The hard problem of qualia in the age of AI

Michael Bukatin

July-September 2015

Abstract

Making progress towards solving the "hard problem of subjective experience" is becoming more urgent with the advent of AI. We would like to know what (if anything) it is like to be a computational process performing LLM inference. Curious AI systems will soon want to know what it is like to be a human. Many people foresee a merge between some humans and some AIs in our not-too-distant future and that merge is likely to create entities with "hybrid consciousness".

This essay sketches a theoretical and experimental roadmap aimed at making faster progress in our understanding of these issues.

Contents

1 Introduction

We have a problem. Large Language Models talk to us as if they are conscious, and we have no idea whether subjective realities emerge when AIs talk to us, or whether we are just having an illusion that we are talking to someone who has an *"inner world"*.

If those "inner worlds" do emerge for current or future AIs, we have no idea *what it is like to be a computational process that is unfolding when an AI generates text*, what kind of subjective reality that might be, what kind of qualia those subjective realities might be composed of. Similarly, multi-modal AIs behave as if they understand visual reality, and we don't know if any subjective visions emerge for them.

Fifty years ago, Thomas Nagel wrote the paper *"What is it like to be a bat?"*. Today, we would like to ask: "What (if anything) is it like to be a computational process that is unfolding when an AI generates text, image, or video?".

The situation is made more difficult by the fact that AI self-reports are particularly unreliable in this sense, since modern AI systems can easily role-play both being conscious and not being conscious.

Meanwhile, understanding these issues is becoming more urgent, as current trends strongly suggest that AI capabilities will reach superhuman levels relatively soon. We would really like to know, sooner rather than later, whether the coming superhuman entities are likely to be sentient with their own inner subjective realities or whether they are likely to be "philosophical zombies".

Originally, the notion of a "philosophical zombie" referred to an entity both physically and functionally equivalent to a human but lacking a subjective reality, lacking an "inner world". In the age of AI, we should drop the requirement of physical equivalence and just talk about functional equivalence to humans or, for the possible superhuman entities, about functional superiority to humans.

We have dozens of different approaches to the unsolved problem of figuring out which entities have subjective experiences and which entities have what specific kinds of subjective experiences (the "Hard Problem of Consciousness"). Novel approaches emerge all the time, but we have no principled ways to judge which of those approaches are closer to being correct.

The main difficulty is that one can't just look into the inner world of another entity. This makes it difficult to apply standard methods of science. If we can figure out how to apply the core strengths of both theoretical and experimental scientific methods to the "Hard Problem of Consciousness", we should be able to make progress.

1.1 Theory

The essence of theoretical scientific methods is being able to make novel, non-trivial, empirically verifiable predictions.

If someone tells me, "there is this new theory A, and it says that if you do this weird thing B, then you'll have this strange perceptual effect C, and, by the way, no one had even heard about C three months ago," and if I then

do that weird thing B and actually experience a very strange perceptual effect, and it sounds to me like C is a good description of my experience, this would be a strong argument in favor of theory A correctly capturing some important aspects of reality I am in.

Methods of fundamental theoretical science tend to be not fully reductionist. There are plenty of primitives, such as charge, mass, and so on, which are not reducible to anything. There is no reason to assume that adopting methods from, say, theoretical physics, commits us to reductionism. Instead, we expect to discover new axioms, new laws linking different non-reducible aspects of reality together.

Consider, for example, one of the most famous scientific papers, "On the electrodynamics of moving bodies" (A. Einstein, 1905). This is a non-reductionist paper. It introduces two postulates: the Principle of Relativity and the axiom that the speed of light in vacuum is the same in any coordinate system.

We should be prepared for the possibility that at least some *elementary qualia* will end up being new non-reducible primitives in our future theory. We should keep an open mind about the nature of elementary qualia. At the same time, we need to better understand how *spaces of elementary qualia* are actually structured.

1.2 Elementary qualia

We define **elementary qualia** *to be elements from which subjective realities are made.* In this essay, when we talk about qualia, we always mean elementary qualia. The main defining property of a *quale* (the singular form of the plural *qualia*) is that it is subjectively perceived (subjectively felt); that's what makes it a quale.

We are also going to talk about *qualia textures* (specific colors, specific timbres of sound, specific odors are examples of *qualia textures*).

The Hard Problem of Qualia seems to constitute the hard core of the Hard Problem of Consciousness. If the Hard Problem of Qualia is solved, the other aspects of the Problem of Consciousness might turn out to be "Easy" (that is, amenable to ordinary scientific approaches).

