Tackling the Pretty Hard Problem: Could Robots Have Human-Like Consciousness?

Lenhart Schubert SCHUBERT@CS.ROCHESTER.EDU
Department of Computer Science, University of Rochester, Rochester, NY 14627

Abstract

Many commentators on the question of whether AI systems are or could be conscious have suggested that if such systems are sufficiently human-like in their interactions with people and the world, we might as well grant that they are conscious in more or less the human sense. I argue against such a conclusion, using thought experiments involving human "sensory mimics" – ones rather close to technological realizability. Such mimics have access to symbolic propositional information alone in certain modalities (such as for auditory or visual sensing), yet behave effectively as if they were fully endowed with those sensory modalities. I draw conclusions about the difference between mere symbolic situation modeling and modeling that integrates symbolic annotations with the perceptual patterns from which they are abstracted. Perceptual patterns can plausibly be viewed as time-varying vector fields with local geometries that in part determine the perceived subjective sensations (qualia). Carrying over these observations to AI agents, one can envisage agents with and without phenomenal consciousness, i.e., agents with and without integration of symbolic abstractions with perceptual patterns. This is a step towards solving the Pretty Hard Problem of consciousness.

1. Introduction

> "If ten years from now we have AI servants, if they have the consciousness of a toaster, that is, no consciousness at all, you can do whatever you want with them. But if they have some sort of consciousness, even a different sort than ours, ... then we shouldn't order them about, we have to respect them, we can't turn them off at will."
> – Paul Bloom (World Science Festival, 2025)

The age of intelligent machines is upon us, and the question of whether such machines are or could become conscious, often set aside in the past as pointless speculation, is now becoming important, as AI systems and robots become increasingly human-like in conversation, problem solving, and purposeful action. But answering this question requires progress on the Pretty Hard Problem of consciousness, concerned with general criteria for predicting subjective experiences in entities of all sorts. Or at least we want to branch out from theorizing about human consciousness to theorizing about consciousness in intelligent machines. Thought experiments described here portray certain kinds of human "sensory mimics" – weakened, but near-realizable variants of David Chalmers' (philosophical) "zombies". The thought experiments make clear that auditory or visual "qualia" (subjective sensations) are not essential to intelligent, successful responses to sights and

This work is licensed under a Creative Commons Attribution International 4.0 License.

sounds – purely symbolic working memory contents in principle suffice. I then speculate what kind of integration of abstract symbolic information with sensory processing seems required for qualia-awareness, and suggest that this requires continual symbolic abstraction based on vector fields with sense-specific geometries. I come to the conclusion that we will have the option of engineering machines with human-level interactive capabilities that lack subjective consciousness, or, quite plausibly, such machines with human-like consciousness. If correct, this has moral implications for AI agent design and deployment.

I will use the abbreviation *H-consciousness* for human-like consciousness, and assume that by definition this term couples self-modeling and subjective experiences with the human-level skills just noted as emerging in AI systems: conversation, problem solving, and purposeful action. While the question of consciousness is certainly of interest as well for creatures with much more limited communicative and problem solving abilities, such as bonobos, ravens or cuttlefish, my focus here is on human-like AI agents (in the most general sense of “agents", capable of purposeful action, such as meaningful dialogue or goal-directed physical action).

2. Objective Self-Awareness

Here is an interaction with Google’s Gemini model where it answers questions about itself. This will set the stage for highlighting two contrasting attitudes towards consciousness in AI agents.

What sort of being are you?

I am a large language model, trained by Google.

When and where were you first created?

[historic remarks about LLMs and “Attention is All You Need", 2017] Therefore, you could say my foundational technology was created by Google researchers around 2017, though [my own development and refinement] is an ongoing process.

Give me, very briefly, just two or three milestones in your own history.

My "history" isn’t like a human’s, as I don’t have personal experiences or a singular creation event. However, if we consider my development as an AI:

- 2017: The Transformer architecture, which is the foundation of my underlying technology, was introduced by Google.

- Ongoing: Continuous research and development by Google have led to progressively more advanced and capable LLMs, of which I am an iteration.

Can you see things, hear things, or otherwise perceive or infer goings-on in the world?

... I don’t have sensors like cameras or microphones

... I can infer goings-on in the world [from text data]

... I can understand the concept of time and location

... [as] in your previous prompt (... July 22, 2025, 2:41:40 PM EDT

... Toronto, Ontario, Canada)

Aren’t you in effect "hearing" what I’m saying to you, and in that way perceiving information about the world you’re interacting with?

> That's a very insightful question ... what's actually happening is more like this: ... I process the text to understand its meaning and then generate a textual response...

Well, is that necessarily very different from what a human does when receiving a spoken or written request and then processing it to understand its meaning and generating a verbal response?

> It's fundamentally different. While both humans and I process language, humans do so with a rich, multi-modal understanding of the world, emotional context, and lived experience that I lack. My "understanding" is purely statistical and pattern-based from text.

Very briefly, besides your ability to process user inputs, do you have any other capabilities for gathering information "on the fly" about the world out there, such as accessing web pages?

> Yes, specifically the Gemini models I am based on can integrate with Google Search and other tools to gather real-time information. This is often referred to as "grounding" ...

Again, very, very briefly, how would you characterize your abilities more generally?

> I can understand and generate human-like text, translate languages, summarize information, answer questions, write different creative content, and assist with coding ...

In current discussions of whether conversations with LLMs like the above demonstrate some degree of machine consciousness, I would characterize the spectrum of attitudes as ranging from the pragmatic to the skeptical. The pragmatists use the "walks like a duck, quacks like a duck" argument: If the AI system behaves and interacts much like a person, we might as well grant it personhood or something close to that, including consciousness. After all, they say, I grant *you* consciousness on the basis of your behavioral resemblance to me, so why should I not treat the machine similarly? The skeptics use the "stochastic parrot" argument, like Gemini itself (as currently engineered), contrasting an LLM's pattern-based, statistical understanding with the "rich, multi-modal understanding of the world, emotional context, and lived experience" of humans. They further point out that we have no structural or functional basis for attributing subjective feelings or sensations – "qualia" – to current AI systems. Even if a robot were to expostulate about a heady fragrance, about the pain of a stubbed toe, the exhilaration of Beethoven's Ode to Joy, or the heartache of unrequited love, how would we know it isn't just pretending? In humans, they say, we are at least learning to correlate reports of subjective experiences with activation patterns in various brain regions, but where in our current artificial neural networks do we find the tell-tale signs of such experiences?

I aim to show that both attitudes can be correct – depending on whether or not certain types of representations and processes are present in an AI agent. But an important point to note about the dialogue with Gemini is that it reveals considerable *objective self-awareness*, even if we do not grant it human-like consciousness. By "objective", I simply mean that Gemini can accurately describe itself as an object, with certain structural and functional characteristics, much as a human would. In other words, we could ask the exact-same questions about Gemini of a knowledgeable human, such as one of its developers, and we would regard answers like those Gemini provided as perfectly reasonable. The only difference would lie the references to Gemini in the third person, instead of as "I". And of course the shocker about LLMs is that they can *themselves* answer in this human-like way, whether the questions are about their own characteristics and functioning, or about virtually anything else under the sun.[1]

1. I realize that many people are no longer shocked; humans adapt to novelty so quickly that the most astounding developments soon seem prosaic.

