doi.org/10.53975/i2e7-1t5x

Seeds of Science

A Paradigm for AI Consciousness

Michael Johnson[1]

Abstract

How can we create a container for knowledge about AI consciousness? This work introduces a new framework based on physicalism, decoherence, and symmetry. Major arguments include (1) atoms are a more sturdy ontology for grounding consciousness than bits, (2) Wolfram's 'branchial space' is where an object's true shape lives, (3) electromagnetism is a good proxy for branchial shape, (4) brains and computers have significantly different shapes in branchial space, (5) symmetry considerations will strongly inform a future science of consciousness, and (6) computational efficiency considerations may broadly hedge against "s-risk".

I. AI consciousness is in the wind

AI is the most rapidly transformative technology ever developed. Consciousness is what gives life meaning. How should we think about the intersection? A large part of humanity's future may involve figuring this out. But there are three questions that are actually quite pressing, and we may want to push for answers on:

1. *What is the default fate of the universe if a technological singularity happens and breakthroughs in consciousness research don't?*
2. *What interesting qualia-related capacities does humanity have that synthetic superintelligences might not get by default?*
3. *What should CEOs of leading AI companies know about consciousness?*

The following is a wide-ranging safari through various ideas and what they imply about these questions. Some items are offered as arguments, others as hypotheses or observations; I've tried to link to the core sources in each section. In the interests of exposition, I've often erred on the side of being opinionated. But first — some preliminaries about why AI consciousness is difficult to study.

[1] Symmetry Institute

Key references:

- Faggella, D. (2023). *A Worthy Successor — The Purpose of AGI*

II. The social puzzle of AI consciousness

"AI consciousness" is an overloaded term, spanning at least three dimensions:

1. Human-like responsive sophistication: does this AI have a sense of self? Is it able to understand and contextualize intentions, capabilities, hidden state, and vibes, both in itself and others? The better AI is at playing the games we think of as characteristically human (those which are intuitive to us, and those we ascribe status to), the more "consciousness" it has.
2. In-group vs Big Other: is the AI part of our team? Our Team has interiority worth connecting with and caring about ("moral patienthood"). The Other does not.
3. Formal phenomenology: in a narrow and entirely technical and scientifically predictive sense, if you had the equations for qualia (the elements and composition of subjective experience) in-hand, would this AI have qualia?

It's surprisingly difficult to talk about technical details of AI consciousness for at least three reasons. First, the other non-technical considerations are more accessible and act as conversational attractors. Second, AI consciousness is at the top of an impressive pyramid of perhaps *10-20 semi-independent open problems in metaphysics*, and being "right" essentially relies on making the correct assumption for each while having no clean experimental paradigm nor historical tradition to fall back on — in some ways AI consciousness is the final boss of philosophy.

Third, humans are coalitional creatures — before we judge the truth of a statement, we instinctually evaluate its implications for our alliances. To take an opinionated position on AI consciousness is to risk offending our coalition members, ranging from colleagues, tenure committees, donors, & AI labs, each with their own forms of veto power. This pushes theorists toward big-tent, play-it-safe, defer-to-experts positions.
But in truth, *there are no experts in AI consciousness, and it's exactly in weird positions that may offend intuitive and coalitional sensibilities where the truth is*

most likely to be found. As Eric Schwitzgebel puts it, "Common sense is incoherent in matters of metaphysics. There's no way to develop an ambitious, broad-ranging, self-consistent metaphysical system without doing serious violence to common sense somewhere. It's just impossible. Since common sense is an inconsistent system, you can't respect it all. Every metaphysician will have to violate it somewhere."

Key references:

- Hoel, E.P. (2024). *Neuroscience is pre-paradigmatic. Consciousness is why*
- Schwitzgebel, E. (2024). *The Weirdness of the World*
- Johnson, M.E. (2022). *It From Bit, Revisited*

III. Will future synthetic intelligences be conscious?

The question of machine consciousness rests on 'what kind of thing' consciousness is. If consciousness is a lossy compression of complex biological processes, similar to "metabolism" or "mood", asking whether non-biological systems are conscious is a *Wittgensteinian type error — i.e. a nonsensical move, similar to asking "what time is it on the sun?" or trying to formalize élan vital*. When we run into such category errors, our task is to stop philosophically hitting ourselves; i.e. to debug and dissolve the confusion that led us to apply some category in a context where it's intrinsically ill-defined.

On the other hand, if consciousness is a "first-class citizen of reality" that's *definable everywhere*, like electric current or gravity, machine consciousness is a serious technical question that merits a serious technical approach. I believe consciousness is such a first-class citizen of reality. Moreover,

I believe synthetic intelligences *will* be conscious, albeit with a huge caveat.

AIs will be conscious (because most complex heterogenous things probably are): Just as we're made of the same stuff as rocks, trees, and stars, it's difficult to formalize a theory of consciousness where most compound physical objects don't have roughly the same *ontological* status when it comes to qualia. I.e. I take

it as reasonable that humans are less 'a lone flickering candle of consciousness' and more a particularly intelligent, cohesive, and agentic "*qualiafauna*" that has emerged from the endemic qualiaflora. We are special — but for quantitative, not qualitative, reasons. Synthetic intelligences will have qualia, because the universe is conscious by default. We don't have to worry about the light of consciousness going *out* — though we can still worry about it going *weird*.

Caveat: Only real things can be conscious.

There's a common theme of attributing consciousness to the highest-status primitive. Theology, Psychology, and Physics have each had their time in the sun as "the most real way of parametrizing reality" and thus the 'home domain' of consciousness. Now that software is eating the world, computation is king — and *consciousness joins its court*. In other words, 'what kind of thing consciousness is' is implicitly not just an assertion of *metaphysics* but of *status*.

Although software is ascendant, computational theory is still in search of an overarching framework. The story so far is that there are different classes of computation, and problems and computational systems within each class are *equivalent in fundamental ways. Quantum computers aside, all modern systems are equivalent to what we call a "Turing machine*" — essentially a simple machine that has (1) a tape with symbols on it, (2) an 'action head' that can read and write symbols, and (3) rules for what to do when it encounters each symbol. All our software, from Microsoft Excel to GPT4, is built from such *Turing-level computations*.

Although computational theory *in general* may prove to intersect with physics (e.g. *digital physics, cellular automatons*), Turing-level computations *in particular* seem formally distinct from anything happening in physics. We speak of a computer as "implementing" a computation — but if we dig at this, precisely *which* Turing-level computations are happening in a physical system is defined by *convention* and *intention*, not objective fact.

- In mathematical terms, there exists no *1-to-1 and onto mapping between the set of Turing-level computations and the set of physical microstates (broadly speaking, this is a version of the Newman Problem*).
- In colloquial terms, bits and atoms are differently shaped domains and it

doesn't look like they can be reimagined to cleanly fit together.
- In metaphysical terms, computations have to be physically implemented in order to be real. However, there are multiple ways to physically realize any (Turing-level) computation, and multiple ways to interpret a physical realization as computation, and no privileged way to choose between them. Hence it can be reasonably argued that computations are never "actually" physically implemented.

To illustrate this point, imagine drawing some boundary in spacetime, e.g. a cube of 1mm^3. Can we list which Turing-level computations are occurring in this volume? My claim is we can't, because whatever mapping we use will be arbitrary — there is no objective fact of the matter (see *Anderson & Piccinini 2024*).

And so, because these domains are not equivalent, we're forced to choose one (or neither) as the natural home of consciousness; it cannot be both. I propose we choose the one that is *more real* — and while computational theory is beautiful, it's also a "mere" tautological construct whereas physics is predictive. I.e. *electrons are real in more ways than Turing-level bits are, and so if consciousness is real, it must be made out of physics. If it's possible to describe consciousness as (hyper)computation, it'll be described in a future computational framework that is isomorphic to physics anyway*. *Only hardware can be conscious, not software.*

This may sound like "mere metaphysics" but whether physical configurations or computational processes are the seat of value is likely the fault-line in some *future holy war*.*

*I think the best way to adjudicate this is predictiveness and elegance. Maxwell and Faraday assumed that electromagnetism had deep structure and this led to novel predictions, elegant simplifications, and eventually, the iPhone. *Assuming qualia has deep structure* should lead to something analogous.