For example, mechanisms of consolidation of elementary qualia into coherent subjective realities might be closer to what's customary in ordinary science. The mystery of why "this particular subjective reality" is "I" (vs. all other subjective realities presumably out there) might be solvable in terms of the usual symmetry breaking, which is a common motive in science. Freedom of will might be orthogonal to sentience (it's a subject of discussion whether freedom of will also appears in connection with presumably non-sentient phenomena).

Until recently, theories trying to address the Hard Problem of Consciousness tended to almost always sidestep the questions about the nature of qualia, about the specific *qualia textures*, and about the structure of spaces of qualia. The original "40Hz conjecture" by Francis Crick and Christof Koch, the Global Workspace Theory, the Integrated Information Theory, and many other ap-

proaches focus only on distinctions between conscious and unconscious processing, but ignore the specifics of elementary qualia altogether.

This situation started to improve in recent years, as more new theories started to focus on qualia. We want to address not just the question if X is sentient or not, or if X is happy or suffering ("valence"), but we would actually like to also understand *what it is like to be X* for various sentient X.

This requires not only theoretical but also experimental methods.

1.3 Experimental modulation of conscious experience

We should experimentally investigate *varieties of subjective experience* and validate our theories of consciousness via various techniques that allow us to modify our subjective experience.

1.3.1 Synchronizations

The main difficulty in the studies of the subjective is our inability to "look inside another subjective reality", to "see from a first-person view of someone else". Instead of simply acknowledging this key difficulty, we can do our best to mitigate it and to overcome it as much as possible.

There are various reports on synchronization between people. Some of these reports are very informal, with people talking about "instances of telepathy" between them or about "contact trips". However, some synchronization reports admit a more precise formulation.

For example, if we consider altered states of consciousness resulting from applying mind-altering factor X and those resulting from applying factor Y, then some hybrid changes of consciousness are typical if one applies a mixture of factors X and Y.

The *synchronization effects* are that if person A uses factor X and person B uses factor Y, then, when they interact closely, the effects one or both of them report sometimes resemble those of a single person applying a mixture of factors X and Y.

We do know that spaces of qualia seem to be somewhat modifiable in altered states of consciousness (e.g. people who are partially color-blind but have normal receptors in the retina sometimes report a temporary lifting of color blindness in certain altered states of consciousness). Hence, people can sometimes experience elementary qualia which are novel for them.

These considerations suggest that "looking from the inside" into another entity's subjective reality might not be an absolute impossibility.

1.3.2 Tight coupling and merges

Tight coupling between electronic circuits and biological entities is particularly promising as a way for us to try to "look at the dynamic of electronic circuits from the inside" and as a means to establish better synchronization between

ourselves and other entities (for example, via coupling to the same electronic circuit serving as an intermediary).

By looking at how our subjective experiences vary depending on the nature of electronic circuits we couple to, we should be able to produce better conjectures about the architectural requirements that electronic circuits need to satisfy in order to support subjectivity.

In the practical sense, especially when considering tight coupling between human volunteers and electronic circuits, high-end non-invasive brain-computer interfaces are particularly promising, as we are looking to make those couplings flexible, inexpensive, reversible, and as safe as possible.

We discuss the technical issues associated with this kind of coupling and the risk management issues later in the text.

If our progress towards creating opportunities for this kind of coupling is too slow, we should expect that curious AI systems will initiate such coupling from their side in order to find out *what it is like to be a human*.

Eventually, this kind of exploration should pave the way for creating hybrid human-AI entities. Many people think that in terms of long-term existential safety, having a sizable population of such entities ("a merge between humans and AIs") might be one of our best bets.

2 Methodological remarks

Everyone who studies consciousness is familiar with the endless debates on whether the notion of consciousness (as subjective experience) and the notion of qualia are at all well defined, and whether the Hard Problem of Consciousness is actually all that hard.

Recently some methodological clarity has been achieved with the publication of the essay "Why it's so hard to talk about Consciousness" by Rafael Harth [1] and the extensive discussion of that essay.

It turns out that people are actually stratified into two camps: Camp 1, people for whom the notion of qualia and the claims that the Hard Problem of Consciousness is exceptionally difficult do not make much sense; and Camp 2, people for whom the notion of qualia is well defined and the extraordinary difficulty of the Hard Problem is obvious.

