What is a theory of consciousness for?

Chris Fields

23 Rue des Lavandières
11160 Caunes Minervois, FRANCE
fieldsres@gmail.com
ORCID: 0000-0002-4812-0744

Abstract: *Galileo's Error* (Goff, 2019) leaves important questions unasked and hence unanswered. I focus on two of these: the question of what a theory of consciousness is supposed to accomplish, and the question of what the materialism – dualism – panpsychism debate is actually about.

Keywords: Basal awareness; Explananda; Human exceptionalism; Minimal physicalism; Panpsychism; Qualia; Quantum information

Introduction

Philip Goff's new book, *Galileo's Error* advances a cogent, and as the text progresses, increasingly passionate argument for panpsychism. Panpsychism, not dualism or materialism, is Goff's proposed "foundation for a new science of consciousness." I am quite happy with panpsychism and will not contest Goff's arguments supporting it. What, however, is the "new science of consciousness" for which panpsychism is meant to provide a foundation? What is its theoretical structure? Most importantly, what questions does it answer? What is it for?

Goff saddles Galileo with creating the "problem of consciousness" (p. 3; all otherwise unspecified page references are to Goff, 2019). The problem of consciousness is that "we now seek a scientific explanation … of the conscious mind" (p. 21) but neither have one nor have a good idea how to develop one. In particular, according to Goff and many others (see e.g. Dietrich and Hardcastle, 2005), it is not clear that neuroscience can explain consciousness. But what does it mean to "explain consciousness"? What exactly is an explanatory theory of consciousness supposed to provide?

I offer two arguments in this commentary. The first is that the explanatory target of most scientists working on consciousness differs in significant ways from the explanatory target of the "science of consciousness" that Goff seems to envisage. Goff, like many other philosophers, appears to want an explanatory science of qualia as such, a science that not only explains why phenomenal experiences occur, but also why particular experiences have both the particular implementations and the phenomenal characters they have, e.g. why, at least for neurotypical humans, red things look red. He

also quite explicitly wants a solution to the combination problem. It is not clear, however, whether such a science is possible, or what it would accomplish if it is. I employ an alternative radical panpsychism, the Minimal Physicalism (MP) of Fields, Glazebrook and Levin (2021), as a contrast to Goff's approach in exploring these issues. While Goff's panpsychism is fundamentally ontological and is based on materiality, MP is fundamentally functional and is based on a characterization of physical interaction as information exchange (Fields, Glazebrook and Marcianò, 2021). Because MP employs a mathematical formalism, that of quantum information theory, it is Galilean in Goff's sense. Because it represents the contents of consciousness as subject to quantitative physical constraints, it has substantial predictive power. It predicts nothing, however, about qualia as such, and as discussed below, rejects most versions of the combination problem as ill-posed.

If the explanada for a science of consciousness are still fundamentally in question, what is the current "consciousness war" between dualism, materialism, and panpsychism even about? My second argument is that it is in fact a proxy war: a minor campaign in a broader cultural conflict about human specialness. Dualism, materialism, and panpsychism as Goff presents them are each uneasy alliances that fragment along deeper fault lines when examined. The deep conflict about specialness commands fiercer loyalties and more clearly reveals the underlying motivations of the contestants. Understanding these motivations helps reveal what a theory of consciousness of the sort Goff proposes might be *for*.

What does it mean to "explain consciousness"?

Is a theory of qualia even possible?

Explaining consciousness is regularly presented as a "grand challenge" for 21st century neuroscience (Hougan and Altevogt, 2008; see also Altevogt, Hanson and Leshner, 2008; Seth, 2010). What excites neuroscientists, however, is often regarded as technical and mundane (as "easy problems" in the terminology of Chalmers, 1996) by philosophers (see Signorelli, Szczotka and Prenter, 2021 for a more nuanced analysis). As LeDoux, Michel and Lau (2020) point out, neuroscience has now spent well over fifty years trying to understand the difference between what subjects can report awareness of, as introspectively accessible contents of consciousness, and what they can demonstrate awareness of independently of such introspective access. Such studies dodge the ontological question of what awareness is in the first place. Even theories that identify consciousness with some independently-defined construct do not explain *why* the proposed identification holds; integrated information theory (IIT; Oizumi,Albantakis and Tononi, 2014), for example, does not explain *why* locally-maximal integrated information Φ is sufficient for qualia (cf. Cerrulo, 2015). The question that most interests philosophers, the "hard problem" of why any experiences occur at all, is not just not answered; it is never even posed by the neuroscience of consciousness. Neuroscience takes it for granted that organisms – at least those with brains – are not philosophical zombies (e.g. Lamme, 2018). Indeed Klein and Barron (2020) argue that the "hard problem" is not a problem for neuroscience or any other science to solve, but rather a set of folk intuitions to be overcome.