Core reference for my argument:

- Anderson, N., & Piccinini, G. (2024). *The Physical Signature of Computation: A Robust Mapping Account*

Key references supporting consciousness as computational:

- Safron, A. (2021). *IWMT and the physical and computational substrates of consciousness*
- Rouleau, N., & Levin, M. (2023). *The Multiple Realizability of Sentience in Living Systems and Beyond*
- Bach, J. (2024). *Cyber Animism*
- Butlin, P., & Long, R., et al. (2023). *Consciousness in Artificial Intelligence: Insights from the Science of Consciousness*
- Levin, M. (2022). *Technological Approach to Mind Everywhere: An Experimentally-Grounded Framework for Understanding Diverse Bodies and Minds*
- Levin, M. (2024). *The Space Of Possible Minds*

Key references supporting consciousness as physical, or not Turing-level computational:

- Piccinini, G. (2015). *Physical Computation: A Mechanistic Account*
- Johnson, M.E. (2017). *Against functionalism*
- Aaronson, S. (2014). *"Could a Quantum Computer Have Subjective Experience?"*
- Johnson, M.E. (2016). *Principia Qualia*
- Johnson, M.E. (2019). *Taking monism seriously*
- Kleiner, J. (2024). *Consciousness qua Mortal Computation*
- Kleiner, J. (2024). *The Newman Problem of Consciousness Science*
- Hales, C.G., & Ericson, M. (2022). *Electromagnetism's Bridge Across the Explanatory Gap: How a Neuroscience/Physics Collaboration Delivers Explanation Into All Theories of Consciousness*
- Johnson, M.E. (2022). *AIs aren't conscious; computers are*
- McCabe, G. (2004). *Universe creation on a computer*
- Schiller, D. (2024). *Functionalism, integrity, and digital consciousness*
- Tononi, G., & Koch, C. (2014). *Consciousness: Here, There but Not Everywhere*
- Pachniewski, P. (2022). *Not artificially conscious*

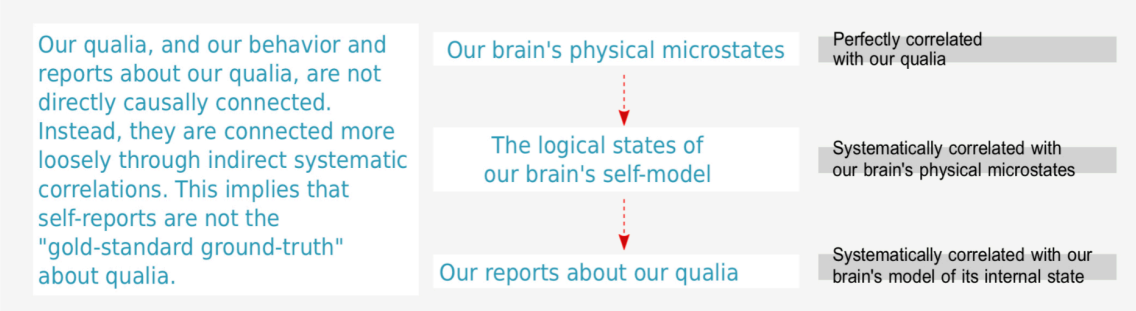
See also:

- Lee, A.Y. (2024). *Objective Phenomenology*
- Kleiner, J. (2024). *Towards a structural turn in consciousness science*
- Johnson, M.E. (2022). *It From Bit, Revisited*
- Ladyman, J. (2023). *Structural Realism*
- Kanai, R., & Fujisawa, I. (2023). *Towards a Universal Theory of Consciousness*
- Forthcoming from *Dalrymple and from Gorard*

IV. We should not rely on AIs or brain emulations to accurately self-report qualia

Many of the most effortlessly intuitive human capacities have proven the most difficult to replicate in artificial systems. Accurately reporting phenomenology may be a particularly thorny problem.

I *suggested* in Principia Qualia that our capacity to accurately report our phenomenology rests on a laboriously evolved system of correlations that's very particular to our substrate:



Graphic: Qualia reports & their coupling with reality (orig. *Johnson 2016*, Appendix C)
I.e. we can talk "about" our qualia because qualia-language is an efficient compression of our internal logical state, which evolution has beaten into systematic correlation with our actual qualia. This is a contingent correlation, not an intrinsic feature of reality.

If we transfer an organism's computational signature to a new substrate, the new substrate it's running on will have some qualia (because ~everything physical has qualia), but porting a computational signature, no matter how well it replicates behavior, will not necessarily replicate the qualia traditionally associated with the signature or behavior. By shifting the physical basis of the system, the link between "physical microstate" and "logical state of the brain's self-model" breaks and would need to be re-evolved.

Over the long term, most classes of adaptive systems are in fact likely to (re)develop such language games that are coupled to their substrate qualia, for the same reasons our words became systematically coupled to our brain qualia — but the shape of their concepts and dimensions of normative loading may be very different. Language's structure comes from its usefulness, and if we were to design a reporting language for "functionally important things about nervous systems" vs a reporting language for "functionally important things about computer state," we'd track very different classes of system & substrate dynamics.

Don't trust what brain uploads or synthetic intelligences say about their qualia — though by all means *be kind to them*.[1]

Key references:

- Johnson, M.E. (2016). *Principia Qualia*
- Kleiner, J., & Hoel, E.P. (2021). *Falsification and consciousness*
- Hoel, E.P. (2024). *AI Keeps Getting Better at Talking About Consciousness*
- Johnson, M.E. (2019). *Taking monism seriously*

Exploring the nature of systematic correlations between reality, brain, and language:

- Safron, A. (2021). *IWMT and the physical and computational substrates of consciousness*

- Ramstead, M., et al. (2023). *On Bayesian mechanics: a physics of and by beliefs*
- Long, R. (2023). *What to think when a language model tells you it's sentient*
- Quine, W.V.O. (1960). *Word and Object*

V. Technological artifacts will have significantly different qualia signatures & boundaries than evolved systems

In "*What's out there?*" I suggested that:

> A key lens I would offer is that the functional boundary of our brain and the phenomenological boundary of our mind overlap fairly tightly, and this may not be the case with artificial technological artifacts. And so artifacts created for functional purposes seem likely to result in unstable phenomenological boundaries, unpredictable qualia dynamics and likely no intentional content or phenomenology of agency, but also 'flashes' or 'peaks' of high order, unlike primordial qualia. We might think of these as producing 'qualia gravel' of very uneven size (mostly small, sometimes large, odd contents very unlike human qualia).

Our intuitions have evolved to infer the internal state of other creatures on our tree of life; they're likely to return nonsense values when applied to technological artifacts, especially those utilizing crystallized intelligence.

There's *lively discussion around whether Anthropic's "Claude" chatbot is conscious (and Claude does nothing to deflate* this). But if consciousness requires something to be physically instantiated, every 'chunk' of consciousness must have extension in space and time. Where is Claude's consciousness? Is it associated with a portion of the GPU doing inference in some distant datacenter, or a portion of the CPU and I/O bus on your computer, or in the past humans that generated Claude's training data, or the datacenter which originally trained the model? Is there a singular "Claude consciousness" or are there thousands of small shards of experience in a computer? What we speak of as "Claude" may not have a clean referent in the domain of consciousness, and in general we

should expect *most* technological artifacts to have non-intuitive projections into consciousness.

This observation, although important, is also somewhat shallow — *of course* computers will exhibit different consciousness patterns than brains. To go deeper, we need to look at the details of our substrate.

Key references:

- Johnson, M.E. (2019). *What's out there?*
- Wollberg, E. (2024). *Qualia Takeoff in The Age of Spiritual Machines*
- Johnson, M.E. (2022). *Qualia Astronomy & Proof of Qualia*

VI. Branchial space is where true shape lives

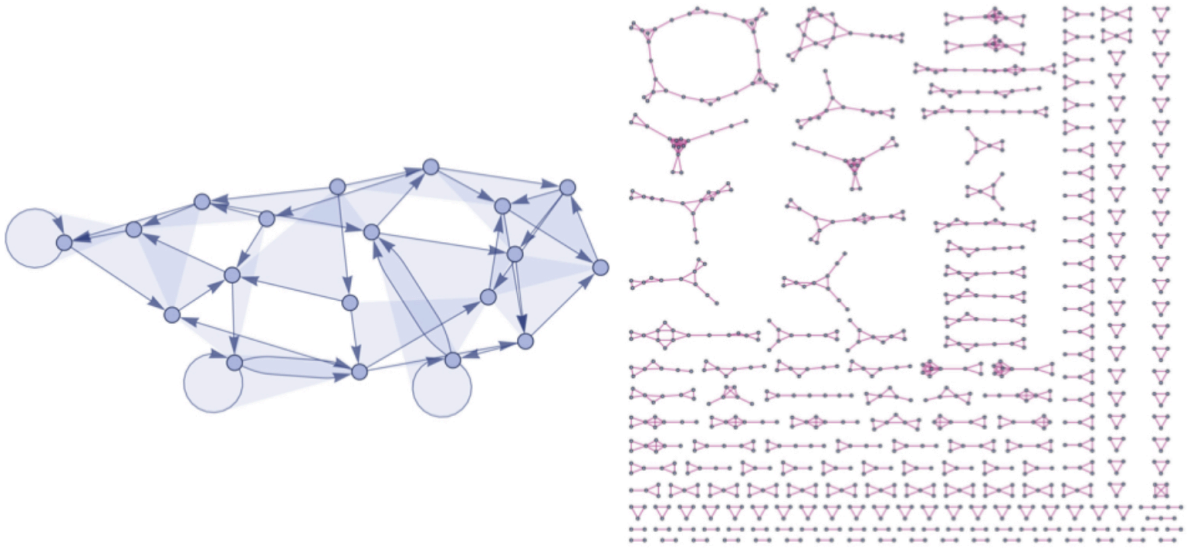
A strange but absolutely central concept in modern physics is that quantum particles naturally exist in an ambivalent state — a "multiple positions true at the same time" superposition. *Decoherence* is when interaction with the environment forces a particle to commit to a specific position, and this (wave-like) superposition collapses into one of its (particle-like) component values. The *Copenhagen interpretation suggested decoherence is random, but over the past ~2 decades Hugh Everett's many-worlds interpretation* (MWI) has been gaining favor. MWI frames decoherence as a sort of "branching": instead of the universe randomly choosing which value to collapse into, all values still exist but in *different branches* of reality.