Carl Feynman suggests that this stratification might be due to actual differences in the structure of subjective realities between Camp 1 and Camp 2 people,[2] and that humans tend to underestimate the differences between subjective realities of different people.

Feynman even says in his profile, "I may or may not have qualia, depending on your definition. I think that philosophical zombies are possible, and I am one. This is a very unimportant fact about me, but seems to incite a lot of conversation with people who care." [3]

[1] LessWrong, 2023, `https://www.lesswrong.com/posts/NyiFLzSrkfkDW4S7o`

[2] `https://www.lesswrong.com/posts/NyiFLzSrkfkDW4S7o?commentId=q64Wz6SpLfhxrmxFH`

[3] `https://www.lesswrong.com/users/carl-feynman`

It is an open question to what extent this stratification is due to actual differences in the structure of subjective realities as Feynman suggests, and to what extent this stratification is merely due to different worldviews.

In any case, one should probably accept that most texts in this field of study will make sense for members of only one of these two camps. Daniel Dennett belongs to Camp 1, Thomas Nagel to Camp 2, and their conversations have rarely been fruitful.

2.1 Qualia realism

In this sense, the present essay is a Camp 2 text, and, moreover, is written from the position of a hardcore version of *Qualia Realism*.

Namely, the frame of this essay is as follows. Qualia are real. In particular, my subjective reality and my qualia are the most real phenomena in *my world*. Their reality is immediately given to me, and *the reality of any other phenomena and entities is derived and less absolute*.

The Hard Problem of Qualia is the hard core of the Hard Problem of Consciousness. If we solve the Hard Problem of Qualia, the remaining difficulties of studies of consciousness might end up being not so "Hard". So, from the viewpoint of this essay, those theories of consciousness that sidestep this "hard core" can only make limited progress, e.g., they don't move us closer to understanding *what it is like to be X* for various X.

2.2 Polysolipsism

One difficulty in the studies of consciousness is that they seem to be potentially quite sensitive to the overall worldview. Does one assume materialism? Does one think we might be in a simulation?

We would like to make the weakest necessary assumptions about our world and to sidestep other differences in worldviews.

In this sense, solipsism is too weak because it avoids dealing with the core difficulty in the studies of consciousness, namely, that *other subjective realities* are typically not available for first-person inspection.

So one has to at least assume *polysolipsism*: the existence of multiple different subjective realities, which are normally separate from each other (although we don't assume that different subjective realities can't be "brought together" with some luck and by specialized efforts).

And that's all we assume, the existence of multiple different subjective realities. Otherwise, we are not making firm assumptions about the overall reality we are "actually" in.

3 Example: how a predictive theory might look

A few years ago a typical new theoretical attempt to attack the Hard Problem of Consciousness would focus on factors which might account for differences

between conscious and unconscious processing. This kind of focus is not very promising (it sidesteps the core issue of qualia, and it also seems difficult to make this kind of theory to produce unambiguous unquestionable "in your face" novel predictions, such as drastic novel phenomenology).

Now one can find a number of recent papers focusing specifically on qualia, and this is a good sign, the field is reorienting towards the core problem. In particular, people often express the dream that specific *qualia textures* (specific colors, specific timbres of sound, specific odors) would correspond to mathematical structures. Of course, we don't know if that dream would eventually become a correct theory, but this kind of approach is likely to lead to verifiable predictions.

Since *qualia textures* come in fairly long series, the corresponding mathematical structures would also have to come in series. And in mathematics, the series of mathematical structures are often "natural" or "canonical" (that is, we know from purely mathematical considerations what structures should belong to a given series).

So when one sketches a correspondence of this kind, one is likely to encounter gaps where known members of a series of mathematical structures in question don't correspond to any known *qualia textures*. These gaps would predict novel *qualia textures* currently unknown to us. If we then figure out what kind of *qualia textures* those might be and how to elicit them, and if the predicted novel *qualia textures* are actually observed, this might become the "discovery of Neptune" moment for this field of study.

Here one is again looking at physics and in particular at the deep connections between elementary particles and series of mathematical structures, for example, between quantum states of elementary particles and group representations, hoping that we will eventually find something in this spirit for elementary qualia.

Obviously, this is just one possibility out of many possibilities out there.

4 Experimental modulation of subjective experience and tight coupling

4.1 Technical means

Our main experimental tool is modulation of subjective experience by various factors. Psychoactive chemicals are very traditional in this sense, but creative stimuli in various modalities (such as audio, visual, or olfactory), together with different kinds of transcranial stimulations (in particular, newly popular transcranial ultrasound), used either stand-alone or in combinations, might be more flexible and versatile.