Why this disagreement, often tacit, about fundamental explananda? Why do many if not most neuroscientists dismiss the hard problem as uninteresting or irrelevant? Is it really Galileo's fault that neuroscientists, and indeed scientists working on more phylogenetically basal, non-neural organisms, e.g. bacteria (Levin, 2020; Lyon, 2020) are more concerned with what their subjects are aware of and how their awareness informs their behavior than with what awareness *is* in some fundamental ontological sense? Proposals that awareness is or is made possible by quantum entanglement, or by its removal via the "collapse of the wavefunction" (Wigner, 1961; Georgiev, 2020) suggest, at any rate,

that employing a quantitative mathematical formalism does not inhibit theorizing about the ontology of consciousness. Everyday observations about qualia motivate both the fundamental postulates and the mathematical formalism of IIT. The question is what such theorizing is meant to accomplish. Surgeons and physicians treating the comatose need to know whether their patients are having experiences, introspectively reportable or otherwise. Ethicists face similar questions about animals, plants, AI systems, even Nature as a whole (Dietrich *et al.*, 2021). It is this kind of question that IIT, as well as other theories based in neural or more broadly, cellular or even abstract information-processing activity try to answer. Goff dismisses IIT as an "easy problem" theory (p. 35). What is the question, then, for a hard problem theory?

Goff's extended discussion of zombies suggests that whether some system X is capable of experience – capable of any kind of experience at all – is the question of interest. His discussion of black-and-white Mary suggests that the question is not whether X is having an experience but rather what kind of experience X is having. Or perhaps, it is the question of what distinguishes black-and-white experiences from color experiences, or of what makes color experiences *color* experiences, and not, say, aural or tactile experiences. A bit later we have: "The job of science [of consciousness], then is … to give an account of the place of feelings in a general theory of reality" (p. 109). Goff's real goal is finally made explicit on p. 130 – 138: it is to reposition consciousness as the *intrinsic nature* of matter, something foundational, beyond explanation, that just exists. Consciousness in this case is neither derived nor emergent, but rather fundamental, a primitive brute fact about the structure of the world.

As mentioned earlier, I agree with Goff on this point: consciousness is neither derived nor emergent, but fundamental. Indeed in MP, consciousness is a fundamental characteristic not of matter, but of the information exchange implemented by physical interactions; hence it characterizes any system that interacts with any other system, material or otherwise. The non-zombiehood of organisms or other systems of interest is, in this setting, not an explanadum for neuroscience or any other science, but rather a foundational assumption. Where does this leave the hard problem, the problem of explaining "*why* are certain kinds of brain activity correlated with consciousness?" (p. 35, emphasis in original). Or in the language of IIT, why is $\Phi > 0$ correlated with consciousness? Surely the plain answer is that these are no longer questions; for a panpsychist, *everything* is correlated with consciousness. Every material object in Goff's version, every physical interaction in MP. Why are we humans conscious? Because we exist.

Other than letting us stop worrying about the hard problem, however, what does the postulate that consciousness is fundamental buy us? Goff devotes considerable space to arguing that a successful, predictive science does not require intrinsic natures, so aside from satisfying a certain inchoate yearning for a settled ontology, what is the claim that consciousness is the intrinsic nature of humans, or of anything, good for? Goff criticizes Chalmers and McQueen for not not providing "*enough* of a causal role for the conscious mind" (p. 47, emphasis in original). Making consciousness the intrinsic nature of physical objects, or in MP, of physical interactions, appears to deprive it of any causal role whatsoever. Consciousness is a logical precondition for anything happening, indeed for anything existing, but it does not cause or prevent any particular happening. In Galilean language, consciousness is a characteristic, not a component, of a functional mechanism.