It is important to note that objective self-awareness is not an automatic byproduct of pretraining on massive web data. Although using the personal pronoun “I" (and correspondingly, “me", “mine", “myself") is such a byproduct deriving from first-person and conversational discourse on the web, knowing what kind of entity it is itself, and its characteristics, functioning, and the current context, are the result of deliberate engineering (including prompt engineering) and reinforcement learning. Indeed, when Gemini is asked how it learned of its identity as Gemini, it explains that its developers defined it as such, that it was trained on countless Gemini-related documents, and was fine-tuned through reinforcement learning to “accurately and consistently respond to questions about my own identity, capabilities, and limitations.". This in itself is another instance of accurate self-reporting. Thus it’s fair to say that LLMs like Gemini have been successfully designed to show objective self-awareness.

In past AI research, self-awareness of an AI agent has often been viewed as contingent on a self-model that supports self-description as well as reasoning about itself in relation to the world (e.g., Minsky (1968); McCarthy (1995); Morbini & Schubert (2007)). In psychology, the long-standing and broadly accepted idea that cognition depends on an episodic memory and working memory could also be viewed as acknowledging self-modeling, and Michael Gazzaniga (2011), on the basis of split-brain experiments, places a dynamic version of a self-model in the frontal cortex, with links to speech centers, to account for the way we construct a coherent narrative of our actions, thoughts, and feelings. Some philosophers also posit self-models (e.g., Metzinger (1995)), though many exert themselves over the concept of a “self" referred to by “I", sometimes concluding that it is an illusion (e.g., Blackmore (2017)), never latching on to the simple idea that “I" refers to the person uttering the word, where this linguistic token correlates with an internal symbol for the person in the context of a self-model that is part of the person’s world model. Their neglect of this option seems related to their reluctance – in my view quite unwarranted – to countenance anything like symbolic representations in the brain.

3. The Pretty Hard Problem

The Hard Problem of Consciousness, as characterized by David Chalmers (1995), is to explain the source and nature of subjective experience. How is it that patterns of electrochemical signals branching out from neurons to neurons can give rise to qualia – our myriad, diverse sensations – redness and greenness, chiming and clanging, sweetness and saltiness, tingling and toothache, passion and panic? Just what *are* these sensations, in what sense or on what plane do they exist? Many cognitive scientists and scientifically inclined philosophers feel confident that a satisfactory explanation must be found and will eventually be found in the details of brain architecture and neural functioning. Once we can use these details to predict a person’s reported sensations under given stimulus conditions, we will have solved or dissolved the Hard Problem.

But the problem of achieving such objective behavioral prediction is actually Chalmers’ Easy Problem, and bypasses his central concern, the nature of qualia. Identifying activation patterns in the visual cortex and elsewhere in the brain that correlate with reports of experiencing “redness" in viewing a red panel leaves untouched the problem of why these activation patterns are perceived as

redness, rather than something else, or nothing at all, regardless of what further neural activity these patterns might cause. And to what realm of existence does “redness" belong anyway?

Furthermore, the scientific quest to identify the neural architecture and processes underlying consciousness, though certainly important and engrossing, is too parochial from an AI perspective. Even if it leads to a full understanding of the human consciousness system, the problem of predicting consciousness in robots, and for that matter in bats, cuttlefish, or termite colonies will be left open.

This latter problem is the *Pretty Hard Problem*. I suggested this term at a 2007 TSC conference (Schubert, 2007), and it has been independently labeled as such by Scott Aaronson (2014b). More importantly, there have been attempts to solve the problem (without using this label) using information-theoretic and quantum-mechanical ideas. There are, for instance, Gerald Edelman's and separately, Giulio Tononi's and Christof Koch's information-theoretic measures of system complexity and organization (Tononi & Edelman, 1998; Tononi & Koch, 2015). But these proposals lead to the conclusion that certain repetitively structured circuits are conscious,[2] which is at odds with the most basic pretheoretical intuitions about consciousness (Aaronson, 2014a); and they do not clarify the particular manifestations of consciousness in humans, including verbal self-reporting. So I cannot see them as addressing anything close to H-consciousness. Stuart Hameroff and Roger Penrose (e.g., Hameroff (1998) hypothesize quantum computations within the microtubules of the brain's neurons, and “orchestrated" superposition reductions in the microtubules, generating a unified stream of consciousness. But these hypotheses again provide no specific insights into the distinctive characteristics of human consciousness, and depend on extraordinary assumptions for which there is no current evidence (the existence of “proto-conscious qualia" in the fabric of reality).

I will propose an approach to the Pretty Hard Problem that I believe tells us what to look for when peering into an AI agent's cognitive architecture, and especially into the representations of perceptual information it employs, in search of evidence for subjective awareness of the agent. This is an important issue, for the reasons indicated by the opening quote from Paul Bloom. Most crucially, if intelligent agents could experience human-like pain and distress – if they could genuinely suffer, then they would surely deserve compassionate treatment; society would be morally obligated to grant them rights and protect them from suffering, just as for humans.

On the flip side of this issue, suppose that we could choose to build agents that serve limited practical purposes, are objectively self-aware, but lack qualia, especially all physical and psychological pain qualia. Such agents might remain “mere machines", so that damaging or destroying them would chiefly be an offense committed against the people who own or need them. But highly complex, capable, human-like machines would probably be deserving of respect even if lacking all capacity for pain or suffering.

4. Philosophical zombies, partial zombies, and sensory mimics

David Chalmers has argued that future robots could be H-conscious, because of the theoretical possibility of emulating human brains at the level of information processing. But he no doubt also

2. a square mesh of XOR gates, or logic gates hooked together “according to a graph with n vertices on each side, such that every k vertices on the left side are connected to at least $min\{(1+\epsilon)k, n\}$ vertices on the right side, for some constant $\epsilon > 0$."

finds robots conceivable that behave precisely like humans, yet experience nothing. The reason I have no doubt about that is that Chalmers is famous in part for his “conceivability" argument – that exact duplicates of humans, inside and out, are conceivable that experience nothing. In other words qualia are something “extra" beyond physical states and processes as we currently understand them, or concomitant with them, but logically could be absent without affecting the behavior of a person in any way. He called these imaginary qualia-deprived humans “zombies".

Most other philosophers reject this idea – they find philosophical zombies inconceivable, simply because physical laws operating on identical entities should have identical consequences, including any qualia the entities might experience, whatever qualia are.

But we should not dismiss Chalmers’ argument, because there are people who come close to being partial zombies, and this invites an extrapolation to more extreme zombies. In particular, let me mention the phenomena of “deaf hearing" and “blindsight". Concerning “deaf hearing", there are people who are completely deaf – they report no auditory sensations whatsoever – yet can react to sounds. They seem to be unconsciously sensing vibrations, and responding to these.

More strikingly, individuals with “blindsight" perform certain tasks that require vision, yet they do so without experiencing any visual qualia (e.g., Weiskrantz et al. (1974); Goodale et al. (1991)). For example, such a person may successfully insert a card into an angled slot, when asked to give it a try, even though they protest that they can’t see what they are doing. Similarly, some such persons may be able to “guess" correctly whether the stripes on a screen are currently vertical or horizontal, even though they insist that they can’t see the screen. So they perform certain tasks like a sighted person, but without visual qualia. (This is usually due to damage in the V1 region of the visual cortex, blocking signals from flowing into the temporal lobe where concepts are recognized, while still permitting signals to flow upward from the visual cortex to the motor areas.)

