masses of the particles m_n, as well as the parameters characterizing their interactions, such as electric charges q_n. The strength and rate of electrochemical processes, including those in human brains and bodies, are calculable in terms of these parameters.

An obvious way that mental aspects could modify physical dynamics is for them to affect the values of these parameters, which would in turn affect the rate of processes in the brain. Given some physical/mental situation **S**, we could imagine context-dependent changes in the values of the physical constants that govern Core Theory dynamics, of the form

$$m_n \rightarrow m_n(\mathbf{S}) \ , \quad q_n \rightarrow q_n(\mathbf{S}). \tag{3}$$

If **S** included mental aspects of our ontology, this would be a mechanism by which those aspects could affect human behavior, such as our testimony concerning our introspective experiences.

As in the case of the Born Rule, we are welcome to contemplate mentally-induced modifications of particle-physics parameters, but a number of questions present themselves. What precisely is meant by the situation **S**, and what kind of dynamics does it have? Naively, changes in the masses or charges of particles would lead directly to violations of conservation of energy and momentum. Are these compensated by transfers of energy between conventional matter and a "mental sector"? These questions are potentially answerable, but they highlight the challenges faced by any proposed theory of this form.

Mental Degrees of Freedom

Panpsychists sometimes analogize consciousness to electric charge, as a property that inheres in appropriate fundamental particles. There are at least two severe limitations to this analogy. First, electric charge is a paradigmatic example of a property with dynamical consequences; placed in an electric field, particles with opposite charges move in opposite directions. In the case where we imagine that the properties associated with consciousness have no dynamical consequences, it is not clear what the analogy is supposed to illuminate. Second, charge is conserved. An elementary particle has a single, unchanging value of charge throughout its existence, whereas it is generally supposed that conscious states can take on different values.

We should therefore distinguish between two alternatives: that hypothetical new mental aspects of our ontology supervene on the physical situation, or that there are independent "mental degrees of freedom" that are not determined by the physical situation. In panpsychist terms, these correspond to the possibility that any particular electron has a definite value of consciousness, versus the idea that any given electron might have multiple conscious states (perhaps even a continuum thereof).

If the new mental aspects of our ontology supervene on the physical aspects, as far as physical behavior is concerned this is indistinguishable from not introducing mental states at all. Consider some causal chain $P_i \rightarrow M_i \rightarrow B_i$, where P_i is some physical state, M_i is some mental state, B_i is some behavior, and "$\rightarrow$" stands for "inevitably leads to." If we think of mental aspects as primary, but nevertheless supervening on the physical situation, we could write $M_i(P_i) \rightarrow B_i$. Either case is functionally equivalent to the shorter chain $P_i \rightarrow B_i$. (This is analogous to "integrating out" heavy or non-dynamical degrees of freedom in quantum field theory.) For all intents and purposes this is equivalent to positing that mental aspects have no effect on physical behavior, a possibility we will consider later in the paper.

Turn instead to the alternative where the same physical configuration might be associated with different mental degrees of freedom. These aspects would be roughly analogous to the spin of an electron, which is some combination of "spin-up" and "spin-down" for any given particle, but can be specified independently of the particle's position.

The idea of new independent mental degrees of freedom runs into serious trouble with the framework of quantum field theory. In conventional field theory, the existence of new degrees of freedom quantitatively affects processes that rely on quantum fluctuations, in which each property value represents a separate contribution that should be added together (e.g. we "sum over spins" in a scattering calculation). But we know empirically how many degrees of freedom actual electrons have – two spin states for the electron, and another two for its antiparticle, the positron. If electrons could also be found in both "happy" and "sad" states, it would have an unmistakable impact on their scattering rates, in flagrant contradiction with experiment.

Our allowed alternatives are to posit that all electrons have the same conscious state – let us be generous and imagine they are all happy – in which case there is effectively no dynamical impact, or that mental degrees of freedom are somehow not like physical ones. In the latter case, we are left with the question of what mental degrees of freedom *are* like, if they are not like physical ones. We can avoid conflict with what we know, but only at the expense of pushing our ideas further away from clarity and tangibility.