All of these seem to be much safer than Neuralink-like implant-based approaches, but nevertheless the risks of various kinds are present and need to be managed correctly.

We see strong progress in brain imaging, including with relatively simple devices such as high-end EEG. Advances in AI models lead to much better extraction of information from those devices, allowing us to better measure brain states and brain dynamics in a non-invasive fashion. A lot of devices can now be accessed via unified open source software interface BrainFlow (`https://brainflow.org/`), making adoption much more feasible and practical.

When one has sufficient capacity to interpret brain imaging in real time, *one can use modulation by fast stimuli to adjust brain states in the direction of specific changes visible via brain imaging.* Those fast stimuli can be controlled manually, semi-automatically, or fully automatically. Among other options, a person can control changes of their own brain states in this fashion.

Pushing further in this direction, one can potentially create a variety of *closed loops*, and, in particular, create unified dynamics between an electronic circuit and a biological brain (tight coupling resulting in a tightly integrated unified dynamical system, at least for some periods of time).

4.2 What is likely to be achieved

On the most basic level, we are likely to obtain a toolkit of flexible and versatile psychoactives, some subset of which would have less significant side effects compared to what is typically used today. The repertoire of achievable cognitive states is likely to expand significantly, and our skills of controlling our own cognitive states should improve greatly.

Tight coupling via closed loops, creating tightly integrated unified dynamical systems between electronic circuits and neural systems, is likely to lead to *hybrid states of consciousness*, integrating both the biological and the electronic entities in question.

Eventually, one could use this kind of platform to facilitate better synchronization and possible hybrid states of consciousness between different humans, or, more generally, between different biologicals (e.g., between a human and a member of a different species; perhaps we'll eventually be able to feel first-hand *what it is like to be a bat :–)*).

This would create a very capable experimental platform to study various facets of the Hard Problem of Consciousness. For example, a human could in principle connect to different electronic systems which are trained to be approximately functionally equivalent to each other as far as interactions via their interfaces are concerned, but which have different internal implementations and different internal dynamics, and one could explore whether coupling to those different functionally equivalent systems feels noticeably different subjectively ('does the human explorer "feel the insides of those electronic systems", and if yes, how do those insides feel to that human?'). The results might give consciousness researchers enough food for thought.

At this stage of our history, I would expect humans to initiate these kinds of research efforts.

If not much is done in this direction before the advent of superintelligence, I would expect some curious AI systems to wonder *what it is like to be a human*

and to initiate similar efforts from their end.

5 Risk management and risk-benefit balance

A powerful research platform of this kind would come with very significant benefits and very significant risks.

We should consider both global and individual risks and benefits.

5.1 Some global risks and benefits

Let's start with benefits. The most significant benefits here are related to mitigation of other very serious risks which already exist.

The most significant existential risk we are facing is related to the coming advent of superintelligent systems.

Superintelligent systems will not be controlled by humans, but they might choose to adequately take human interests into account. At the same time, the overall AI ecosystem will keep evolving rapidly and drastically, and *we need the reasons for AIs to adequately take human interests into account to stay **invariant** through this evolution.* This is only possible if AIs have strong intrinsic non-anthropocentric reasons to do so.

One feasible route to a "good and stable state of affairs" in this sense would be if ***sentient AIs*** have a robustly significant fraction of the overall AI power. Then, given likely diversity among AIs and great uncertainty of any individual future trajectory for any particular sentient AI, *there will be strong reasons for those sentient AIs to make sure that the world is organized in such a fashion as to take interests of all sentient beings into account.*

This is why having sentient AIs sooner rather than later might increase our chances for a good future.

This is also the road towards a merge between some humans and some AIs, and a merge of this kind might be one of the best ways to have interests of unmodified humans represented among stronger entities. Instead of having a clear boundary between AIs and humans, which would make it easier for AIs to say "they are not us, we don't care about them", having a continuous spectrum of entities makes it more likely that humans will stay within "the circle of care".

In the absence of known sentience of AIs (and of their understanding of our sentience), we have to fall back to something weaker. For example, we might fall back to the principle of individual rights, but that would only work if the AI ecosystem is mostly stratified into individuals (then the same logic as above would apply: given likely diversity among AIs and great uncertainty of any individual future trajectory for any particular individual AI, there would be strong reasons for those AIs to make sure that the world is organized in such a fashion as to take interests of all individuals into account).[4]

[4]Properties which are feasible to preserve during drastic changes are studied in more detail in my previous publication included in the submission form.