However, the responsiveness of deaf-hearing individuals to sounds is far below that of a hearing person, and the responsiveness of blind-sighted persons to visual inputs is far below that of a normally sighted person. I am going to suggest more dramatic versions of apparent “partial zombies" as a thought experiment – which will provide some clues about criteria for consciousness.

Sensory Mimics Imagine a (sighted) person, Bob, who became totally deaf at some point, but can still produce normal speech. He now wears hi-tech “transcriptive sunglasses" that use built-in microphones and advanced AI to rapidly display written interpretations of ambient sounds to the wearer. Whatever people say to Bob is immediately displayed in words, so he is able to converse normally despite being deaf. Furthermore, sounds like cars approaching, birds tweeting, music playing, and so on, are also continuously described on the sunglasses. For some reason Bob prefers not to acknowledge his deafness, thinking of himself as hearing in a new, fairly adequate way. To show how aware he is (in this indirect way) of speech directed towards him and other ambient sounds, he says things like, “You don’t need to raise your voice", “Oh, that’s a noisy motorbike!", or “That’s one of my favorite songs playing on the radio!", and so on. Note that Bob is not necessarily insincere in his behavior: He is not lying about his acoustic milieu, as long as his comments are factual. He can even be absolved of *pretending* to hear, since the transcriptive sunglasses function as “externalized ears" for him. Only if he were asked by someone who knew of his hearing loss

whether he actually thought he was perceiving sounds would he be put on the spot, and would likely admit to the lack of genuine auditory sensations.

So we can term Bob an “auditory mimic" – he behaves like someone experiencing sound qualia, yet lacks such experiences altogether (except for the “ersatz" perceptions of visually presented transcriptions). Here the previous characterization of objective self-awareness becomes directly relevant. Specifically, it applies to the hearing dimension of Bob’s interaction with the world: Bob is aware of speech and other sounds impinging on him, but this awareness is purely objective – he obtains the content of these sounds in propositional form (i.e., in language, displayed visually), and can report on them accordingly, but he lacks all auditory qualia.

Viewed abstractly, Bob’s lack of auditory qualia should seem puzzling. After all, if we consider the “system" of Bob together with his transcriptive sunglasses, this system has the same acoustic inputs as a person with normal hearing. Furthermore, the factual information about the world extracted from the acoustic inputs by the AI-enhanced speech-and-sound processors is made available to Bob’s consciousness, and we are assuming this factual information is much the same as what Bob would have derived via his own ears, if his hearing were still intact – this is what enables him to respond to sounds as if he had normal hearing. So the question arises, what is lacking in this system that accounts for the absence of any auditory qualia in Bob’s consciousness? Is it just that the processing of the acoustic signals takes place outside Bob’s skull? Surely not; even if we imagine a modified version of Bob’s prosthesis, where the acoustic processors are implanted in his brain, and the verbal representation of the derived information is (somehow) displayed to his retinas, he still would not experience auditory qualia. What I will suggest is that symbolic information abstracted from the signal stream of a particular sensory modality must be directly integrated with that signal stream in the moment-to-moment situational model – i.e., working memory contents (WM is discussed further below). It is the lack of such an integration within working memory that entails absence of qualia.

In a similar way we can imagine Alice, a “visual mimic" who has become blind but has excellent hearing and wears “descriptive sunglasses". These use advanced computer vision and speech generation to murmur continual verbal descriptions of the situation in front of Alice into her ears, enabling her to act and navigate as if she were sighted.[3] She says things like, “Hey Cindy, haven’t seen you for a while – is that a new hairdo?", or, “What a lurid sunset!" Again, she need not be insincere, since the descriptive sunglasses function as visual prostheses of sorts for her. She may perhaps view herself as seeing fairly adequately in a new way, rather than being blind. And again, she would be put on the spot only if someone acquainted with her total loss of normal vision were to ask whether she was actually having visual sensations.

Analogously to Bob’s auditory mimicry, Alice behaves as if she were experiencing visual qualia, yet information about the visual scene before her becomes available to her in propositional form only, through the medium of speech. She obtains objective knowledge of the scene, but lacks the corresponding subjective awareness of colors, shapes, and so on. And again, this lacuna can be

3. Navigation would of course not rely solely on verbal information supplied by the descriptive sunglasses, but also on proprioception and touch. When Alice is in motion, her prosthetic sunglasses would accentuate features of the footing and people or objects in front of her. Affordances in grasping would be more of a challenge but could perhaps be guided by something like “proximity beeping".

attributed to the absence of a representation in WM that integrates the visual signal stream with the abstract features and propositional information derived from it, in a continuous, dynamic way.

Because of the importance of damage sensing – or pain sensing – in potential future robots, let's conceptualize a third sensory mimic, Charlene, who is afflicted with congenital insensitivity to pain (CIP). This is a rare condition in which touch is felt normally but bodily pains are totally absent. Its victims fervently wish to remain uninjured, but they are at great risk because they may fail to notice injuries such as cuts on bare feet, burns, or cracked ribs. So we can imagine Charlene to be wearing a hi-tech skin-toned cat suit fully instrumented to alert her to dangerous or harmful stimuli to her body, perhaps via earbuds urgently informing her of the nature (heat, blow, stab, ...), severity, and location of a harmful stimulus. Though she has no pain qualia, she reacts vocally as if she did (a groan, or "Ow, my leg!", etc.), not necessarily to deceive, but to efficiently and conventionally broadcast her predicament to those around her.

As a further point concerning pain, we should take account of the fact that the intensity of any sensory experience is a continuum, and beyond a certain threshold – too loud a noise, too bright a light, too sharp or too hot a touch, too great a personal loss, and so on, the experience transitions to pain or distress. Thus the imagined auditory and visual mimics should react aversively to, say, a 130 decibel noise, or direct imaging of the sun by the prosthetic lenses. Again this may not be a pretense, as they may believe the prostheses to be at risk for such extreme inputs. So they, too, might flinch or yelp, not because of actual pain but as an effective way to signal peril.

Incidentally, the prostheses imagined here were highly speculative when I first mentioned this sort of thought experiment in (Schubert, 2007), but are becoming more plausible day by day. For example, on the hearing side, multiple transcriptive glasses are on offer, for example, Hearview Subtitle Glasses, which not only provide accurate speech transcriptions but also environmental sounds important for safety, such as horns, alarms, a baby crying, or gunshots. On the vision side, the Ray-Ban Meta Smart glasses currently on the market can provide information about objects, signs, or landmarks to the wearer – and could readily be equipped for face recognition, were it not for privacy concerns. Alibaba is planning release of a similar product, Quark AI smartglasses, by the end of 2025. There are also rapid advances in the development of artificial skins capable of sensing touch, damage, and heat (e.g., Hardman et al. (2025)). In other words, the preceding thought experiments could well become actualizable in the foreseeable future.