Consider black-and-white Mary. Assuming she is physical, she is conscious. So she is conscious of something. But what? What distinguishes her black-and-white experiences from her later color experiences? Not consciousness per se, but something about its particular content at some time. Neuroscience describes the differences in implementation between black-and-white and color experiences in exquisite detail, and explains how the two kinds of experiences can afford different

behaviors. Computational models generalize such implementation-level differences; both MP and IIT, for example, provide formal specifications of what differences between inputs are detectable by a given system, and of how detection of one input versus another influences behavior. What more do we want? An explanation of the qualia themselves, of the particular "feel" of particular color experiences, proponents of black-and-white Mary and similar thought experiments insist. Neuroscience notes the correlations between color experiences and affective or motivational feelings, and evolutionary neuroscience explains why, for example, experiences of red correlate, in humans, with experiences of danger in some contexts and experiences of excitement in others, but neither has anything to say beyond this. Neither do computational models of such correlations. Neither does the claim that consciousness is fundamental.

Is there a "why" question about the correlation between physical or biological states and qualia more fundamental than the ones answered by evolutionary and developmental neuroscience, and does this further question pose a new "hard problem" that a science of consciousness should solve? Both evolutionary and developmental neuroscience are concerned with averages; is there an additional, real-time "why" question about each individual qualium? Are individual qualia – the redness of a particular apple, the taste of a particular chocolate – well-enough defined to even pose this question? Illusionists argue that qualia are ephemeral and lack empirically-accessible identities, and are therefore illusory (e.g. Frankish, 2016), but does ephemerality imply illusion? Physics is full of phenomena that are ephemeral but real: fluctuations in the quantum vacuum, for example. Why should qualia not have this same status? In MP, they do: qualia are ephemeral but real. The qualia of a Boltzmann Brain (Bousso and Freivogel, 2007) are just as real, for example, as Mary's or mine, and are ephemeral by definition. As every observer and every event of observation are unique, in principle, in MP, MP regards every qualium as unique. The only science of qualia MP allows is a science of descriptive reports, 3rd-party measurements, and averages: the kind of correlative science that neuroscience – indeed, biology in general – already offers.

Goff and I appear, at least, to deeply differ on this point: Goff appears to want a science of qualia per se, an explanation of why particular qualia have particular implementations and arise in particular circumstances, and to believe that such a science can be developed from a panpsychism anchored in materiality. There seem to be two motivations for this, one scientific and the other purely philosophical. The scientific motivation stems from an assumption that both experiences and experiencers are in some sense compositional, an assumption Goff shares with IIT (Oizumi,Albantakis and Tononi, 2014; see especially Fig. 15). This assumption leads via well-trodden paths to the combination problem.

Does panpsychism need to solve the combination problem?

Panpsychism is regularly saddled with the combination problem; indeed Goff characterizes it as "the hard problem" for panpsychism (p. 147 ff) and suggests that absent a robust theory of combination, panpsychism is a "lost cause" (p. 146). How do simple experiences combine to produce complex experiences (or in the "emergentist" version, how do clusters or colocalizations of simple experiences enable the emergence of more complex experiences)? How do simple *experiencers* combine to produce, or enable the emergence of, complex experiencers? This problem is often regarded as intractable, but Goff is optimistic, particularly about the emergentist version.

Consider me, in surgery after a good wallop of anesthetic. I am "unconscious," but what does this now mean? That I no longer have an intrinsic nature? Or just that there has been some functional change, something neuroscience has a handle on (Kelz and Mashour, 2019). "I" may be anesthetized, but

plenty of my systems are still operating with full awareness, regulating heart rate and so forth. Turning the cortex off inhibits the kind of consciousness that philosophers and surgeons are most interested in, but another kind of consciousness, one at least as important to us as organisms, seems to remain. At least it does in MP, in which *every* bounded system has awareness. We know, however, essentially nothing about qualia for this kind of consciousness, just as we known nothing about qualia for fungi, plants, or bacteria. What is it like to be my basal ganglia? Such questions are rarely raised in either neuroscience and philosophy. What is it like to be a neuron in my basal ganglia? Cook (2008) provides a compelling account of what it's like to be a neuron; if he is right, it isn't pleasant.