Strong Emergence

One approach to the relationship between consciousness and physics is to appeal to *strong emergence* – the idea that legitimately new behaviors arise in collective phenomena that cannot be derived in terms of the individual behaviors of constituent parts of the system. Strong emergence is sometimes invoked as a way to allow for specifically mental causal powers (O'Connor and Wong 2005). It is worth spending a moment on the relationship between strong emergence and the underlying framework of effective quantum field theory – namely, they are entirely incompatible.

As discussed above, the Core Theory provides a comprehensive specification of the quantum-field dynamics within its domain of applicability, which includes any processes between known particles with energy transfers less than a hundred billion electron volts. The field equations are precisely local: the unitary dynamics of each field at any one spacetime point are influenced only on the values and derivatives of the other fields at the same point, and not directly by what is happening elsewhere. Electrons and other particles obey the same equations whether they are inside a rock or inside a human brain.

The strong emergentist must therefore deviate from the paradigm of effective field theory entirely, while maintaining the empirical successes of the Core Theory. The most direct way to do this would be to postulate a new restriction on the domain of applicability that is not given in terms of energy transfers in particle interactions, but on some explicitly macroscopic criterion. For example, one could hypothesize that quantum field theory breaks down when the number of particle excitations in a region surpasses a certain number, or (probably more relevant to consciousness) when the configuration of such particles reaches a certain quantifiable degree of complexity or information-processing capacity. The effective masses and couplings of elementary particles might, for example, be modified as in (3), where the situation **S** could involve a quantitative measure of consciousness from an approach such as Integrated Information Theory (Tononi et al. 2016).

One is, of course, free to contemplate whatever extravagant deviations from contemporary physics as one likes. Particle-physics experiments typically examine the interactions of just a few particles at a time, so new physical laws that only kick in for complex agglomerations of particles are not necessarily ruled out by data we currently have. It's worth noting, however, how profound a departure such laws would represent. The most fundamental principle of quantum field theory is locality: fields at any one point in spacetime are only influenced by the values and derivatives of other fields at that same point, not the behavior of fields at other points. Modifying the dynamical equations in ways that were sensitive to the complexity of a configuration of surrounding particles would represent a dramatic overthrow of this principle.

Moreover, based on purely physical grounds rather than consciousness-based motivations, our expectation that the laws of quantum field theory might break down in biological organisms would be very low indeed. To we macroscopic people, the 10^{27} particles in a human brain seems like a lot, certainly far greater than the number

physicists typically collide in high-energy accelerators. But the density of those particles is very low by particle-physics standards. To be conservative, we might take as a standard length scale the Compton wavelength of the electron, $\lambda_e = 2 \times 10^{-10}$ cm. The volume of a human brain is 1,260 cubic centimeters, or about 10^{32} cubic Compton wavelengths. The number of particles is therefore less than 10^{-5} per standard volume. (It would be enormously smaller had we used the Compton wavelength of the proton or neutron, not to mention the Planck length.) From the point of view of particle physics, a brain is not a densely packed system; indeed, it's practically empty space. There is no physical rationale for expecting the dynamics of the Core Theory to break down in such an environment, regardless of how complex the overall situation is. For any particular electron or nucleus, almost all of the rest of the brain is so far away as to be essentially irrelevant.

This is not to say that the concept of strong emergence, and the related phenomenon of "downward causation," might not be relevant in other contexts, where the "micro" theory is something other than elementary particles. If a complex system consists of a collection of smaller systems that are themselves complex, a purely local theory might not suffice, and the best description of the overall dynamics could conceivably involve microphysical dynamics that depend on macrophysical contexts in interesting ways (Flack 2017). In the phenomenon of "quorum sensing," for example, gene expression in bacteria is affected by their overall population density (Miller and Bassler 2001). In such cases, the subsystems themselves are extended objects with nontrivial internal dynamics, which can effectively measure and record information about their environment. Because of that feature, the criterion of locality is significantly less severe.

Quantum field theory is a very different situation, where subsystems (elementary particles) have no internal structure. In that case, the locality of interactions is exact; what matters to the dynamics of a field at each point is only the other fields at that same point, not anything elsewhere. Any new context-dependent behavior departing from the predictions of (2) would be a violation of our expectations from effective quantum field theory, not a supplement to them.

Passive Mentalism and Zombies

We finally turn to the possibility that there exist purely mental aspects of the basic ontology of the world that have no effect at all on physical dynamics. We can label this alternative "passive mentalism."