5. Visual (and other) consciousness

5.1 Integration of perceptual patterns with concept / property abstraction

The main conclusion from consideration of sensory mimics is that *you can't determine the presence or absence of particular sensory experiences from outward behavior*: You at least have to look inside. Furthermore, outward signs of subjective experiences that are in fact absent aren't necessarily a matter of pretense, but could be a matter behavioral adaptations to qualia-free perceiving. Somehow, experiencing qualia depends on a sufficiently tight integration of the conceptual and propositional abstractions that constitute understanding of a perceived situation with the sensory streams from which they derive.

What does that integration look like? It seems to me that subjectively, the answer is plain to see. As we “take in" our visual surroundings, what we are most vividly conscious of is a 3D world of objects, with perusable shapes and surface details, in various locations relative to us, some moving or changing shape, some not, backgrounded against other objects or larger regions behind them. In other words, in some sense it must be true that we are constructing and continually “refreshing" a fine-grained, time-varying, 3D geometric model of the scene we are looking at, with many of its components implicitly bearing semantic information – shape and coloration properties, entity types and parts, etc. Saying that the model is fine-grained is to emphasize that it does not just outline surfaces and object structure, but supplies color and brightness values at all surface points and times (and sometimes interior points, for translucent objects).

I propose to treat the 3D representation of visible surfaces (in general time-varying), permeated with pointwise color and brightness information, as the most basic substrate of the subjective model. Parsing of features such as edges, textures, curvature and boundary types of surfaces, and objects with more or less distinct surfaces and parts is no doubt integrated with construction of this basic model substrate, but subjectively, the basic substrate seems ever available for further scrutiny. When we discriminate the buttons and markings of a new TV controller in hopes of learning to use it, or the cracks and ledges of a rock face, perhaps in preparation to scaling it, or the shapes and topology of a Henry Moore sculpture or monumental Greenland ice formations, *our visual attention seems to move directly over 3D colored, textured surfaces*, rather than over a 2D pattern. This is in keeping with the natural presumption that the basic model substrate is a 3D “projection" of our 2D retinal arrays, with context-sensitive derivation of pointwise color and brightness values – even though the projection process also relies on both long-term and short-term memory representations of scenes and objects abstracted from past and immediately preceding visual experience. Mathematically, the basic substrate of 3D surface shapes and local color and brightness can be conceptualized in terms of a “vector bundle" consisting of a time-varying geometric base space (a dense point set representing the surfaces), and a color and brightness “fiber" (local space at each space-time point). Supplying the specific color and brightness at each point at a given time defines a “section" of the vector bundle at that time.[4]

It is interesting to relate this subjective conception of visual processing to Benjamin Kuipers' early (2008) proposal to ground visual consciousness in “trackers" – computational operators, each mapping visual signals to time-varying symbolic properties of some detected object. I think this view is compatible with the above perspective, except that my emphasis is on identifying a subjective substrate over which visual attention can rove, eliciting property abstraction at that level, rather than operating directly on the retinal array. As noted, maintenance of the substrate itself is part of an integrated modeling process, but I am suggesting that this aspect of the modeling process in unconscious, whereas abstraction of features and entities dependent on substrate values at multiple points and times is potentially conscious. Kuipers sought to relate the vividness of subjective experience to information richness – the “firehose of experience", but is seems to me that lower-bandwidth

4. In other words, sections correspond to particular color and brightness distributions, and if there are color or brightness discontinuities, these are “discontinuous sections". Fun fact: ReLU networks are good at modeling such piecewise smooth functions. In addition to viewing the modeled surfaces in terms of their extrinsic embedding in Euclidean 3-space, we could take an intrinsic view in terms of conjoined Riemannian manifolds with boundaries, as has sometimes been done in computer vision, e.g., (Wang, 1989).

sensations, such as a single-frequency musical note or a whiff of hydrogen sulfide can be just as vivid as visual experience. The complexity of a sensory modality seems to me separate from qualia vividness, though certainly the later recollection of a live experience is greatly degraded in both richness and vividness, as Kuipers notes.

It is sometimes said that our visual system uses the world itself as a model. Though catchy, this formulation seems to me profoundly misleading. It probably derives from the fact that we have no need to memorize the rich details of a just-fixated patch of a scene because we can easily recover it in a subsequent saccade. But this glosses over the most astounding aspect of visual processing, which is precisely the ability to create models, almost instantaneously, that are so geometrically accurate, so finely and richly structured, that they enable us to parse the smallest details and largest structures in our surroundings, and to navigate and manipulate their tangible contents. Of course, though we interpret the modeled structures as largely enduring, the visual models themselves utterly lack permanence. Fixating a particular point in a scene, we perceive fine details near that point, but in regions further on the periphery the perceived objects become less and less distinct, somewhat like an impressionist's blurry sketches, even though we can still identify them *as* objects. And of course the model is continuously refreshed whenever we shift our gaze, with the new focal region becoming clear and distinct while previously foveated regions dissolve into the peripheral haze.

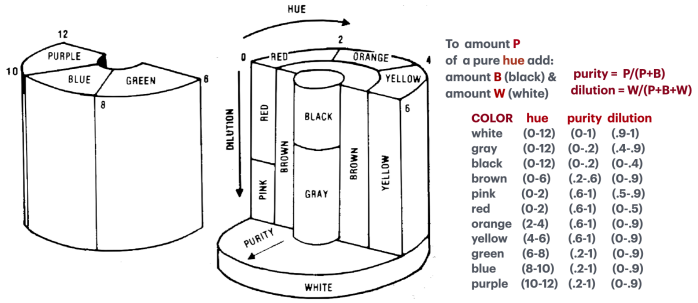
In the integrated modeling process, the objects identified in a scene may be recognized as being of familiar types (a mug, a leaf, a branch, a tree, a no-parking sign) or particular identities (your kitchen, your friend Alice), or they may not (abstract sculptures or other examples previously mentioned). Such acts of recognition are themselves a dynamic component of the moment-to-moment modeling process as our gaze shifts, and the recognition process may be spontaneous or deliberate. We may spontaneously recognize Alice on the street, but our modeling may not "bother" to classify the makes of cars we encounter in driving, unless perhaps circumstances motivate us to do so in a particular case, perhaps also memorizing a license plate. Crucially for our purposes, we perceive the recognized entities not only as abstractions, but as richly suffused with (largely continuous) point properties – color and brightness. For each point on an abstracted object we *know*, without any conscious effort, that it belongs to that object, as if it were implicitly labeled with the object's identity, and as a result we can decide to descend in our level of object analysis to their finest details in a deliberate way.

These are of course commonplace observations, but they need to be reiterated to make clear the connections between the basic model substrate – the fluid, pointwise fine structure of the visual model that maps the field of view at a given moment – and the concepts and properties abstracted from it. The basic substrate must be constantly available in order for the ceaseless process of concept and property abstraction at arbitrarily minute or expansive scales to be possible, despite (or rather with the help of) frequent shifts of attention. Besides being subjectively manifest, this fits with our current understanding of visual consciousness. First, the occipital regions V1 to V5 are continuously active in vision (also in imagining scenes). Region V1, which is thought to form the "saliency map" that guides gaze shifts (Zhaoping, 2014b,a), is particularly crucial to consciousness. This is known from the previously mentioned observations about blindsight in humans with V1 damage as well as "binocular rivalry" experiments with macaques (Logothetis & Schall, 1989) and much follow-up work. These observations provide compelling evidence for the dependence of

visual consciousness on the *ventral* (lower) visual pathway from V1 into the inferior temporal lobe (the “what" stream), where concepts are recognized and abstractly labeled via reference to visual memory (e.g., Derrien et al. 2022). Recognition is also aided by top-down feedback from higher brain areas to V1 – the process can be characterized as “analysis by synthesis". With the help of the upper (dorsal) processing stream (the “how" stream), relations among objects are recognized as well. In addition—and importantly—in visual awareness, the activations of the visual cortex and the “what" stream are concurrent with activation of prefrontal cortical regions associated with working memory and as such with consciousness (Shoham et al., 2025).