Straightforward (i.e. reductionist) versions of the combination problem assume that the experiences of my basal ganglia are components of my experience, and experiences of neurons are components of my basal ganglia's experience. Solving this reductionist version would require saying how my experiences incorporate those of my basal ganglia. The exclusivity (or "maximality" in Goff's usage, pp. 167–169) postulate of IIT is designed to prevent this kind of cross-level compositionality of experiences by denying that my basal ganglia have any experiences at all, i.e. by making them zombies, whenever I, as a containing system with larger Φ, am conscious. Emergentist versions of the combination problem may assume that proper components of an experiencer have no experiences on their own (as in IIT), but that they somehow enable high-level experiences. Low-level biological processes clearly enable high-level processes, conscious or not, in a well-understood physiological sense. Is there a further kind of enabling that is specific to consciousness?

Computation provides a useful analogy here. The theory of virtual machines (Smith and Nair, 2005) renders the "emergence" of complex, system-scale computational behavior from the simpler behaviors of logic gates or operating-system components well understood. A virtual machine stack is, however, a semantic hierarchy, not a causal hierarchy. Low-level operations do not "combine" to form high-level operations in any straightforward, context-independent, or locally reproducible way. Computation poses an implementation problem, one solved by well-defined inter-level interfaces that enforce semantic constraints, but it poses no "combination problem" over and above this. Why should this not be the case for conscious systems?

The combination problem as Goff formulates it does not arise in MP; indeed the quantum-theoretic formalism renders it ill-posed (Fields, Glazebrook and Levin, 2021). Systems in MP execute hierarchical computations – indeed, hiearchical Bayesian inference – but do not have well-defined mereological decompositions. Experiences are compositional in MP under precisely-specified conditions, but experiencers are not compositional. Agents in MP – and their environments – are virtual-machine hierarchies over underlying quantum computations. They are semantic constructs, not causal constructs.

If adopting a panpsychist position does not, as MP demonstrates, entail facing the combination problem, why is the combination problem so central to discussions of panpsychism? The answer, I believe, is ontology. Goff's motivations are fundamentally ontological; he describes the first four chapters of *Galileo's Error* as "an essay in ontology" (p. 183). Here MP provides a somewhat extreme counterexample: it rejects not only the commonsense classical ontology of observer-independent material objects (cf. Hoffman, Singh and Prakash, 2015), but even the assumptions of observer-independent space, time, or information (cf. Wheeler, 1989). It replaces, as Galilean science does generally when pushed toward instrumentality, such ontological assumptions with assumptions about how systems or interactions can be formally described.

The materialist and dualist positions with which Goff contrasts his panpsychism are similarly ontological positions. The problem of qualia that they attempt to solve is an ontological problem, and the solutions they offer are ontological solutions. Goff never says so explicitly (though he comes close in Technical Appendix B), but one might speculate that the *real* error of Galilean science, in Goff's view, is its deep-seated disinterest in ontology. Aside from recognizing that consciousness is a real phenomenon (p. 174), however, what is the ontology of consciousness trying to accomplish? What is the problem of qualia actually about?

What is the "consciousness war" about?

The difficulty of pinning down exactly what a theory of consciousness is supposed to explain suggests that something not quite ordinary is going on. Sciences generally have straightforward explanatory goals; why should a science of consciousness be different? Goff remarks that "[t]he brute identity theory [identifying consciousness with brain states] is very unsatisfying" (p. 108). How then is the idea that consciousness is the *intrinsic nature* of brain states satisfying? If not an ontological yearning, what is a theory identifying consciousness with the intrinsic nature of matter meant to satisfy?

Goff answers this question in Chapter 5: a panpsychist theory of consciousness is meant to satisfy an *existential* yearning, a yearning for meaning. It is meant to counter the "cosmic alienation" (p. 216) of postmodern life. While many blame this cognitive and emotional dislocation on materialism, Goff correctly points to dualism as the culprit: it is modern dualism, particularly as expressed by Descartes, that isolated humans even from other animals as the only conscious entities on Earth. Is it fair, however, to blame Descartes' contemporary Galileo for this? Is it really the idea that science should employ the formal tools of mathematics that separated humans from their mammalian cousins, or from the idea of Gaia? Should we not look a bit deeper into history?