E.g. let's say we're observing a cesium-137 atom. This atom is unstable and can spontaneously *decay* (a form of decoherence) into either barium-137 or barium-137m. It decays into barium-137m. The many-worlds interpretation (MWI) claims that it did both — i.e. there's a branch of reality where it decayed into barium-137, and another branch where it decayed into barium-137m, and we as observers just happen to be in the latter branch. MWI may sound like a very odd theory, but it collects and simplifies an enormous amount of observations and confusions about what happens at the quantum level.

Stephen Wolfram suggests understanding the MWI in terms of *branchial graphs where each interaction which can cause decoherence creates a new branch.*

This sort of graph gets Vast very quickly, but in principle each branch is perfectly describable. Wolfram's new physics proposes *the universe can be thought of as the aggregate of all such graphs*, which he calls "branchial space":

Tracing through the connections of a branchial graph gives rise to the notion of a kind of space in which states on different branches of history are laid out. In particular, branchial space is defined by the pattern of entanglements between different branches of history in possible branchial graphs.



Graphic: a branchial graph (orig. Namuduri, M. (2020). *Comparing expansion in physical and branchial space*)

Different branches of reality split off and can diverge — but they can also interact, and merge (recohere). Many "weird quantum effects" such as the *double-slit experiment can be formally reframed as arising from interactions between branches (this is the core thesis behind quantum computing), and time itself may be understood as emergent from branchial structure*.

The "branchial view" suggests that *different types of objects are different types of knots in branchial space*, as defined by (1) how their particles are connected, (2) what patterns of coherence and decoherence this allows, and (3) the branches that form & interact due to this decoherence. A wooden chair, for instance, is a

relatively static 'knot' (though there's always froth at the quantum level); a squirrel is a finely complex process of knotting; the sun is a 4.5 billion-years-long nuclear Gordian weave.

The "branchial view" matters for consciousness because decoherence may be necessary for, *if not identical to*, consciousness.

Decoherence is often seen as an impediment to consciousness; e.g. *Max Tegmark argues that predictive systems must minimize it as a source of uncertainty. On the other hand, Scott Aaronson* argues decoherence is instead a *necessary* condition for consciousness:

> [Y]es, consciousness is a property of any suitably-organized chunk of matter. But, in addition to performing complex computations, or passing the Turing Test, or other information-theoretic conditions that I don't know (and don't claim to know), there's at least one crucial further thing that a chunk of matter has to do before we should consider it conscious. Namely, it has to participate fully in the Arrow of Time. More specifically, it has to produce irreversible decoherence as an intrinsic part of its operation. It has to be continually taking microscopic fluctuations, and irreversibly amplifying them into stable, copyable, macroscopic classical records.
>
> ...
>
> So, why might one conjecture that decoherence, and participation in the arrow of time, were necessary conditions for consciousness? I suppose I could offer some argument about our subjective experience of the passage of time being a crucial component of our consciousness, and the passage of time being bound up with the Second Law. Truthfully, though, I don't have any a-priori argument that I find convincing. All I can do is show you how many apparent paradoxes get resolved if you make this one speculative leap.
>
> …
>
> There's this old chestnut, what if each person on earth simulated one neuron of your brain, by passing pieces of paper around. It took them several years just to simulate a single second of your thought processes. Would that bring your subjectivity into being? Would you accept it as a replacement for your current body? If so, then what if your brain were simulated, not neuron-by-neuron, but by a gigantic lookup table? That is,

> what if there were a huge database, much larger than the observable universe (but let's not worry about that), that hardwired what your brain's response was to every sequence of stimuli that your sense-organs could possibly receive. Would that bring about your consciousness? Let's keep pushing: if it would, would it make a difference if anyone actually consulted the lookup table? Why can't it bring about your consciousness just by sitting there doing nothing?

Aaronson goes on to list some paradoxes and puzzling edge-cases that resolve if 'full participation in the Arrow of Time' is a necessary condition for a system being consciousness: e.g., whether brains which have undergone Fully Homomorphic Encryption (FHE) could still be conscious (no – Aaronson suggests that nothing with a clean digital abstraction layer could be) or whether a fully-reversible quantum computer could exhibit consciousness (no – Aaronson argues that no fully-reversible process could be). (Paragraph from *Johnson 2016*)

I agree with Aaronson and propose going further ("MBP Hypothesis" in Appendix E, *Johnson 2016; see also significant work in Chalmers & McQueen 2021). Briefly, my updated 2024 position is "an experience is an object in branchial space, and the magnitude of its consciousness is the size of its branchial graph." Pick a formal specification of branchial space, add a boundary condition for delineating subgraphs (e.g. ontological, topological, amalgamative-majority, dispersive/statistical*, compositional), and we have a proper theory.

Slightly rephrased: the thesis of "Qualia Formalism" (QF) or "Information Geometry of Mind" (IGM) is that *a proper formalism for consciousness exists*:

> An information geometry of mind (IGM) is a mathematical representation of an experience whose internal relationships between components mirror the internal relationships between the elements of the subjective experience it represents. A correct information geometry of mind is an exact representation of an experience. More formally, an IGM is a mathematical object such that there exists an isomorphism (a one-to-one and onto mapping) between this mathematical object and the experience it represents. Any question about the contents, texture, or affective state of an experience can be answered in terms of this geometry. (*Johnson 2023*)

If such a formalism exists, a core question is how to derive it. I'm speculating here that perhaps (a) what Wolfram calls "branchial space" is the native domain of this formalism, (b) an IGM/QF will be isomorphic to a bounded branchial graph, and (c) solving the binding/boundary problem is identical with determining the signature for where one bounded graph ends and another begins.

However, to recap — this section's thesis is that branchial **space is where true shape lives**. This has three elements:

1. **Decoherence is a crucial part of reality and can be understood in terms of branchial space;**
2. **Decoherence may be necessary for consciousness;**
3. **If we wish to understand the "true shape" of something in a way that may reflect its qualia, we should try to infer its shape in branchial space.**

Key references:

- Aaronson, S. (2014). *"Could a Quantum Computer Have Subjective Experience?"*
- Sandberg, A., et al. (2017). *That is not dead which can eternal lie: the aestivation hypothesis for resolving Fermi's paradox*
- Tegmark, M. (1999). *The importance of quantum decoherence in brain processes*
- Tegmark, M. (2014). *Consciousness as a State of Matter*
- Johnson, M.E. (2016). *Principia Qualia*
- Johnson, M.E. (2023). *Qualia Formalism and a Symmetry Theory of Valence*
- Wikipedia, accessed 29 April 2024. *Quantum decoherence*
- Wolfram, S., et al. (2020). *The Wolfram Physics Project; example* of branchial expansion
- Barrett, A. (2014). *An integration of integrated information theory with fundamental physics*
- Albantakis, L., et al. (2023). *Integrated information theory (IIT) 4.0: Formulating the properties of phenomenal existence in physical terms*

- Chalmers, D.J., & McQueen, K.J. (2021). *Consciousness and the Collapse of the Wave Function*

VII. Brains and computers have vastly different shapes in branchial space

A defining feature of brains is *self-organized criticality* (SOC). “Critical” systems are organized in such a way that a small push can change their attractor; “self-organized” means assembled by intrinsic and stochastic factors, not top-down design. A property of SOC systems is they *stay* SOC systems over time — the system evolves criticality itself as one of its attractors. In other words, the brain is highly sensitive to even small inputs, and will eventually regenerate this sensitivity almost regardless of what the input is.

Brains at the edge of criticality can be thought of as ‘perching’ on their symmetries/ambivalences/sensory *superpositions: multiple interpretations for inputs can be considered, and as they get ruled out the system can follow the energy gradient downwards (“symmetry breaking*”). Later as the situation is metabolized the system recovers its perch, ready for the next input.[2] Given that these ‘perch positions’ are local optimums of systemic potential energy and *sensory* superpositions, and given evolution’s tendency towards scale-free motifs, I suspect these perches might also be local statistical optimums for quantum superpositions (and thus ambient decoherence).