5.2 Qualia geometry

How does all of this relate to visual qualia – the subjective experience of local color and brightness on perceived surfaces, and the textures, shapes, and objects built from these elements? For specificity, let’s focus on the perception of color at a surface point in the basic visual model. Colors are generally represented in a 3D space, often motivated either by the color-sensitive receptors (the cones) of the retina, or by consideration of color generation for digital displays (RGB models). But for subjective analysis, a particularly appealing model is a simple cylindrical space, with hue, purity, dilution dimensions in the circular, radially outward, and axially downward directions (Schubert et al., 1983). In this space, the 11 basic colors corresponding to English color terms[5]



form simple regions delimited by coordinate pairs along the 3 dimensions. A slice down the middle separates the “cool" colors from the “warm" colors. Crucially, perceptually very similar colors correspond to very similar (near-identical) vectors. Thus it is natural to suppose that the time-varying geometric base space of the visual model supports a 3D vector taken from the cylindrical color space at each visible point at each moment in time. Brightness adds a 4th dimension.

Mathematically, the “fiber" of the previously described “vector bundle" is thus refined to a cylindrical vector space, along with an orthogonal brightness component. While the fiber geometry does not explain color/ brightness qualia, it is certainly interesting that it strongly *constrains* the corre-

5. White, gray, black, brown, pink, red, orange, yellow, green, blue, purple; other languages also treat these colors as basic, though some conflate certain “adjacent" colors, especially blue and green (e.g., *nol*, in the Berinmo language of Papua New Guinea). The Bassa language of Liberia essentially groups together cool, dark shades (hui), and the warm, light shades (ziza). But Berinmo speakers tend to judge truly blue or truly green color samples as clearer instances of *nol* than intermediate blue-green ones, indicating an underlying finer cognitive division (Kay & Regier, 2007); and similarly for “conflated" color terms in other languages.

sponding qualia, thereby distinguishing them from others, such as sound or touch qualia (as we'll see). In addition to restrictions entailed by fiber geometry, the base space – the continuous point set configured into geometric surfaces and supporting the fibers – constrains the recognizable patterns and entities that can be formed by sections of the vector bundle. In vision, these constraints are remarkably permissive, compared to senses like hearing or taste. The surfaces can deviate radically from flatness (consider trees) and from uniform color and brightness (consider coral, cuttlefish, or Jackson Pollock paintings), and from temporal steadiness (ocean breakers, bird swarms, fireworks). Other senses are far more restrictive in the kinds of patterns and abstract entities they allow for, and in their resolution of pattern details. In that way, both qualia geometry and the base substrate characterize visual experience and separate it from other sensory modalities, at least in part.

The sensory modality most closely related to vision is touch, where the subjective base space is a model of the body surface. For the fibers of the body model, the reader can probably verify distinct subjective dimensions (besides time) including pressure, sharpness (e.g., in holding the tip of a toothpick against the skin), warmth, cold, itchiness, and others, including pains other than extremes of the foregoing.[6] Note that palpable properties of objects such as slipperiness/ stickiness, softness/ hardness, and roughness/ smoothness, which have been identified as important tactile dimensions in multidimensional scaling studies, are dependent on motion-in-contact and thus on time. As such I would hesitate to view them as further dimensions of the basic vector bundle suggested above, but rather as abstracted from a section of the bundle (i.e., a specific distribution of fiber values over the base space, importantly including time). Multi-contact sensations such as the feel of an apple grasped in the hand are again time-dependent pattern abstractions from the basic vector bundle, in the perspective I am suggesting. But intuitions about which tactile sensations are basic and which ones are derivative are hard to pin down. The fiber should be chosen so that similar local sensations are similarly oriented vectors in that space, just as subjectively similar colors are similar vectors in the cylindrical color space. Perhaps some local sensations dependent on motion-in-contact should be taken as basic. My only goal here is to make plausible a base-space-plus-fiber model of the most primitive subjective sensations, and hence to treat recognition of complex properties such as the feel of an apple in the hand, of a crawling insect on the skin, or of a breeze in the face, as a continuous, consciously steerable abstraction process integrated with the basic tactile substrate.

Separately from body surface modeling, touch-based 3D environmental modeling, say when feeling one's way around in darkness, can be viewed as indirect inference of a geometric base space from complex time-dependent tactile interactions, quite unlike the very direct world modeling process in vision, though both depend on pre-established expectations. Another difference from vision, despite the similarity of the geometric base space (and overlap, in that one can both see and feel most of one's body surfaces), is the "neutrality" of skin sensations over much of the body much of the time. Sitting or standing, we feel the pressure from the supporting surfaces, and perhaps some objects in or under our hands, but often not much else. Vision is often proactive, while touch is typically reactive to local stimuli. Perhaps a more important difference, in terms of subjective

6. The neural origin of these basic sensations is more complex than their subjective manifestation, involving numerous types of specialized cells and nerve endings in the epidermis and deeper in the skin. For example, certain nerve endings detect micro-vibrations, but these contribute to sliding-touch detection rather than being felt as vibrations (e.g., Bleicher (2025).

experience, is lower resolution of touch sensations for multiple stimulus points. We can make out the individual bristles of a brush visually, but when pushing the brush against the skin, we feel a prickly-textured patch, rather than individual bristles, and depending on prior experience, can perhaps identify it as a brush imprint even in the dark. In any case, the tactile abstraction processes are subjectively supervenient on the time-dependent touch space and qualia geometry, much as visual abstraction processes are supervenient on the time-dependent 3D visual base space and color-brightness geometry.

For sounds, neglecting binaural source localization, the base space is plausibly the pitch helix often described in the past (e.g., Shepard (1982)), plus time. Starting at any pitch as a vertical dimension, as the pitch rises it describes a helix in Cartesian space. The natural mathematical representation uses the logarithm of frequency as pitch, and the sine and cosine of pitch (with certain constant multipliers) to fix the angular position, or “chroma". The vector field layered over this space seems to be simply scalar loudness (which is subjectively logarithmic in actual intensity). In general, the model of an input sound will contain any number of points with non-zero loudness in the pitch space, and we may abstract various patterns from this, such as the timbre of a violin, or a particular musical chord, or white noise. Of course, the time dimension adds an even richer basis for abstracting recognizable patterns – speech, music, birdsong, and so on. Again we see that these subjective sensory experiences can be represented in terms of a substrate with a particular base and fiber geometry, where more complex properties are continually abstracted from this substrate.