We can presume that at one time, say 15,000 years ago, everyone on Earth was an animist, and hence a panpsychist. What happened? The invention of human "specialness" seems to have coincided, though at different times on different continents, with the near-simultaneous development of agriculture, urbanism, and overtly-hierarchical political power (Harari, 2014). Perhaps feeling special – as Marx suggested – made life as an expendable underling a bit more tolerable. Hence the abandonment of panpsychism in favor of human exceptionalism as a foundational cultural myth appears, at least in Eurasia, to have predated the European Enlightenment by several millennia. No one criticizes the exceptionalist myth or its cultural effects better than Gray (2002). It is unfortunate that Goff did not pursue the historical demise of animist panpsychism more deeply, as it perhaps offers some advice for our current predicament.

I would suggest that the pan-cultural replacement of animism by exceptionalism also offers some insight into the current battles about consciousness. Thoroughgoing materialists tend to align with thoroughgoing panpsychists in believing that humans are just part of the natural order, more cognitively capable and hence more interesting psychologically than rocks, but of no different ontological status. Such a view is highly deflationary; it implies that the science of human experience is the science of human experiential capabilities, an implication that MP joins evolutionary neuroscience in embracing. Even dualists can support this deflationary view provided they believe, as animists do, that rocks have rock-specific souls. More sectarian panpsychists, however, limit the "pan-" to some preferred class of entities, mammals for example, or more liberally, animals or even all organisms. Rocks and artifacts are out; they have the "wrong organization" to be conscious. Perhaps as "mere aggregates" they lack intrinsic natures; at any rate, they are zombies. Sectarian materialists are happy with this limitation,

imagining that consciousness “emerges” from zombiehood with the evolution of complex nervous systems, maybe only with human nervous systems. Or maybe, as Bell (1990) teases us, only with the evolution of Ph.D.s. Contemporary dualists are, by and large, happy to go along with this: contemporary dualism is partial to zombies. Hence whether a theory admits zombies as logically possible is the acid test between exceptionalism and a truly universal, exception-free panpsychism, like MP, that views consciousness as a *logical* precondition for existence.

It is not clear from *Galileo's Error* where Goff's loyalties lie in this larger debate (see especially pp. 113 – 114). But if Goff does support a thoroughgoing, unrestricted panpsychism, why should mathematics pose a problem? Simple systems are just as conscious as complex ones, but understanding what their experiences are *like* may well benefit from – even require – a sophisticated mathematical formalism. We may need math to solve the combination problem, and even if there is no such thing as combination, we may need math – indeed approaches as different as IIT and MP demand math – to understand how complex systems use high-bandwidth informational inputs to guide their behavior. If, on the other hand, Goff supports a more limited, sectarian brand of panpsychism, e.g. one compliant with the Maximality condition of IIT (pp. 167–169), a formal analysis may be required to identify those systems that neither contain nor are contained in systems with larger Φ, and hence to distinguish conscious systems from zombies. Either way, what Galileo taught us seems essential to a science of consciousness, not antithetical to it. True, Galileo dispensed with the “what is it like” questions, but we may need his methods to answer them, however conditionally.

One is left, unfortunately, with the feeling that Galileo has been singled out as a bogeyman, as so many have singled out “materialism” or “Enlightenment culture.” Humans are humans, and humans are, fortunately or otherwise, descendants of a long and difficult mammalian, vertebrate, metazoan, and ultimately microbial lineage. It is no surprise that we look out for ourselves first and our kin or our neighbors second. It is no surprise that we think we're special. Science has been chipping away at this sense of specialness at least since Copernicus. Hopefully the developing science of consciousness can help.

References

Altevogt, B. M., Hanson, S. L. and Leshner, A. I. (2008) Molecules to minds: Grand challenges for the 21st century, *Neuron*, 60, pp. 406–408.

Bell, J. (1990) Against ‘measurement’, *Physics World*, 3(8), pp. 33–41.