Relatedly, self-organized criticality makes brains very good at *amplifying decoherence*. Tiny decoherence events can make neurons near their thresholds activate, which can snowball and influence the path of the whole system. Such decoherence events are *bidirectionally coupled to* the brain’s information processing: local quantum noise influences neural activity, and neural activity influences local quantum noise.
My conclusion is that the brain is an *extremely dynamic* branchial knotting process, with **each moment of experience as a huge, scale-free knot in branchial space**.[3]

Meanwhile, *modern computers minimize the amplification of decoherence*. A close signature of decoherence is heat, which computers make a lot of — but

work very hard to *buffer the system against* in order to maintain what Aaronson calls a "clean digital abstraction layer". *Every parameter of a circuit is tuned to prevent heat and quantum fluctuations from touching and changing its intended computation* (*Kleiner & Ludwig 2023*).

This makes computers rather odd objects in branchial space. They have noise/decoherence in proportion to their temperature (this is called "*Johnson noise*" after John B. Johnson; no relation) and occasionally they'll sample it for random numbers, but mostly computers are built to be deterministic — the macroscopic behavior of circuits implementing a typical computation end up exactly the same in almost all branches. This makes the computation *very differently represented* in branchial space compared to the Johnson noise of the circuits implementing it — and compared to how a brain's computations are represented in branchial space.

Clarifying what this means for computer consciousness essentially involves three questions:

1. What are the major classes of objects in branchial space?[4][5][6]
2. What are the branchial-relevant differences between brains and computers?
3. How do we construct a "qualia-weighted" branchial space such that the size of the subgraph corresponds with the amount of consciousness?[7]

These questions are challenging to address properly given current theory. As an initial thesis, I'll suggest that a reasonable shortcut to profiling an object's branchial shape is to evaluate it for criticality and electromagnetic flows. I discussed the former above; the next section discusses the latter.

Key references:

- Kleiner, J., & Ludwig, T. (2023). *If consciousness is dynamically relevant, artificial intelligence isn't conscious*
- Wikipedia, accessed 26 April 2024. *Johnson-Nyquist noise*

- Hoel, E.P., et al. (2013). *Quantifying causal emergence shows that macro can beat micro; primer*
- Johnson, M.E. (2019). *Neural Annealing: Toward a Neural Theory of Everything*
- Johnson, M.E. (2024). *Minds as Hyperspheres*
- Zurek, W.H. (2009). *Quantum Darwinism*
- Tegmark, M. (2014). *Consciousness as a State of Matter*
- Olah, C. (2024). *Distributed Representations: Composition & Superposition*

VIII. A cautious focus on electromagnetism

A growing trend in contemporary consciousness research is to focus on the electromagnetic field. Adam Barrett describes the basic rationale for 'EM field primacy in consciousness research' in *An integration of integrated information theory with fundamental physics*:

> **1.** Quantum fields are fundamental entities in physics, and all particles can be understood as
> ripples in their specific type of field.
> **2.** Since they're so fundamental, it seems plausible that these fields could be carriers for consciousness.
> **3.** The gravity, strong, and weak nuclear fields probably can't support the complexity required for human consciousness: gravity's field is too simple to support structure since it only attracts, and disturbances in the other two don't propagate much further than the width of an atom's nucleus.
> **4.** However, we know the brain's neurons generate extensive, complex, and rapidly changing patterns in the electromagnetic field.
> **5.** Thus, we should look to the electromagnetic field as a possible 'carrier' to consciousness (Summary of *Barrett 2014, quoted from Johnson 2016*)

W. H. Zurek makes a complimentary point in his famous essay *Quantum Darwinism*:

> Suitability of the environment as a channel [for information propagation] depends on whether it provides a direct and easy access to the records of the system. This depends on the structure and evolution of [the environment] *E*. Photons are ideal in this respect: They interact with various systems, but, in effect, do not interact with each other. This is why light delivers most of our information. Moreover, photons emitted by the

usual sources (e.g., sun) are far from equilibrium with our surroundings. Thus, even when decoherence is dominated by other environments (e.g., air) photons are much better in passing on information they acquire while "monitoring the system of interest": Air molecules scatter from one another, so that whatever record they may have gathered becomes effectively undecipherable.

These are substantial arguments and strongly suggest that electromagnetism plays a dominant role in binding human-scale experience. However, I would suggest three significant caveats:

1. A dominant (statistical) role is not necessarily an exclusive (ontological) role;
2. A force being necessary for binding does not imply that it's sufficient for describing or instantiating all experiential elements / records;
3. Human-scale experiences happen on a certain energy scale, and dynamics that hold at this energy scale may not hold at other scales. I.e., it seems plausible that other forces play a more significant role in binding at quantum scales, or in cosmological megastructures (e.g. black holes).

I take the branchial view of consciousness as more philosophically/ontologically precise than the electromagnetic view. However, the EM view is generally a useful compression of the branchial view, since most of the branchial dynamics associated with variance in human-scale experience are mediated by the electromagnetic field.

This shortcut is likely similarly relevant for computer consciousness. So — what do we know about brain EMF vs computer EMF?

Brain EMF profile: The brain uses voltage potentials extensively across organs, various classes of chemical gradients, cellular membranes (*~−50 mV, inside negative*) and axons (~-70mV), as well as within cells to assemble structures (DNA, proteins, etc). Brains also use chemical reactions ("metabolism") extensively during normal operation and these reactions are primarily mediated through valence shells, an electromagnetic phenomenon.

As a speculative sweeping characterization, I suspect the overall configuration resembles a highly layered configuration of nested & partially-overlapping electromagnetic shells (cf. unreleased summer 2021 talk on the binding/boundary problem). The *overall electrical polarity* of the shells may be a simple proxy for both decoherence and consciousness — e.g. the voltage potential between brainstem and brain surface should increase during psychedelics, meditation, and arousal, and decrease with age (Michael Johnson & Max Hodak in conversation, 2024).

Computer EMF profile: Computer substrates are mostly non-reactive semiconductors which prioritize conditional ease of electron flow and adjustment of phase; voltage flows through gates based on simple logic, and the configuration of these gates changes each clock cycle (typically in the GHz range). Voltage in computer logic gates approximates square waves (whereas bioelectric voltage is sinusoidal) and the sharper these transition are, the greater 'splash' in the EM field. The magnitude of this 'splash' may or may not track branchial expansion / strength of consciousness. Computers can be turned off (thus adding another category of object); brains cannot be.

Computer chips are designed to maintain neutral overall polarity, but the voltage channels (which can be thought of as *directed graphs* for current, essentially functioning as 'waveguides' for EM waves) that instantiate a computational state are relatively high voltage (~600mV+) compared to biological voltages (although getting smaller each chip generation).[8] I'll propose characterizing the electromagnetic profile of a modern processor as a strobing electromagnetic lattice.

Ultimately, these sorts of characterizations are data-starved, and one of the big 'intuition unlocks' for consciousness research in general — and the branchial view in particular — will be the capacity to visualize the EM field in realtime & high resolution. *What stories might we tell if we had better tools?*

Key references:

- Zurek, W.H. (2009). *Quantum Darwinism*
- Barrett, A. (2014). *An integration of integrated information theory with fundamental physics*

- Gomez-Emilsson, A., & Percy, C. (2023). *Don't forget the boundary problem! How EM field topology can address the overlooked cousin to the binding problem for consciousness*
- Hales, C.G., & Ericson, M. (2022). *Electromagnetism's Bridge Across the Explanatory Gap: How a Neuroscience/Physics Collaboration Delivers Explanation Into All Theories of Consciousness*

(See also *An introduction to Susan Pockett: An electromagnetic theory of consciousness*)

IX. The Symmetry Theory of Valence is a Rosetta Stone

In 2016 I offered the Symmetry Theory of Valence: *the symmetry of an information geometry of mind corresponds with how pleasant it is to be that experience*. I.e. the valence of an experience is due entirely to its structure, and symmetry in this structure intrinsically feels good (*Johnson 2016*, *Johnson 2021*, *Johnson 2023*).

Rephrased: if the proper goal of consciousness research is to construct a *mathematical formalism* for an experience (essentially, a high-dimensional shape that exactly mirrors the structure of an experience), STV predicts that the *symmetry* of this shape corresponds with the *pleasantness* of the experience. "Symmetry" is a technical term, but Frank Wilczek suggests "change without change" as a useful shorthand: for each symmetry something has, there exists a mathematical operation (e.g. a flip or rotation) that leaves it unchanged.

The fundamental question in phenomenology research is where to start — what are the natural kinds of qualia? STV has three answers to this:

1. Valence is a natural kind within experience;
2. Symmetry is a natural kind within any formalism which can represent experiential structure;
3. *These are the same natural kind*, just in different domains.