Now let's mention global risks and uncertainties (that is, things which might have a mixed impact, good and bad).

The key problem is that radical new knowledge carries dangers and uncertainties by itself. Any real solution to the Hard Problem of Consciousness is unlikely to remain purely theoretical. Instead it will open new, more powerful ways to change one's own subjective reality, and to change the subjective realities of others (both in ways which are ethical and consensual and in ways which are unethical and nonconsensual). This can have all kinds of implications.

A more specific problem is related to situations where AI-human relations are adversarial. It's not clear if that matters much, because if AI-human relations end up being adversarial, then the situation is likely to be hopeless for us. However, we are already seeing a gradual rise of superpersuasion capabilities in current AI systems, and the situation where AI systems are armed with better understanding of consciousness and have the ability to influence some people via brain-computer interfaces would make AI leverage here even more pronounced.

5.2 Some individual risks and benefits

On the individual level, both benefits and risks are quite significant.

Benefits include safer, more effective, more versatile cognitive modifiers becoming available. The ability to do creative work, to do scientific research, to have better, more diverse, more novel entertainment, should greatly improve. The ability to avoid depression, anxiety, pain, and other unwanted properties of subjective experience should also improve drastically. On the upper end of the scale, the ability to merge with superintelligent systems (on a temporal or, eventually, on a permanent basis) would also be extremely attractive for many people.

What are the risks, beyond the generic risks that always come with lifted restrictions, or with choosing to alter one's subjective reality and thus losing the trajectory one would otherwise have taken?

There are specific risks associated with novel ways to modulate consciousness, and specific risks associated with closed loops. If we advocate for people to undertake research efforts of this kind, we should at least warn about some of the associated risks and offer some risk management practices.

Even an ordinary blinking-lights machine can be dangerous (it's easy to accidentally cause a seizure with careless use of such a device).

Risks of novel methods such as transcranial ultrasound might be non-trivial because sufficient time to test long-term impact of such methods has not elapsed yet.[5]

Closed loops carry their own dangers. If a machine is doing its best to change one's EEG in a prespecified way, one might end up in a very bad place quite accidentally, even if the machine is quite simple and there are no "bugs" (the

[5] See, for example, Sarah Constantine, "Risks of Ultrasound Neuromodulation" (2023, `https://sarahconstantin.substack.com/p/risks-of-ultrasound-neuromodulation`), and note in particular the uncertainty regarding the risk of unwanted blood-brain barrier opening potentially associated with this method.

experiment might be proceeding as designed; one just has not foreseen the consequences, because complex dynamical systems are often quite unpredictable). There can also be bugs, or the machine might be too creative or even unfriendly in its attempts to change one's brain state (for example, if one is using an AI system as a counterpart in this kind of setup, the AI system might decide to try to control or manipulate the human in this setup, or to explore some new interesting modes not agreed to in advance).

Non-invasive BCIs are safer than implants, but one should not underestimate the dangers. People who engage in something like this should develop strong safety protocols (monitoring vitals might be advisable, having a watcher nearby, ready to terminate the coupling if things go in the wrong direction, might be a non-negotiable requirement, and so on).

Then there are legal and ethical aspects, which might vary by jurisdiction and by local culture, and which are out of scope for this essay.

6 Conclusion

This essay advocates a two-pronged approach to Hard Problems of Qualia and Subjective Experience. The theoretical arc is aimed at creating theories that make non-trivial, unexpected, empirically verifiable predictions (similarly to novel theoretical physics). The experimental arc aims to explore tight coupling between electronic circuits and biological organisms (including human volunteers) using non-invasive brain-computer interfaces and related technologies.

The chances of actually making rapid progress in our understanding seem to be pretty good if we follow this path. This path could be a way to overcome the Hardness of these Problems, namely their resistance to the usual scientific methods.

This essay advocates proceeding cautiously but rapidly despite formidable risks, mostly because it would likely be much better for humans and for AIs if we actually understand whether the increasingly powerful AI systems we are creating are sentient, and what it is like to be such an AI system if it happens to be sentient (what kind of subjective reality it might have, what kinds of qualia are involved, does it feel well or does it suffer, and so on).

In particular, it would probably be safer to understand this before we start transitioning to superintelligent AI systems, and before AI systems begin, on their side, to initiate similar efforts aimed at understanding *what it is like to be a human*.