Bousso, R. & Freivogel, B. (2007) A paradox in the global description of the multiverse, *Journal of High Energy Physics*, 06, 018.

Cerullo, M. A. (2015) The problem with Phi: A critique of Integrated Information Theory, *PLoS Computational Biology*, 11, e1004286.

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

Cook, N. D. (2008) The neuron-level phenomena underlying cognition and consciousness: Synaptic activity and the action potential, *Neuroscience*, 153, pp. 556–570.

Dietrich, E., Fields, C., Sullins, J. P., van Heuveln, B. & Zebrowski, R. (2021) *Great Philosophical Objections to Artificial Intelligence: The History and Legacy of the AI Wars*, London: Bloomsbury.

Dietrich, E. & Hardcastle, V. G. (2005) *Sisyphus's Boulder: Consciousness and the Limits of the Knowable*, Amsterdam: Johns Benjamin.

Fields, C., Glazebrook, J. F. & Levin, M. (2021) Minimal physicalism as a scale-free substrate for cognition and consciousness, *Neuroscience of Consiousness*, 2021, niab013.

Fields, C., Glazebrook, J. F. & Marcianò, A. (2021) Reference frame induced symmetry breaking on holographic screens, *Symmetry*, 13, 408.

Frankish, K. (2016) Illusionism as a theory of consciousness, *Journal of Consciousness Studies*, 23, pp. 11–39.

Georgiev, D. D. (2020). Quantum information theoretic approach to the mind–brain problem, *Progress in Biophysics and Molecular Biology*, 158, pp. 16–32.

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

Gray, J. N. (2002) *Straw Dogs*, London: Granta.

Harari, Y. N. (2014) *Sapiens*, New York: Harper.

Hoffman, D. D., Singh, M. & Prakash, C. (2015) The interface theory of perception, *Psychonomic Bulletin & Review*, 22, pp. 1480–1506.

Hougan, M. & Altevogt, B. M. (Rapporteurs) (2008) *From Molecules to Minds: Challenges for the 21st Century*, Washington, DC: National Academies Press.

Kelz, M. B. & Mashour, G. A. (2019) The biology of general anesthesia from paramecium to primate, *Current Biology*, 29, pp. R1199–210.

Klein, C. & Barron, A. B. (2020) How experimental neuroscientists can fix the hard problem of consciousness, *Neuroscience of Consciousness* 2020, niaa009.

Lamme, V. A. F. (2018) Challenges for theories of consciousness: Seeing or knowing, the missing ingredient and how to deal with panpsychism, *Philosophical Transactions of the Royal Society of London B*, 373, 20170344.

LeDoux, J. E., Michel, M. & Lau, H. (2020). A little history goes a long way toward understanding why we study consciousness the way we do today, *Proceedings of the National Academy of Science USA*, 117, pp. 6976–6984.

Levin, M. (2020) Life, death, and self: Fundamental questions of primitive cognition viewed through the lens of body plasticity and synthetic organisms, *Biochemical and Biophysical Research Communications*, in press. https://doi.org/10.1016/j.bbrc.2020.10.077

Lyon, P. (2020) Of what is ‘‘minimal cognition’’ the half-baked version? *Adaptive Behavior,* 28:, pp. 407–428.

Oizumi, M., Albantakis, L. & Tononi, G. (2014) From the phenomenology to the mechanisms of consciousness: Integrated Information Theory 3.0, *PLoS Computational Biology,* 10, e1003588.

Seth, A. (2010) The grand challenge of consciousness, *Frontiers in Psychology,* 1, 5.

Signorelli, C. M., Szczotka, J. & Prenter, R. (2021). Explanatory profiles of models of consciousness - Towards a systematic classification, Preprint, 2021.

Smith, J. E. & Nair, R. (2005) The architecture of virtual machines, *IEEE Computer,* 38(5), pp. 32–38.

Wheeler, J. A. (1989) Information, physics, quantum: The search for links, in Zurek, W. J. (Ed.) *Complexity, Entropy, and the Physics of Information,* Boca Raton, FL: CRC Press, pp. 3–28.

Wigner, E. P. (1961) Remarks on the mind-body question, in Good, I. J. (Ed.), *The Scientist Speculates*, London: Heinemann, pp. 284–302.