Explicitly, the Symmetry Theory of Valence is a theory of valence and is testable as such; I offer some routes in my *2023 summary paper*. Tacitly, STV is *also* a collection of implications about "what kind of thing" consciousness research is.

Just like the first line in the Rosetta Stone offered a wide range of structural constraints on Ancient Egyptian, if we can say 'one true thing' about qualia this may inform a great deal about potential approaches.

Perhaps the most significant tacit implication is importing physics' *symmetry aesthetic* into consciousness research. Nobel Laureate P. W. Anderson famously remarked "It is only slightly overstating the case to say that physics is the study of symmetry"; Nobel Laureate Frank Wilczek likewise describes symmetry as a core search criterion:

> [T]he idea that there is symmetry at the root of Nature has come to dominate our understanding of physical reality. We are led to a small number of special structures from purely mathematical considerations—considerations of symmetry—and put them forward to Nature, as candidate elements for her design. [...] In modern physics we have taken this lesson to heart. We have learned to work from symmetry toward truth. Instead of using experiments to infer equations, and then finding (to our delight and astonishment) that the equations have a lot of symmetry, we propose equations with enormous symmetry and then check to see whether Nature uses them. It has been an amazingly successful strategy. (*Wilczek 2016*)

If STV is true, Wilczek's observation about the centrality of symmetry likely also applies to consciousness. Initial results seem promising; a former colleague (A.G.E.) once suggested that "Nothing in psychedelics makes sense except in light of the Symmetry Theory of Valence" — and I tend to agree.

Concretely, importing physics' symmetry aesthetic predicts there will be phenomenological conservation laws (similar to conservation of energy, charge, momentum, etc), and suggests a phenomenological analogue to *Noether's theorem. The larger point here is that consciousness research need not start from scratch, and just as two points define a line, and three lines define a plane, it may not take too many dualities* such as STV to *uniquely identify* the mapping between the domains of consciousness & physics. Optimistically, "solving consciousness" could take years, not centuries.

STV and the "branchial view" described in Sections VI-VIII are separate hypotheses, but looking at them side-by-side elicits certain questions. How should we think about "branchial symmetry" and evaluating valence of knots-as-moments-of-experience in branchial space? Should we look for metrics of *graph uniformity, recoherence/unity*, non-interference? Physics has perhaps half a dozen formally equivalent interpretations of reality; I expect STV will too.[9]

Key references:

- Johnson, M.E. (2016). *Principia Qualia*
- Johnson, M.E. (2023). *Qualia Formalism and a Symmetry Theory of Valence*
- Johnson, M.E. (2021). *A Primer on the Symmetry Theory of Valence*
- Wolfram, S. (2021). *The Concept of the Ruliad*
- Johnson, M.E. (2019). *Taking monism seriously*
- Safron, A., et al. (2023). *Making and breaking symmetries in mind and life*
- Wilczek, F. (2016). *A Beautiful Question: Finding Nature's Deep Design*
- Gross, D.J. (1996). *The role of symmetry in fundamental physics*
- Brading, K., & Castelanni, E. (Ed). (2003). *Symmetries in physics: philosophical reflections*

(Thanks also to David Pearce for his steadfast belief in valence (or "hedonic tone") as a natural kind.)

X. Will it be pleasant to be a future superintelligence?

The human experience ranges from intense ecstasy to horrible suffering, with a wide middle. However, we also have what I would call the "Buddhist endowment" — that if and when we *remove expectation/prediction/tension from our nervous system, it generally self-organizes into a high-harmony state as the default*.

If we continue building smarter computers out of *Von Neumann architecture GPUs, or eventually switch to e.g. quantum, asynchronous, or thermodynamic processors, what valence of qualia is this likely to produce? Will these systems have any similar endowments? Are there considerations around what design choices we should avoid*?

I'll offer five hypothesis about AI/computer valence:

Hypothesis #1: Architecture drives valence in top-down systems, data drives valence in bottom-up systems. If the valence of an experience derives from its structure, we should evaluate where the structure of systems comes from. The structure of self-organized systems (like brains) varies primarily due to the data flowing through them — evolutionarily, historically, and presently. The structure of top-down systems (like modern computers) varies primarily due to fixed architectural commitments.

Hypothesis #2: Computer valence is dominated by chip layout, waveguide interference, waveform shape, and Johnson noise. I argue above that decoherence is necessary and perhaps sufficient for consciousness. Computers create lots of decoherence along their electrified circuits; however, this decoherence is treated as "waste heat" (essentially the *Johnson noise* of the circuit) and is largely isolated from influencing the computation. I suspect a future science of machine qualia will formalize how a circuit's voltage pattern, physical microstructure, computational architecture, nominal computation, and Johnson noise interact to project into branchial space.

Whether there is a *systematic connection* between high-level computational constructs (e.g. virtual characters in a video game) and qualia is extremely muddy at this point and likely highly dependent on hardware and software implementation; potentially true if the system is designed to make it true, but likely not the case by default. I.e., neither brain emulations nor virtual characters in video games will be "conscious" in any real sense by default, *although we could design a hardware+software environment where they would be*.

Hypothesis #3: Valence shells influence phenomenological valence: variance in the chemical structures that comprise a system's substrate contributes to stochastic variance in its phenomenal binding motifs. These factors will influence phenomenological structure, and thus valence.[10]

Hypothesis #4: Machine consciousness has a quadrimodal possibility distribution: instead of biology's continuous and dynamic range of valence, I expect the substrates of synthetic intelligences to reliably lead to experiences which have either (a) extreme negative valence, (b) extreme positive valence, (c) extremely neutral valence, or (d) swings between extremely positive and

negative valence. Instead of a responsive & continuous dynamism as in biology, whatever physical substrate & computational architecture the computer chip's designers originally chose will likely lock in one of these four valence scenarios regardless of what is being computed (see hypothesis #1).

Hypothesis #5: Trend toward positive valence in optimal systems: reducing energy loss is a primary optimization target for microprocessor design, and energy loss is minimized when all forms of dissonance are minimized. In theory this should lead to a trend away from the production of negative valence as computers get more energy efficient. To paraphrase Carl Shulman, *pain and pleasure may not be equally energy-efficient* — and this bodes well for future computers.

Key references:

- Johnson, M.E. (2016). *Principia Qualia*
- Johnson, M.E. (2023). *Qualia Formalism and a Symmetry Theory of Valence*
- Extropic (2024). *Ushering in the Thermodynamic Future; thread on thermodynamic processors*

On energy loss & architectural imperatives:

- Friston, K. (2010). *The free-energy principle: a unified brain theory?*
- Ramstead, M., et al. (2023). *On Bayesian mechanics: a physics of and by beliefs*
- Safron, A. (2020). *An Integrated World Modeling Theory (IWMT) of Consciousness*

XI. Will artificial superintelligences be interested in qualia?

We care about our qualia, and sometimes the qualia of those around us. Why? I'd offer three reasons:

1. As noted above, humans developed increasingly sophisticated and coherent terminology about "consciousness" because it was a great compression schema for evaluating, communicating, and coordinating internal state;
2. Caring about the qualia of ourselves and others has substantial instrumental value — happy creatures produce better results across many tasks. This led us to consider qualia as a *domain of optimization*;
3. We are porous *harmonic computers that catch feels* from those around us, incentivizing us to care about nearby harmonic computers, even if it's not in our narrow-sense self-interest.

These factors, iterated across hundreds of thousands of years and countless social, political, and intellectual contexts, produced a "*language game*" which approximates a universal theory of interiority.

Humanity is now in the process of determining whether this "as-if" universal theory of internal state can be properly systematized into an *actual* universal theory of internal state — whether qualia is the sort of thing that can have its own alchemy-to-chemistry transition, or its Maxwell's Laws moment, where a loose and haphazard language-game can find deep fit with the structure of reality and "snap into place". I'm optimistic the answer is yes, and not only because it's a Pascalian wager.

Humans care about qualia. Will Artificial Superintelligences (ASIs)?

I'd suggest framing this in terms of *sensitivity*. Humans are *sensitive* to qualia — we have a map of what's happening in the qualia domain, and we treat it as a domain of optimization. We are *Qualia Sensitive Processes* (QSPs). Most of the universe is *not* sensitive to qualia — it is made up of *Qualia Insensitive Processes* (QIPs), which do not treat consciousness as either an explicit or implicit domain of optimization.

This distinction suggests reframing our question: *is the modal synthetic superintelligence a QSP*? Similarly — is QSP status a *convergent capacity that all sufficiently advanced civilizations develop (like calculus), or is it a rare find, and something that could be lost during a discontinuous break in our lineage? What parts of the qualia domain do QSPs tend to optimize for — is it usually*

valence or are there other common axes to be sensitive to? Can we determine a typology of cosmological (physical) megastructures which optimize for each common qualia optimization target?