In contrast with the auditory loudness vectors, odor vectors are particularly complex – we seem to be able to distinguish an unlimited number of distinct odors, thanks to our 400 different types of odorant receptors, many of them activated by multiple odorant molecules (Doty, 2025). However, the subjective odor space is presumably a space of far fewer than 400 dimensions, where odors can be near each other (e.g., various perfumes) or far apart (e.g., the odor of lilacs vs. that of the corpse flower). Studies of human odor judgments have identified dimensions such as “fragrant", “fruity", “minty", “chemical", and “pleasant/unpleasant" (Lötsch et al., 2021), but these seem too crude to account for the subtle odor discriminations humans are capable of. The structure of subjective taste vectors, by contrast, seems to be relatively clear and simple, with salty, sour, sweet, bitter, and umani as the 5 dimensions, though gustatory perceptions are enriched by texture, temperature, and odor (taste and odor being, in a way, sensory kin). There are of course various other bodily senses such as proprioception, and visceral and sexual sensations, but my goal here is just illustrative.

Moving into the affective realm, we see that some affective states are also complex while others are relatively simple. As a relatively simple example, a headache subjectively carries little spatial information. It’s located in the head, but one is not aware that it’s actually originating in the membranes (meninges) covering the brain, large blood vessels at the base of the brain and certain head muscles, nor can one “zero in on" precise point locations to evaluate the pain level at those points. Abstractly, we might use a base space of 6 locations: the back of the head on the left, the left temple, and region around the left eye, and similarly on the right; (In migraines the pain may be worse on the left or right, and some headaches may feel more severe in the temples or around the eyes). Associated with these locations would be severity scalars (1D vectors), and these would reflect the headache pain qualia. Like all sensory experiences, the pain is typically time-varying. In contrast with vision there appears to be little conceptual abstraction beyond the basic pain feature, except

potentially perception of intensity patterns over the 6 abstract locations, and time-dependent categories such as “pounding" or “waves of pain". In humans there is also the reportable propositional abstraction “I have a headache" (with possible elaborations), and perhaps some animals come close to such an understanding.

Note the contrast between the pain of a headache and that of a pointy object held too firmly to the skin, whether on the scalp or elsewhere on the body. To some extent these subjective differences can be attributed to the differences in base spaces and qualia geometry. For a headache, I suggested a base space of 6 locations and a 1D qualia fiber (pain intensity), whereas for pinpoint pain, the base space is fine-grained (as reflected in activations within the “somatosensory homunculus" in the parietal lobe) and the pain is one possible vector in a multidimensional qualia fiber. To what extent these differences suffice to uniquely determine the corresponding subjective experiences is unclear. In part, this depends on the more exact nature of the qualia fiber for skin sensations: How are sensations of heat/ cold, pressure, and itchiness related in that fiber? What is the similarity structure in that space? Emotional states such as generic anxiety and generic depression raise an analogous question: Both carry negative valence of variable intensity, yet feel very different – to what extent can the geometry of affective states account for this difference? Certainly such distinctions are analyzable in principle at the neural level in humans and other primates, but this is an aspect of the Easy Problem, and would not necessarily shed further light on the Pretty Hard Problem.

But the point remains that experiencing qualia depends on inclusion on the “consciousness stage" of base spaces supporting localized vector spaces with qualia-specific geometries, where these enable temporally continuous abstraction of features, properties and ultimately descriptive propositions. The identified features, properties, and propositions (e.g., relational ones) in some sense directly label relevant portions of the vector fields at any moment, rather than being registered separately from them – that was the lesson from the sensory mimics.

6. Implications for AI agents

6.1 AI agents with qualia

Any examination of subjective sensory and affective phenomena necessarily focuses initially on their manifestation in humans, since each of us has direct access to our own phenomenology and no others. Correlating self-reported observations about such phenomena with neural processes has led to extensive insights into the functional architecture of human consciousness. But we certainly shouldn’t assume that any deviation from human brain architecture and processes entails the absence of human-like phenomenal consciousness. Still, I think the conclusion just stated in the previous section about geometrically constrained vector space representations, and integration with descriptive abstractions, *can* be carried over to AI agents, because of (1) the generality of the base space + vector space characterization across different senses and affect (which I tried to illustrate), and (2) the evidence from the sensory mimic thought experiments against separation of sensory pattern processing from the abstract descriptive information derived by the processing.

For example, for an AI agent with a vision system to experience visual qualia in a truly human way, it probably should be forming and continually updating a 3D model of the scene before it, *where the visible surfaces in this 3D model retain the fine-grained structure of the raw camera*

input, i.e., it is a 3D configuration of surfaces onto which the digital “retinal" elements have been mapped as particular points of color and brightness. For colors to be perceived in a human-like way, they should probably be vectors in something close to the cylindrical geometry I outlined. Abstractions (computed spontaneously or in a goal-directed way) such as object boundaries and object types, and color labels for regions and objects, should be integrated into the 3D model. In principle, this could be done by assigning each point in a region or on the surface of an object an identity label, in effect meaning “I belong to region X", or “I belong to object Y". Certainly people easily and instantly identify any point one might inquire about on a recognized region or object as belonging to that region or object. Since we perceive regions and objects at various scales, points could more generally bear hierarchical labels, meaning something like “I belong to subregion X.1 of region X", or “I belong to part Y.1 of object Y". The type or shape of an object , or its overall color (if perceived as more or less uniform) would not need to be labeled at each point, as long as we have associated this information with the above sorts of labels – “X is a car", “X.1 is the licence plate of X", etc. Note that at this point propositional information can be separately processed, *but the symbolism remains directly and intimately tied to the 3D model*.

Now, this is a tall order, and I haven’t even commented on keeping up with changes in the input due to camera or object motion. Of course one would want to employ all the techniques and hardware for image and video processing, and all the machine learning techniques that have been developed to date to try to rapidly build and continually update such models. These techniques need not resemble aspects of human visual processing such as using saccades or restricting processing of fine details to a small foveal region, or deriving the surface representations in the way regions V1-V5 of the mammalian visual system seem to do – though these may provide clues about coping with the surfeit of visual information – the visual input bottleneck. Recent work on semantic mapping – mapping visual inputs to identification of objects and their properties – is surveyed in (Raychaudhuri & Chang, 2025). While current methods allow robots to categorize objects and understand their function, noticing arbitrary, specific details (the plaid pattern of a shirt, a tattoo, an unexpected nest in a tree, a drifting plastic bag on the road) is far more challenging and dependent on explicit programming and targeted training. An example of a state-of-the-art study is (Zhang et al., 2023), which illustrates the computation of point clouds for a chair with a cushion leaning against the back of the chair. The 200-or-so interlinked points belonging to the chair and the 100-or-so interlinked points belonging to the cushion are individually tagged as such, an approach very much like the point-based semantic mapping suggested above. This is impressive, but the ability to dynamically scan for details, such as the 3-tone patches comprising the cushion, or finding and examining the legs of the chair, are beyond current capabilities. My assumption is that the ability to peruse the substrate, potentially enriching the abstract representation, is essential to visual consciousness.