Passive mentalism opens the door to contemplating the possibility of philosophical *zombies*: creatures that have exactly the same physical behavior as ordinary human beings, but lack any inner conscious experiences (Campbell 1970, Chalmers 1996, Kirk 2005). In particular, we can imagine two kinds of possible worlds, both of which exactly obey the dynamics of the Core Theory: one of which is a purely physical one without any additional aspects, and the other of which includes those features that make true conscious experience possible.

It is tempting to think of zombies as a kind of automata, going through life without feeling or affect. Maybe they would come across as generally a little deadpan or at least even-keeled. But that would be wrong; by hypothesis, zombies behave *exactly* like conscious human beings. They laugh, cry, lament when their hearts are broken, and cheer when their team wins. If we take ourselves to live in the possible world that does include conscious experiences, there is a zombie possible world with exactly the same set of people have exactly the same discussions and interactions. It contains a zombie Philip Goff who wrote precisely the text of *Galileo's Error*, and a zombie Sean Carroll who wrote precisely this article; those poor souls (as it were) just don't really possess any conscious experiences, despite what they may say about the matter when they are asked. In zombie world, there are plenty of people who sincerely testify that they are experiencing the redness of red, but those utterances have absolutely no connection with an actual experience of redness.

The zombie thought experiment is sometimes developed into an argument against physicalism, which goes something like this:

- Zombies are conceivable.
- Conceivability implies possibility.
- If zombies are possible, consciousness is not physical.
- Therefore, consciousness is not physical.

The idea behind the third premise is that the conceivability of zombies implies that whatever consciousness is, if we can in principle imagine the same physical situation with or without conscious experiences, consciousness can't be reducible to physical behaviors. It's a valid argument, which has engendered a great deal of discussion about the nature of conceivability and its relation to possibility (Chalmers 2002), which I won't delve into here.

One might reasonably suggest that any construal of consciousness in which conscious experience has no causal impact on human behavior is not capturing what we really care about. We generally think that we not only experience things, but that we react to those experiences, talk about them, and so on. But such reactions and speech acts are behaviors, which for now we are assuming are entirely described by the Core Theory. The passive-mentalism notion of consciousness floats freely from all that, leaving no imprint on our actions in the world. Let us put aside these worries for the time being, and think more carefully about the physicalist alternative.

Is Physicalism Conceivable?

The zombie argument against physicalism gains leverage from the idea that zombies are conceivable. There is an obvious related question that garners less attention: is *physicalism* conceivable? By "physicalism" here I don't simply mean a world with only physical

properties, but specifically physicalism about consciousness. Can we conceive of a world where the ontology consists of nothing other than some notion of physical "stuff" (or the specific quantum fields in the Core Theory) without any inherently mental aspects, but which nevertheless accounts for consciousness as we experience it?

Asking this question invites us to contemplate more carefully what a physicalist picture would entail, without necessarily going into the specific way in which actual human consciousness relates to functions in the brain. The basic idea is that of *weak emergence*. There exists a description of the system at a fine-grained level – in this case that of quantum field theory – that is complete and accurate on its own terms. And there is coarse-grained level of description, complete and more or less accurate on its own terms, which might involve entirely distinct ontological categories from those of the fine-grained theory. But the descriptions are compatible, in the sense that there is a well-defined map (typically many-to-one) from appropriate states in the fine-grained theory to those in the coarse-grained theory. In this picture, there is a set of processes undergone by certain states in the microphysical ontology, all of which correspond in the macroscopic description to "a person experiencing the redness of red."

Goff (2019) refers to the "brute identity theory," according to which conscious states simply *are* states of the brain. This is not exactly how I would put it, and while the difference is slight it might be worth drawing out. It's not that conscious states "are" states of the brain; it's that certain states of the brain *correspond to* certain conscious states. This terminological nuance highlights the idea that there are multiple vocabularies we can use to describe the same underlying physical situation, each of which can stand on its own feet without reference to the others (we could profitably talk about conscious states before we knew anything about neurons or electrons), but which are compatible with each other in the sense that there are well-defined maps between them (even if the current state of human knowledge is unable to specify what those maps are).