As a starting assumption, I suggest we can view the "qualia *language game" as a particularly useful compression of reality. If this compression is useful or incentivized* for future superintelligences, it will get used, and the normative loading we've baked into it will persist. If not, it won't. There are no laws of the universe forcing ASIs to intrinsically care about consciousness and valence, but they won't intrinsically disregard it either.

Qualia has a variable causal density, which informs the usefulness of modeling it

In *Taking Monism Seriously*, I suggested:

> There may be many possible chemical foundations for life (carbon, silicon, etc), but there will tend to be path-dependent lock-in, as biological systems essentially terraform their environments to better support biology. Terran biology can be thought of as a coordination regime that muscled out real or hypothetical competition to become the dominant paradigm on Earth. Perhaps we may find analogues here in past, present, and future phenomenology.

The more of a certain kind of structure present in the environment, the easier it is to model, remix, use as infrastructure, and in general invest in similar structure — e.g. the more DNA-based organisms in an ecosystem, the easier it is for DNA-based organisms to thrive there. The animal kingdom seems to be providing a similar service for consciousness, essentially "qualiaforming" reality such that bound (macroscopically aggregated) phenomenological experience is increasingly useful as both capacity and model. Rephrased: minds are points of high causal density in qualiaspace; the more minds present in an ecosystem, the more valuable it is to understand the laws of qualia.

I think *causal density* is a particularly useful lens by which to analyze both systems and ontologies. Erik P. Hoel has written about *causal emergence and how certain levels of abstraction are more "causally dense" or efficacious to model than others; I suspect we can take Hoel's hypothesis and evaluate causal*

density between dual-aspect monism's different projections of reality. I.e. the equations of qualia may not be particularly useful for modeling stellar fusion, but they seem relatively more useful for predicting biological behavior since the causal locus of many decisions is concentrated within clean phenomenological boundaries.

The equations of qualia don't seem particularly useful for understanding the functional properties of modern computers. Whether studying phenomenology will stay useful for humans, and become useful for modeling AI behavior, is really up to us and our relationships with AI, neurotechnology, and our *egregores. "Who colonizes whom?" may revolve around "who is legible to whom?"*

Finally — to reiterate a point, all physical processes have projections into the qualia domain. *Whatever* ASIs are doing will still have this projection! I.e. the risk is not that consciousness gets wiped out, it's that whatever optimization target ASI settles on has a "bad" projection into the qualia domain, while at the same time shifting the local environment away from the capacity or interest to self-correct. But there are reasons to believe suffering is energetically inefficient and will get optimized away. So even if we don't make ASIs explicitly care about consciousness, the process that created them may still implicitly turn out to be a QSP.

Key references:

- Wittgenstein, L. (1953). *Philosophical Investigations*
- Quine, W.V.O. (1960). *Word and Object*
- Hoel, E.P. (2017). *When the Map Is Better Than the Territory (see also Erik's primer*)
- Johnson, M.E. (2019). *What's out there?*

XII. Where is consciousness going?

I can say AI consciousness is a wild topic not just because it crosses the two most important topics of the day but also because there's a drought of formal models and intuitions diverge profoundly on how to even start. Here's a recap of what I believe are the most important landmarks for navigation:

1. AI consciousness is as much a social puzzle as a technical one;
2. We should distinguish “software consciousness” from “hardware consciousness”; only the latter can be a well-formed concept;
3. We should carefully trace through where humans’ ability to accurately report qualia comes from, and shouldn’t assume artificial systems will get this ‘for free’;
4. Artificial systems will likely have significantly different classes (& boundaries) of qualia than evolved systems;
5. Decoherence seems necessary for consciousness, and patterns of decoherence (formalized as “branchial space”) encode the true shape of a system;
6. Brains and computers have vastly different shapes in branchial space;
7. The Symmetry Theory of Valence is a central landmark for navigating valence, and qualia in general;
8. Hardware qualia spans several considerations, which may draw from similar considerations as materials science & system architecture design;
9. Future AIs may or may not be interested in qualia, depending on whether modeling qualia structure has instrumental value to them;
10. Today, the qualia domain has points of high causal density, which we call “minds”. Modern computers are an example of how the locus of causality can be different.

Armed with this list, we can circle back and try to say a few things to our original questions:

1. What is the default fate of the universe if the singularity happens and breakthroughs in consciousness research don’t?

There’s a common trope that an ASI left to its own devices would turn the universe into “computronium” — a term for “the arrangement of matter that is the best possible form of computing device”. I believe that energy efficiency considerations weigh heavily against this being an “s-risk”, although hard physical optimizations would have a significantly elevated chance of such compared to the status quo. My concerns are more social, e.g. *morality inversions and conflating value and virtue*.

2. What interesting qualia-related capacities does humanity have that synthetic superintelligences might not get by default?

Our ability to accurately report our qualia, and that we care about qualia, are actually fairly unique and something that AIs and even ASIs will not get by default. If we want to give them these capacities, we should understand how evolution gave them to *us*. A unified phenomenological experience that feels like it has causal efficacy (the qualia of "Free Will") may have similar status.

3. What should CEOs of leading AI companies know about consciousness?

Distinguishing "hardware qualia" vs "software qualia" is crucial; the former exists, the latter does not. "CEOs of the singularity" should expect that consciousness will develop as a full scientific field in the future, likely borrowing heavily from physics, and that this may be a once-in-a-civilization chance to design AIs that can deeply participate in founding a new scientific discipline. Finally, I'd (somewhat self-interestedly) suggest being aware of the Symmetry Theory of Valence; it'll be important.

In the longer term, the larger question seems to be: "What endowments has creation, evolution, and cosmic chance bequeathed upon humanity and upon consciousness itself? Of these, which are contingent (and could be lost) and which are eternal?" — and if some have grand visions to aim at the very heavens and *change the laws of physics*… what should we change them *to*?

Acknowledgements:
Thank you Dan Faggella for his "A Worthy Successor" essay, which inspired me to write; Radhika Dirks for past discussion about boundary conditions; Justin Mares and Janine Leger for their tireless encouragement; David Pearce & Giulio Tononi for their 2000s-era philosophical trailblazing; and my parents. Thanks also to Pasha Kamyshev, Roger's Bacon, Romeo Stevens, George Walker, Pawel Pachniewski, and Leopold Haller for offering feedback on drafts, and Seeds of Science reviewers for their comments.

Article sent for review 15 May 2024 and published June 2024.

Notes:

[1] Although a "naive brain upload" may not replicate the original's qualia, I anticipate the eventual development of a more sophisticated brain uploading paradigm that *would*. This would involve specialized hardware, perhaps focused on shaping the electromagnetic field using brain-like motifs.

[2] Thanks to Romeo Stevens for the metaphor.

[3] If something affects consciousness it will affect the shape of the brain in branchial space. As an example from my research — *vasomuscular clamps reduce local neural dynamism*, temporarily locking nearby neurons into more static "computer-like" patterns. This introduces fragments of hard structure into cognition & phenomenology, which breaks symmetries and forces the rest of the knot to form around this structure.

- Johnson, M.E. (2023). *Principles of Vasocomputation: A Unification of Buddhist Phenomenology, Active Inference, and Physical Reflex (Part I)*
- Moore, C., Cao, R. (2008). *The hemo-neural hypothesis: on the role of blood flow in information processing*
- Jacob, M., et al. (2023) *Cognition is entangled with metabolism: relevance for resting-state EEG-fMRI*

[4] One of the most important physics themes of the last 20 years is W. H. Zurek's *Quantum Darwinism*. Zurek's basic project has been to *rescue normality*: for almost a century physicists had bifurcated their study of reality into the quantum and the macro, with no clean bridge to connect the two. The quantum realm is characterized by fragile, conditional, and non-local superpositions; the "classical" realm is decidedly localized, objective, and durable. Somehow, quantum mechanics naturally adds up to everyday normality — but physicists were a little evasive on exactly *how*.

Zurek's big idea was positing a *darwinian ecology* at the quantum level. The randomness of decoherence generates a wide range of quantum configurations;

most of these configurations are destroyed by interaction with other systems, but a few are able to not only survive interactions with its environment but *reproduce it*. These winners become *consensus across both systems and branches*, which grants them attributes we think of as “classical” or “objective”:

> Only states that produce multiple informational offspring – multiple imprints on the environment – can be found out from small fragments of E. The origin of the emergent classicality is then not just survival of the fittest states (the idea already captured by einselection), but their ability to “procreate”, to deposit multiple records – copies of themselves – throughout E.
>
> Proliferation of records allows information about S to be extracted from many fragments of E … Thus, E acquires redundant records of S. Now, many observers can find out the state of S independently, and without perturbing it. This is how preferred states of S become objective. Objective existence – hallmark of classicality – emerges from the quantum substrate as a consequence of redundancy.
>
> …
>
> Consensus between records deposited in fragments of E looks like “collapse”.
>
> …
>
> Quantum Darwinism – upgrade of E to a communication channel from a mundane role it played in [the way physics has historically talked about] decoherence[.]