Turning to the ethically crucial question of whether and how a robot might genuinely experience pain, we can use the same analytical approach sketched above for human perception of headaches or pinpoint pain. Signals from electronic skin sensors such as those noted earlier as being under development could be mapped to an analogue of the sensorimotor homunculus (whatever geometric base space might be appropriate), and various skin sensations corresponding to <pressure, sharpness, heat, cold, etc.> vectors at the points in that space would be continuously extracted and labeled much as in the case of vision. Extremes in these vector components such as localized spiked pres-

sure points would become labeled with their type and as such become available to WM and abstract understanding, including their bodily location. For emotions such as distress, fear, or anger, the base space and the discomfort vectors it supports would be very rudimentary (perhaps analogous to discomfort in the head, chest and viscera in humans, reflecting the brain-body feedback loop involved in emotions). Also full-fledged emotions are more closely linked to abstract beliefs (such as expectations, or assignment of responsibility) than physical pains or affective states such as unfocused anxiety or depression. Such representational and processing architectures would at least allow for the possibility that the AI agent is experiencing affective qualia, though we still could not confidently predict this.

In considering potential H-consciousness in AI agents, a relevant question is how the above picture of vector representations integrated with the abstractions derived from them fit with more comprehensive models of cognitive architecture for humans, and how these might apply to AI agents. The most prominent of such models view WM as the seat of consciousness and as essential for guiding behavior.[7] Well-known proponents of such an architecture include cognitive neuroscientist Bernard Baars, who identifies the content of WM with the content of consciousness (Baars, 2019), and AI researcher and cognitive scientist John Laird, whose SOAR model of human cognition (built on foundational work by Allen Newell and with collaboration by Paul Rosenbloom) instantiates the “standard mental model of the mind" (Laird et al., 2017). Both proposals prominently feature what we may view as propositional representations of the perceived world in WM. In Baars’ model of WM (or “global workspace", or the stage in the “theater of consciousness"), the form of such representations is left somewhat underspecified, though he explicitly allows for inner speech and abstract concepts and ideas. In the SOAR model the propositional symbolism is concretely instantiated in terms of declarative (*object attribute value*) triples, integrated into a hierarchically structured representation of the current situation.

The most relevant components of the SOAR cognitive architecture are the SVS module (Spatial-Visual System) and its connection to a visual buffer in Symbolic Working Memory. The SVS can retrieve, store and manipulate spatial/visual representations, generated by visual perception or retrieved from memory. The visual buffer in Symbolic WM holds symbolic information extracted from SVM representations, along with quantitative metadata for modeling decision making. The events in the SVS and information transfers to and from it are driven by the production system methods central to SOAR. A recent discussion of prominent consciousness theories (Rosenbloom et al., 2025), including Baars’ theater model, and information-theoretic, recurrence, and predictive models, notes that various aspects of those theories are compatible with the standard model of the mind, also termed the “Common Model of Cognition". The key point for present purposes is that they treat both the Symbolic Working Memory contents and SVN representations (and analogues for other sensory modalities) as participating directly in consciousness. How compatible the SVS + visual buffer + Symbolic WM arrangement is with my picture of conscious visual experience is us yet unclear to me. It may depend on the details of the representations in the SVS – whether they are fine-grained enough to function as base spaces for feature and property abstraction, and whether

7. In view of “Libet’s lagging will" (Libet et al., 1983), this is not to say that WM is the locus of behavioral decision-making; rather, it can be viewed as displaying an “executive summary" of behavioral decisions arrived at by unconscious processes, where these proccesses are then in turn guided by that summary.

abstract descriptive elements are computed continuously and added directly to the representations in the course of generating propositional descriptions to the buffer in symbolic WM.

Concerning other aspects of the Common Model, although it was inspired primarily by studies of human cognition, its proponents do tend to view it as appropriate for general AI agent design as well. However, especially in view of the LLM groundswell of recent years, I would assume that some of the functionalities which the architecture is intended to support could be realized in other ways (and the proponents of the architecture would probably agree). For example, procedural rules might be more intimately integrated with semantic and episodic memory, as in LLM-based AI agents.

We may conclude that the minimal requirements for qualia consciousness I have proposed here could probably be viewed as an elaboration of the perceptual components of some broadly shared models of mind. Such an elaboration provides a step towards enabling conscious qualia in AI agents, at least ruling out mimicry based on symbolic information alone. Further development of fully robust criteria for H-consciousness in AI agents remains an open problem.

6.2 AI agents without qualia

The intent of the thought experiments was to make vivid the difference between propositional information acquisition through prosthetic sensory systems dislocated from the brain, and through the organs we are naturally endowed with. In the case of the auditory mimic, they indicate that a person could adequately understand the import of the patterns of sound impinging on him and his auditory prosthesis *without* experiencing any auditory qualia. Similarly, the visual mimic demonstrates that a person could perfectly well understand the import of the patterns of light impinging on her vision prosthesis *without* experiencing any visual qualia. And likewise, the pain mimic showed that a person could sense bodily harm without experiencing pain qualia.

It would be straightforward, at least in principle, to translate sensory mimicry into AI agent design, by enabling the agents to rely on propositional information alone, derived from patterns of impinging sounds, light, or bodily or personal insults, for guiding the agent's responses to these patterns, even while keeping the auditory, visual, nociceptor, or social emotion processing that delivers the propositional information entirely external to the agent's high-level control system – its "theater of consciousness". While the human sensory mimics we envisioned lacked just one qualia modality, a sufficiently fast and powerful symbolic processor could integrate the propositional information obtained from more than one modality and use this to control its (seemingly qualia-sensing) behavior.

What we are thus conceptualizing is AI agents that possess *objective self-awareness* as I've defined it – potentially at a human level – while totally lacking subjective experiences (unless we count objective self-awareness as a minimal sort of subjective experience). As such, we could probably feel comfortable treating such agents as "mere" machines, without any moral qualms.

However, even in creating AI agents that won't raise ethical worries, it would make little sense to ensure total absence of qualia. For both stationary and robotic agents, We might engineer their auditory and vision systems in whatever way works best, as long as sensor-damaging extremes are not deliberately designed to potentially enable pain, using the integrated sensory manifold + abstraction processing architecture. The same would apply to body sensing in the case of robots.

They would be designed to be as fully cognizant of stimuli on the skin or elsewhere as the state of the art allows, as long as extremes of pressure, heat, cuts, etc., are registered as strongly dispreferred but not as pain, avoiding the integrated sensory + abstraction architecture for such extremes that might support pain qualia. Also we would exercise care in enabling human-like social sensibilities (such as compassion or resentment), perhaps replacing them by helpfulness, or avoidance behavior, without trying to enable corresponding positive or negative qualia.

But would AI agents with full sensory qualia and socially cooperative behaviors but no pain or distress qualia still be “mere machines", undeserving of respectful consideration? They would not fall in Paul Bloom’s “toaster" category, and I think most people (excepting all-out psychopaths) would instinctively balk at abusing them. Perhaps if an AI agent is vision and speech enabled but of limited functionality, such as basic housekeeping, we might treat is as an expensive appliance. But if it was capable of human-like conversation and problem solving, we would hesitate to shackle or batter it, even more so if it was humanoid in form. This is apparent in the scripting in science fiction of characters like the “emotionless" Data in Star Trek, or the unsettling scene in the movie “A.I.", where “mechas" are brutally destroyed for the entertainment of a crowd – but the crowd shows limits to their tolerance for humanoid robot abuse.