Emergent concepts, even in this weakly-emergent sense, capture something true and real; they are neither illusory nor arbitrary (Dennett 1991, Carroll 2016). Conscious states, in particular, describe real phenomena, and play causal and predictive roles in the world. If we are told "she was conscious of being watched," we immediately have somewhat reliable expectations for how she will behave in response to future events. Subsequent identification of corresponding microstates in a more fine-grained description doesn't affect the reality of the emergent concepts in an appropriate domain of applicability. And the recognition of these causal powers doesn't imply that emergent concepts need to be accounted for by enriching the underlying ontology; rather, it counts strongly against any purported explanation that separates the concepts from their causal role.[1]

---

[1] Philosophers such as Putnam (1975) and Kripke (1980) have used examples of weak emergence – for example the relationship between "water" and "H_2O" – to argue for the existence of *a posteriori* necessities. The point I am making here is weaker, and independent of that discussion; simply that the concept of water plays a sensible role in a higher-level ontology whether or not we know that it corresponds to H_2O at the microphysical level.

If we grant that physicalist consciousness is conceivable, it follows that appropriate behavior of physical matter, without any specifically mental aspects, would amount to what we think of as "consciousness." In that case, it follows that zombies are *not* conceivable after all, or at least they are not conceivable in a way that implies possibility. This point has been made in different ways in the literature; see Balog (1999, 2012), Frankish (2007), Brown (2010), and Campbell et al. (2017).

When faced with a purported zombie – who acted in every way conscious, sincerely assured us that they had conscious experiences all the time, and behaved accordingly – someone who accepted the possibility of emergent physicalist consciousness would readily categorize such a creature as, in fact, conscious. (To that person, "consciousness" is a label we attach to such creatures, as a useful concept in our emergent ontology.) We could therefore run the following argument:

- Physicalist consciousness is conceivable.
- Conceivability implies possibility.
- If physicalist consciousness is possible, zombies are impossible.
- Therefore, zombies are impossible.

Contrasting this with our previous zombie deduction, we see that the zombie argument is not by itself an argument against physicalism. It is a way of clarifying the idea that one will judge zombies to be conceivable if and only if one judges physicalist consciousness to be inconceivable. Neither argument is likely to have much persuasive power among people who are already satisfied with the other side.

The zombie thought experiment should, if anything, push our credences more in the direction of physicalism. Given the existence of physically identical possible worlds with and without consciousness, an agent will reasonably want to know which world they find themselves in. The answer typically comes down to our first-person experiences, and the idea we can interrogate the reality of our own conscious experiences through introspection to decide that our experiences are real. But that's just what a zombie would decide – at least, it's what they would say they had decided, were we to ask them. Unless our thoughts are completely uncorrelated with what physically happens in our brains, the correct conclusion of the zombie scenario is that introspection about our conscious experiences is unreliable. But such introspection is the entire reason we felt the need to develop non-physicalist accounts of consciousness in the first place. In that sense, the zombie argument against physicalism is self-undermining. (For a contrary view see Chalmers 2003a.)

The zombie scenario posits that we can conceive of persons who behave exactly as we do, but who lack inner experience. To pull off this trick, it is necessary to invoke strategies to completely sequester consciousness from anything that people say or do. The cost is that what ends up being described is not what we usually think of a person at all. Within a

passive-mentalist approach, a person is not an integrated whole of phenomenal experience and behavior. Rather, they are effectively a zombie carrying around a sealed box labeled "mental stuff." And their physical selves will never know what's inside the box. Were they allowed to look inside and become aware of the mental aspects of their existence, the knowledge they gained would inevitably affect their behavior, which is against the rules. The fact that passive mentalism admits the conceivability of zombies implies that what it purports to explain is not consciousness as we know it.

Conclusions

The temptation to augment the ontology of the world with specifically mental aspects stems from a conviction that describing the mere behavior or function of matter cannot be sufficient to account for consciousness or innate nature. As characterized by Levine (1983), there seems to be an "explanatory gap" between physical states and conscious experiences. Physicalism posits that a conscious experience is an emergent phenomenon that arises in higher-level models of the same underlying processes described by physics. To a panpsychist, as Goff (2019) says about the brute identity theory, this "is very unsatisfying." Arguably it is this "satisfaction gap," more than any explanatory or ontological gap, that prompts the introduction of intrinsically mental concepts and categories into fundamental ontology.