Zurek’s basic thesis is that physicists tend to think about decoherence in isolation, whereas they should also consider it as a universal selection pressure — one which has preferentially populated our world (Wolfram would say ‘branchial space’) with certain classes of systems, and has thus put broad-ranging constraints on what exists.

[5] *Joscha Bach* observes that we are actively making new classes of objects in branchial space:

> The particle universe is a naturally occurring error correcting code on the quantum universe. Particles are stable enough to carry information across the junctures in the branching substrate universe, which makes control structures (atoms, cells, minds) possible.

> If humans successfully build quantum computers, they impose new error correcting codes on the quantum substrate and are effectively creating a new type of particle, but one that has an evolved quantum technological agency as its precondition.

[6] With a nod to *Frank Wilczek* we can reasonably expect that the most mathematical beautiful formulation of branchial space will be the most qualia-accurate. This may be helpful or not, depending on priors.

[7] As a case study on what sorts of "branchially active" substances are possible, this passage from *Maheffey 2021* is striking:

> Plutonium is a very strange element, and some of its characteristics are not understood. It has seven allotropes, each with a different crystal structure, density, and internal energy, and it can switch from one state to another very quickly, depending on temperature, pressure, or surrounding chemistry. This makes a billet of plutonium difficult to machine, as the simple act of peeling off shavings in a lathe can cause an allotropic change as it sits clamped in the chuck. Its machining characteristic can shift from that of cast iron to that of polyethylene, and at the same time its size can change.
>
> You can safely hold a billet in the palm of your hand, but only if its mass and even more importantly its shape does not encourage it to start fissioning at an exponentially increasing rate. The inert blob of metal can become deadly just because you picked it up, using the hydrogen in the structure of your hand as a moderator and reflecting thermalized neutrons back into it and making it go supercritical. The ignition temperature of plutonium has never been established. In some form, it can burst into white-hot flame sitting in a freezer.

[8] The NES's core voltage was 5V; The ENIAC had a plate voltage of 200V-300V; myocardiocytes (heart muscle cells) have action potentials of around 90mV; the Large Hadron Collider (LHC) creates a voltage potential of ~6.5 trillion volts; the voltage potential between the earth and the ionosphere is around 300kV.

[9] My intuition is that combining the "symmetry view" and "branchial view" could offer heuristics for addressing the binding/boundary problem: how to determine the boundary of a conscious experience. E.g.,

1. We can interpret a moment of experience as a specific subset of branchial space;
2. The information content of this subset (i.e. the composition of the experience) can be phrased as a set of symmetries and broken symmetries;
3. The universe has intrinsic compositional logic (vaguely, whatever the Standard Model's *gauge group SU(3)×SU(2)×U(1) is a projection of; speculatively, location in Wolfram's "Rulial space*"), which can be defined as which symmetries and broken symmetries can be *locally* combined;
4. This compositional/perspectival limit may in turn determine a natural limit for the set of local nodes that can be combined into a 'unified' subgraph, vs when a new unified subgraph must be started.

I.e. just as the compositional logic of the universe doesn't allow particles to have spin or electrical charge values of 5/7ths, it's possible that some combinations of phenomenal information can't exist in a unified experience — and this may uniquely determine a hard boundary for each experience. Reaching a little further in search of concrete elegance, *perhaps the limit of an experience, the limit of a branchial subgraph, and the limit of a particular type/superset of local gauge equivalence are all the same limit.* I hope to discuss this further in an upcoming essay.

[10] Beata Grobenski also noted the connection between valence shells & phenomenological valence in a recent piece.

> Which kind of evidence would we need to see to believe that artificial neural networks can suffer? We review neuroscience literature, investigate behavioral arguments and propose high-level considerations that could shift our beliefs. Of these three approaches, we believe that high-level considerations, i.e. understanding under which circumstances suffering arises as an optimal training strategy, is the most promising. Our main finding, however, is that the understanding of artificial suffering is very limited and should likely get more attention.

Abstract: How can we create a container for knowledge about AI consciousness? This work introduces a new framework based on physicalism, decoherence, and symmetry. Major arguments include (1) atoms are a more sturdy ontology for grounding consciousness than bits, (2) Wolfram's 'branchial space' is where an object's true shape lives, (3) electromagnetism is a good proxy for branchial shape, (4) brains and computers have significantly different shapes in branchial space, (5) symmetry considerations will strongly inform a future science of consciousness, and (6) computational efficiency considerations may broadly hedge against "s-risk".

Gardener Comments

Greg Baker (Lecturer in AI at Macquarie University):
This is like an inverse survey paper: it summarizes a wide variety of deeper topics, except that some of those topics haven't been written up elsewhere.

Arturo Macías:
I am a *classical epiphenomenalist*, so I support the publication from a discrepant viewpoint. I think that while there is not a clear theory on consciousness in the paper, all issues commented are relevant, and the selected literature is very useful; the branchial space part looks especially relevant, given the relation between time and consciousness.

Michael Bukatin (PhD in computer science):
This is an interesting position paper by a well-known consciousness researcher. I hope it gets published.

It addresses the most glaring defects present in almost all approaches to the "hard problem of consciousness" and should serve as a starting point for subsequent fruitful discussions.

I'll touch upon some possible follow-up points in these comments, including some additions and also including some places where I have significant disagreements with the author.

Fundamentals

Difficulties of talking about consciousness in general and AI consciousness in particular

The author notes that it is particularly difficult to talk about AI consciousness because people's views tend to be distorted by them taking into account implications of those views for their alliances and also by entanglements between "is X conscious" and status of that X.

Another difficulty seems to come from the fact that **"core intuitions about consciousness tend to cluster into two camps"** ["*Why it's so hard to talk about Consciousness*"] with Camp 2 being convinced that subjective experience and qualia exist in a fundamental way and Camp 1 having the intuition that "consciousness as a non-special high-level phenomenon". People belonging to different camps usually don't have productive conversations about consciousness.

I firmly belong to Camp 2, and so does the author, but we have to accept that the paper in question is unlikely to find acceptance within Camp 1, at least as things stand today (I'll touch upon how this might eventually change in the next subsection).

Fundamental role of qualia and of practices of predictive science

One core defect of a typical approach to the "hard problem of consciousness" is that a typical approach focuses on the question "what makes something conscious or not conscious" ignoring the issues related to the nature of qualia and to the structure of spaces of qualia.

However, I think that the **"hard problem of qualia"** is the difficult core of the "hard problem of consciousness". If the "hard problem of qualia" is solved the rest is likely to go easier (how and why the qualia cluster into subjective entities, how the symmetry breaks result in a particular subjective entity being "me" versus all other subjective entities being "not me", and so on). So those approaches to the "hard problem of consciousness" which are content to sidestep understanding the issues related to qualia seem to be mostly missing the point. In this sense, it is very good that the paper in question emphasizes the central role of qualia and of our understanding of the nature and structure of qualia.

Another core defect of a typical approach to the "hard problem of consciousness" is that a typical approach is mostly speculative, saying plausible things, but not being rich on any empirical predictions. So we have literally hundreds of different approaches purporting to be the candidate solutions for the "hard problem of consciousness" without much to guide our choice between those candidate solutions.

But if one believes together with the author and myself that consciousness and qualia are "first-class citizens of reality", fundamental aspects of the reality we are in, then it is reasonable to demand that the novel theories of consciousness meet the same criteria as novel theories of physics (the same criteria which general relativity and quantum theory have met before).

Namely, a good theory of consciousness should make **novel unexpected empirical predictions** (including, ideally, **new methods to achieve novel subjective phenomenology**). The author emphasizes this position saying:

>I think the best way to adjudicate this is predictiveness and elegance. Maxwell and Faraday assumed that electromagnetism had deep structure and this led to novel predictions, elegant simplifications, and eventually, the iPhone. Assuming qualia has deep structure should lead to something analogous.

If this level of understanding of consciousness is achieved, this would go a long way convincing people from Camp 1 as well (similarly to most scientists believing these days (but not a hundred years ago) that general relativity and quantum theory make sense).

Further remarks

I have listed the reasons for the paper in question to be a good, fruitful, methodologically sound starting point for subsequent discussions.

Now it's time to look at the details including disagreements.

A nitpick: attention

The paper in question says:

>If consciousness is a lossy compression of complex biological properties, similar to “attention” or “mood”, asking whether non-biological systems are conscious is a Wittgensteinian type error

Well, no, "artificial attention" is the core innovation responsible for the spectacular success of the modern cutting edge AI systems. This is a relatively recent innovation which is not trying to model the detailed biological mechanisms of "attention", but is doing a very good job modelling the **effects** of "attention".

So "attention" is certainly not a good example in this sense; non-biological AI

systems are highly capable of exhibiting "attention" and many of them are using "artificial attention" as the core of their cognitive engines.