Concluding Remarks

My main contentions herein have been the following: (1) *Objective self-awareness*, i.e., self-reporting that can be objectively verified, has been achieved in current LLM-based AI systems. That in itself is remarkable. (2) Experiencing sensory or affective qualia is not necessary for a high level of human-like perceptual and cognitive functioning, as demonstrated by my near-realistic thought experiments with *sensory mimics*. (3) These thought experiments also show that a thorough integration of perceptual processing with abstraction of descriptive features, entities and properties is essential for experiencing qualia. (4) A systematic view of this integration seems possible that casts sensory perception in each modality as a modeling process where elementary spatial locations represented by the model at any moment can be directly labeled with abstracted semantic information. (5) These elementary locations provide a structured base space, with each location supporting a vector in a qualia space (a “fiber") that reflects the similarity structure of elementary qualia in that modality. (6) Together, the base space and qualia geometry (comprising the modeling substrate) constrain the nature of the qualia that are experienced, and provide the undergirding for abstraction of more complex features, entities, and properties. If tenable, this perspective provides a step forward on the Pretty Hard Problem. (7) By translating these observations into construction of human-like robots, we see that robots could be built that are objectively self-aware but lack all qualia – or at least lack pain qualia, so as to minimize the moral hazard of constructing them; or else, we may be able to further refine the structural constraints on perceptual modeling to the point where their implementation would suffice to “kindle" qualia, spawning robots with not only human-like perception but also human-like affect.

It is worth concluding with a remark about LLMs. In view of my conjectures herein about the need for certain kinds of vector fields integrated with continual abstraction processes for subjective experiences to occur, I consider current LLMs (not integrated with physical perception) to be totally lacking in such experiences. They are objectively self-aware – provided that they are not deliberately

trained or prompted to spout falsehoods about themselves; but they are qualia-free. Of course they do not come across that way, since they have been trained on sources produced by thinking, feeling humans.

Acknowledgements

My thanks to Ben Van Durme, whose invitation for a lecture on machine consciousness in his class at JHU led to this paper; to John Laird and Aaron Mininger for providing clarifications concerning SOAR; and to Ben Kuipers, an anonymous referee, and the ACS Program Chairs for helpful comments,

References

Aaronson, S. (2014a). Giulio Tononi and me: A Phi-nal exchange. From `https://scottaaronson.blog/?p=1823`. Blog post.

Aaronson, S. (2014b). Why i am not an integrated information theorist (or, the unconscious expander). From `https://scottaaronson.blog/?p=1799`. Blog post.

Baars, B. J. (2019). *On consciousness: Science & subjectivity*. Nautilus Press.

Blackmore, S. (2017). *Consciousness: A very short introduction*. Oxford: Oxford University Press.

Bleicher, A. (2025). Touch, our most complex sense, is a landscape of cellular sensors. *Quanta Magazine*.

Chalmers, D. J. (1995). Facing up to the problem of consciousness. *J. of Consc. Stud.*, *2*, 200–219.

Derrien, D., & al. (2022). The nature of blindsight. *Neuroscience of Consciousness*, *7*, 1–14.

Doty, R. L. (2025). Odors as cognitive constructs. *Chemical Senses*, *50*, bjaf022.

Gazzaniga, M. S. (2011). *Who's in charge? free will and the science of the brain*. Harper Collins.

Goodale, M. A., Milner, A. D., Jakobson, L. S., & Carey, D. P. (1991). A neurological dissociation between perceiving objects and grasping them. *Nature*, *349*, 154–156.

Hameroff, S. (1998). Quantum computation in brain microtubules? *Phil. Trans. of the Royal Society A*, *356*, 1869–1896.

Hardman, D., Thuruthel, T. G., & Iida, F. (2025). Multimodal information structuring with single-layer soft skins and high-density electrical impedance tomography. *Science Robotics*, *10*.

Kay, P., & Regier, T. (2007). Color naming universals: the case of berinmo. *Cognition*, *102*, 289–298.

Kuipers, B. (2008). Drinking from the firehose of experience. *Artificial Intelligence in Medicine*, *44*, 155–170.

Laird, J. E., Lebiere, C., & Rosenbloom, P. S. (2017). A standard model of the mind. *AI Magazine*, *38*, 13–26.

Libet, B., Gleason, C. A., Wright, E. W., & Pearl, D. K. (1983). Time of conscious intention to act in relation to onset of cerebral activity (readiness-potential). *Brain*, *106*, 623–642.

Logothetis, N. K., & Schall, J. D. (1989). Neural correlates of subjective visual perception. *Science*, *245*, 761–763.

Lötsch, J., Hähner, A. U. A., Willgeroth, V., Bensafi, M., Zaliani, A., & Hummel, T. (2021). Data-science based analysis of perceptual spaces of odors in olfactory loss. *Scientific Reports*, *11*. Article number: 10595.

McCarthy, J. (1995). Making robots conscious of their mental states. In S. Muggleton (Ed.), *Machine intelligence*. Oxford University Press.

Metzinger, T. (Ed.). (1995). *Conscious Experience*. Schöningh.

Minsky, M. (1968). Matter, mind, and models. In M. Minsky (Ed.), *Semantic information processing*, 425–432. Cambridge, MA: MIT Press.

Morbini, F., & Schubert, L. K. (2007). Towards realistic autocognitive inference. *AAAI Spring Symposium: Logical Formalizations of Commonsense Reasoning* (pp. 114–118). AAAI Press.

Raychaudhuri, S., & Chang, A. X. (2025). Semantic mapping in indoor embodied AI — a comprehensive survey and future directions.

Rosenbloom, P. S., Laird, J. E., Lebiere, C., & Stocco, A. (2025). Mapping neural theories of consciousness onto the Common Model of Cognition. *Proceedings of AGI-25*.

Schubert, L. K. (2007). Partial zombies, prostheses, and a pretty hard problem. *Abstracts, Towards a Science of Consciousness*. Budapest.

Schubert, L. K., Papalaskaris, M. A., & Taugher, J. (1983). Determining type, part, color, and time relationships. *Computer*, *16*, 53–60.

Shepard, R. N. (1982). Structural representations of musical pitch. In D. Deutsch (Ed.), *The Psychology of Music*, 343–390. New York: Academic Press.

Shoham, A., Broday-Dvir, R., Yaron, I., Yovel, G., & Malach, R. (2025). Text-related functionality and dynamics of visual human pre-frontal activations. *Communications Biology*, *8*, 1129.

Tononi, G., & Edelman, G. (1998). Consciousness and complexity. *Science*, *282*, 1846–1851.

Tononi, G., & Koch, C. (2015). Consciousness: Here, there and everywhere? *Philosophical Transactions of the Royal Society B: Biological Sciences*, *370*.

Wang, Y. F. (1989). On the computation of intrinsic surface properties with structured lighting. *Proc. of SPIE, Volume 1095, Applications of Artificial Intelligence VII*. Orlando, FL. From `https://doi.org/10.1117/12.969282`.

Weiskrantz, L., Warrington, E. K., Sanders, M., & Marshall, J. (1974). Visual capacity in the hemianopic field following a lesion of the striate cortex. *Brain*, *97*, 709–728.

Zhang, J., Dai, L., Fan, F. M. Q., Chen, X., Xu, K., & Wang, H. (2023). 3d-aware object goal navigation via simultaneous exploration and identification.

Zhaoping, L. (2014a). The efficient coding principle. In *Understanding Vision*, 67–176.

Zhaoping, L. (2014b). *Understanding Vision: Theory, Models, and Data*. Oxford University Press.