Any discussion of mental aspects of ontology must specify one of two alternatives: changing the known laws of physics, or positing that these aspects exert no causal influence over physical behavior. We cannot rule out the first option either through pure thought or by appeal to existing experimental data, but we can ask that any modification of the Core Theory be held to the same standards of rigor and specificity that physics itself is held to. The point of expressions like (1) and (3) is not that mentally-induced modifications of physical parameters are impossible, but that a promising theory of consciousness should be specific about how they are to be implemented.

The passive mentalism option, where mental aspects have no impact on physical behavior, seems even less promising. "Behavior" should not be underrated; the behavior of physical matter is literally "what happens in the universe." Crying at a funeral is behavior, as is asking someone to marry you, as is arguing about consciousness. No compelling account of consciousness can attribute a central explanatory role to metaphysical ingredients that have no influence on these kinds of behaviors.

We don't know everything there is to know about the laws of physics, and there is always the possibility of a surprise. But the solidity of our confidence in the Core Theory within its domain of applicability stands in stark contrast with our fuzzy grasp of the nature of consciousness. The most promising route to understanding consciousness is likely to involve further neuroscientific insights and a more refined philosophical understanding of weak emergence, rather than rethinking the fundamental nature of reality.

Acknowledgements

It is a pleasure to thank Philip Goff for stimulating this work and for enlightening discussions. I'd also like to thank Jenann Ismael, Barry Loewer, Alex Moran, David Papineau, Alex Rosenberg, and Eric Schwitzgebel for helpful comments and pointers to references. This work is supported in part by the Foundational Questions Institute.

References

Aristotle (2002). *Metaphysics,* Book 3, 996b. Translated by Sachs, Joe (2nd ed.). Santa Fe, N.M.: Green Lion Press.

Balog, K. (1999). "Conceivability, Possibility, and the Mind-Body Problem," *The Philosophical Review*, 108: 497-528.

Balog, K. (2012). "In Defense of the Phenomenal Concept Strategy," *Philosophy and Phenomenological Research*, 84: 1-23.

Brown, R. (2010). "Deprioritizing the A Priori Arguments against Physicalism," *Journal of Consciousness Studies*, 17 (3-4): 47-69.

Campbell, D., J. Copeland and Z-R Deng 2017. "The Inconceivable Popularity of Conceivability Arguments," *The Philosophical Quarterly*, 67: 223—240.

Campbell, K. (1970). *Body and Mind*. London: Macmillan.

Carroll, S.M. (2016). *The Big Picture: On the Origins of Life, Meaning, and the Universe Itself.* New York: Dutton.

Carroll, S.M. (2021a). "The Quantum Field Theory on Which the Everyday World Supervenes." Submitted to *Levels of Reality: A Scientific and Metaphysical Investigation* (Jerusalem Studies in Philosophy and History of Science), eds. O. Shenker, M. Hemmo, S. Iannidis, and G. Vishne. https://arxiv.org/abs/2101.07884.

Carroll, S.M. (2021b). "Reality as a Vector in Hilbert Space." Submitted to *Quantum Mechanics and Fundamentality: Naturalizing Quantum Theory Between Scientific Realism and Ontological Indeterminacy*, ed. V. Allori. https://arxiv.org/abs/2103.09780.

Chalmers, D.J. (1996). *The Conscious Mind: In Search of a Fundamental Theory*. New York and Oxford: Oxford University Press.

Chalmers, D.J. (2002). "Does Conceivability Entail Possibility?" in T. Gendler and J. Hawthorne (eds.), *Conceivability and Possibility.* Oxford: Oxford University Press.

Chalmers, D.J. (2003a). "The Content and Epistemology of Phenomenal Belief," in Q. Smith & A. Jokic (eds.), *Consciousness: New Philosophical Perspectives*. Oxford: Oxford University Press.

Chalmers, D.J. (2003b). "Consciousness and its Place in Nature," in Stich, Stephen P.; Warfield, Ted A. (eds.). *The Blackwell Guide to Philosophy of Mind*. Malden, MA: Blackwell Publishing Ltd.

Chalmers, D. J. (2010). *The Character of Consciousness*. Oxford University Press.

Chalmers, D.J. and McQueen, K.J. (2014), "Consciousness and the Collapse of the Wave Function," in S. Gao (ed.), *Consciousness and Quantum Mechanics*, Oxford: Oxford University Press.