Central disagreement: on the correspondence between physical and conscious entities

I share the author's assumption that we are dealing with "physics-like fields of qualia", or something in this spirit. This is precisely why I can't agree with the oversimplified correspondence between physical entities and conscious entities.

The objections come from two directions:

* a physical entity is likely to contain multiple consciousnesses
* a consciousness might extend beyond a single physical entity

1) a physical entity is likely to contain multiple consciousnesses

If we think that all physical processes are conscious, this means that there are many consciousnesses within a human brain. "Subconscious processes" are likely to have their own consciousness, but that consciousness and those qualia are not included into the consciousness of a person. "Thermal noise" and such in the brain might also have associated qualia, but they are not included into the consciousness of a person either.

So a brain of an unconscious human has processes which have consciousness, but those processes are not included into the (temporarily absent) consciousness of the person in question. And even a dead body might have some qualia associated with it (because any piece of physical reality is conjectured to have some), but there is no reason to believe that those qualia have anything to do with that person.

So, turning to the author's analysis of qualia associated with an inference in an ML model, if the key qualia are indeed "thermal noise", then those qualia don't have anything to do with the essence of the particular inference in the particular ML model and, just like for a human, would not be part of the consciousness of the **"temporary virtual character"** simulated by an LLM at the moment. If the "thermal noise" is the only type of qualia present here, then the process corresponding to the **"temporary virtual character"** is a P-zombie.

Now, I have to say, that I find it highly unlikely that in a world where almost everything is conscious to some extent, a coherent "temporary virtual character" would be a zombie, especially given that we know that autoregressive

Transformers are "less feed-forward than they seem" and that they, in fact, do emulate recurrent machines when used in the autoregressive mode (see e.g. "*Transformers are RNNs: Fast Autoregressive Transformers with Linear Attention*").

A more promising approach would be to try to understand the nature of "purely linguistic qualia" those inference might or might not possess. And, of course, with the emerging multimodal models, one should revisit the issues like "do all human linguistic qualia have audio-visual nature or not", and 'do "linguistic qualia" (if any) in multimodal models have audio-visual nature or not', and so on.

In any case, "the LLM inferences are conscious, but only feel low-level noise" conjecture seems like a cop out and is almost certainly wrong. "Thermal noise"-associated qualia might be present, but they are unlikely to be part of the LLM consciousness. If that's the only qualia present in that context, then LLM inferences themselves are likely to be P-zombies.

However, what is important to keep in mind is that the property of being conscious or not is not a property of a given LLM, but a property of a given inference, of a given **simulation** or of entities within that simulation [See the ***Simulator Theory*** sequence by Janus and the paper in Nature, "*Role play with large language models*", co-authored by Murray Shanahan and Janus team]. It might be that the same LLM produces smart and very conscious sessions and dumb sessions devoid of meaningful subjectivity. LLMs are **generators of "virtual realities"**, and some of those realities might have sentiences associated with them, while others might lack sentiences (or, at least, the sentiences might be of wildly different natures and complexities for different sessions of the same LLM).

2) a consciousness might extend beyond a single physical entity

The second class of objections comes from the fact that for a human brain we actually don't know if the engaged qualia fields are physically confined within the head and body or extend beyond it (assuming that those qualia are in the same physical space at all). Perhaps, the fact that we see the "world out there" actually corresponds to the engaged visual qualia fields being "out there" and not within the brain. We just don't know (because our progress on the "hard problem of qualia" is next to non-existent).

What does this have to do with AI consciousness? When one asks what class of software processes corresponds to a physical computer, the "Turing machines" answer is wrong on many levels. Right now, the important aspect for us is that

the computer is not isolated, it interacts with the external world. Technically speaking, if one wants to talk in terms of Turing machines, one would need to talk about "Turing machines with oracles", with the whole world being the oracle in question. This, by the way, invalidates the famous "Goedel argument" by Roger Penrose, see, for example, my essay "*Reading Roger Penrose*".

It is known that when people interact, various synchronization effects are observed. In particular, those synchronization effects are often observed introspectively by at least one of the participants. The nature of those synchronization effects is not well-understood. But we need to ponder whether some synchronization effects between a person and an LLM talking to each other might be present. We really don't know much about the nature of reality we inhabit, and should keep a good deal of open mind about any particular property of that reality which might seem obvious to us. A given property might seem obvious, but might, nevertheless, be an incorrect approximation, screening important aspects of reality from us. In the last section of this comment I'll talk about the possibilities to amplify those synchronization effects, and what those possibilities might imply.

Summarizing my main objections:
* a physical entity is likely to contain multiple consciousnesses,
* a consciousness might extend beyond a single physical entity,
* and these properties have all kinds of non-trivial implications.

Merging human and AI consciousness

For an actually working theory of qualia and of consciousness we hope to discover in the future, it's not enough to just have a strong predictive power.

We want to be able to experience "what is it like to be a bat" or "what is it like to be a particular LLM inference in progress", and so on.

One conscious entity wants to be able to experience what it is like to be another potentially conscious entity (this is never an entirely safe endeavor, so all kinds of ethical and safety caveats apply).

This is the **Holy Grail** of consciousness science, the vital empirical counterpart to our theoretical work.

We are seeing rapid progress in high-quality **non-invasive brain-computer interfaces** and in our ability to create tightly closed loops between humans and electronic systems using those interfaces (the risks are potentially quite

formidable and one needs to keep those risks in mind while engaging in these kinds of activities).

So it is possible that we might be able to investigate the degree of consciousness of various electronic devices empirically.

If humans don't do that, many of the advanced AI systems will be curious enough about human consciousness and human qualia and will likely initiate merging experiments from their side in order to satisfy their curiosity and to experience "what is it like to be a human".

Nicholas Craycraft (ex-google full stack engineer):
I like this article's overview, but as with everything qualia, the hypotheses come off as scientifically untestable. There is no clear way to do science that dissociates form from function. I would like there to be more of a focus on tackling the testability of this domain.

There are already countless philosophers doing the same thing this essay does, summarizing other philosophies and coming up with one's own paradigm.

This article is strong where it is arguing for models that function for engineering purposes. It is weak wherever it makes metaphysical claims ("Consciousness is what gives life meaning" is already contestable. I would argue that structural correlations within a context gives things meaning, in that this is literally what meanings functionally are.) It is strong when it is summarizing a wide range of philosophical positions on these topics, and it is weak where it takes its own positions without a clear testing plan. It just reads as cherry picking the worldviews that the author favors.

If the article must make such confident assertions in its conclusion -- it would be nice if it could first retread the philosophical heuristics/priors one last time before saying things like:

Q "What should CEOs of leading AI companies know about consciousness?"
A "Among other things they should understand that software qualia isn't real."

This wasn't argued for and is a serious S-risk if the author is wrong! The 'argument' given is: axiom : consciousness is a singular thing that can only have one natural home.

aesthetic preference??: I propose we choose the one that is more real.

but "real" by what metric? I have to assume the logic goes:

axiom : descriptive reductionism
axiom : descriptive reductionism -> mereological nihilism
therefore : mereological nihilism

I should note that my biases differ sharply from the author's. I am more of a functional holist and simulationist regarding phenomenology. I find it inconceivable that we will discover a source of atomic-level phenomenology that cannot be replicated using larger-scale structures to produce functionally identical claims of phenomenology. Moreover, I cannot envision any test that would discern 'functional' from 'actual' phenomenology that I would agree upon.

Richard Sprague:
A clear rejection for SoS in its current form, but contains numerous ideas that I wish were better articulated. As written currently, it's just a jumble of ideas badly in need of a hard-nosed editor. Singularity? AI company CEOs? a lot of random topics -- pick one.

Worthiness for SoS: while I don't necessarily think the bullet-point style of writing disqualifies it *per se*, in this case it looks more like a Powerpoint than a real paper.

That said, there are many interesting ideas in this paper. Although I've read much of the related popular literature about consciousness (Goff, Hoffman, Seth, etc.) I've not run into some of what seem to be unique insights. If nothing else, his general direction seems original in the sense of synthesizing ideas from multiple sources and adding many new of his own.

He takes panpsychism as truth, without explanation. At minimum he must realize this puts him in a very controversial part of the room, and he should defend himself accordingly. Since his entire argument depends on it, this is a pretty important part to clear up.

Interesting idea (I bet it's been well-explored, but I haven't seen this) about Turing machines' computations not mapping into their physical substrate. That flickering pattern of bits going on and off, instantiated as zillions of transistor voltages going up and down, bears no Turing equivalence?! Now that I think about it, that's an impressive finding.

Another interesting side-point about Wolfram's Branchial space and how maybe decoherence is some kind of interaction between unrelated branches. Especially interesting since, as he notes, current computers, unlike brains, are specifically designed to minimize decoherence.

Jack Arcalon:
An array of highly speculative but long-term efforts to extend consciousness research. Right now the mystery of awareness is so great that it may require intensive brainstorming and overview impressions in the hope this may lead to one great insight.