Dennett, D.C. (1991). "Real Patterns," *The Journal of Philosophy,* 88, pp. 27-51.

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

Flack, J. C. (2017). "Coarse-graining as a downward causation mechanism." *Philosophical Transactions of the Royal Society A: Mathematical, Physical and Engineering Sciences*, *375*(2109), 20160338.

Frankish, K. (2007). "The Anti-Zombie Argument," *The Philosophical Quarterly,* 57: 650-666.

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

Goff, P. (2019). *Galileo's Error: Foundations for a New Science of Consciousness*. New York: Vintage Books.

Jin, F., et al. (2017). "Experimental test of Born's rule by inspecting third-order quantum interference on a single spin in solids." *Physical Review A*, *95*(1), 012107.

Kirk, R. (2005). *Zombies and Consciousness.* Oxford: Clarendon Press.

Kripke, S. (1980). *Naming and Necessity*. Cambridge: Harvard University Press.

Levine, J. (1983). "Materialism and qualia: the explanatory gap." *Pacific Philosophical Quarterly*, 64: 354-361.

Loewer, B. (1995). "An Argument for Strong Supervenience," in E. Savellos and U. Yalcin (eds.), *Supervenience: New Essays*. Cambridge: Cambridge University Press.

Maudlin, T. (2019). *Philosophy of physics: Quantum Theory*. Princeton: Princeton University Press.

Miller, M.B. and Bassler, B.L. (2001). "Quorum Sensing in Bacteria," *Annual Review of Microbiology*, 55: 165-199.

Moran, A. (2021). "Panpsychism and grounding the qualitative," this volume.

Norsen, T. (2017). *Foundations of Quantum Mechanics: An Exploration of the Physical Meaning of Quantum Theory*. Springer.

O'Connor, T. and Wong, H.Y. (2005). "The Metaphysics of Emergence", *Noûs*, 39(4): 658–678.

Papineau, D. (1995). "Arguments for Supervenience and Physical Realization," in E. Savellos and U. Yalcin (eds.), *Supervenience: New Essays*. Cambridge: Cambridge University Press.

Papineau, D. (2020). "The Problem of Consciousness," in U. Kriegel (ed.), *The Oxford Handbook of the Philosophy of Consciousness*. Oxford: Oxford University Press.

Penrose, R. (1989). *The Emperor's New Mind: Concerning Computers, Minds and The Laws of Physics*. Oxford: Oxford University Press.

Penrose, R. (2014). "On the Gravitization of Quantum Mechanics 1: Quantum State Reduction." *Foundations of Physics* 44, pp. 557–575.

Penrose, R. and Hameroff, S. (2011). "Consciousness in the Universe: Neuroscience, Quantum Space-Time Geometry and Orch OR Theory," *Journal of Cosmology*, **14**.

Putnam, H. (1975). 'The Meaning of "Meaning"', in K. Gunderson (ed.), *Language, Mind, and Knowledge: Minnesota Studies in the Philosophy of Science*, 7, Minneapolis: University of Minnesota Press, pp. 131–193.

Rovelli, C. (2021). "Relations and Panpsychism," this volume.

Russell, B. (1927). *The Analysis of Matter.* London: Kegan Paul.

Smolin, L. and Verde, C. (2021). "Physics, Views, and Qualia," this volume.

Stapp, H. (2001). "Quantum Theory and the Role of Mind in Nature." *Foundations of Physics*. **31** (10): 1465–1499

Strawson, G. (2006). "Realistic monism: Why physicalism entails panpsychism". *Journal of Consciousness Studies*. **13**, pp. 3-31.

Tegmark, M. (2000). "Importance of quantum decoherence in brain processes." *Physical Review E*. **61** (4): 4194–4206.

Tononi, G., Boly, M., Massimini, M., & Koch, C. (2016). "Integrated information theory: from consciousness to its physical substrate." *Nature Reviews Neuroscience*, *17*(7), 450-461.

Wigner, E. P. (1961). "Remarks on the mind-body question," in I.J. Good (ed.), *The Scientist Speculates*. Heineman.

Wilczek, F. (2015). *A Beautiful Question: Finding Nature's Deep Design*. New York: Penguin.