9
What of the Hard Problem?

9.1 What a Conscious Experience Is Like

Some readers may feel that we have thus far left out the single most important issue: the qualitative aspects of subjective experience.

To many, this Hard Problem—of accounting for the subjective, phenomenological nature of conscious experience—is why we are here in this business in the first place. To some, the supposed deep conceptual mystery is a given. But at times, we also struggle to explain to the uninitiated what the problem really is, without resorting to some philosophical jargon.

It may be telling that one of the most useful phrases in these situations is "what it is like" to have a certain experience (Nagel 1974). We explain to our friends: there is *something it is like* to be in that sharp pain in the finger. That horrible feeling is much more than a piece of information telling you that something is wrong with your finger. It *feels like something* to be in that brain state. Accounting for that "what-it-is-like-ness" in objective scientific terms is the Hard Problem.

9.2 A Structural-Relational View

So what is it like to feel that sharp pain in the finger? Well, it feels somewhat like a dull pain, but is more pinpointed, more concentrated, right? It is not exactly like a dull pain, but it certainly is closer to that than to a gentle stroke. Another way to put it may be to say that the sharp pain at the tip of my index finger is very much like the same pain in my thumb. The locations are different, but the subjective *quality* of the pain is similar. Neither is it like the taste of baked potatoes at all. Nor is the pain anything like the sound of the cello. But if you force me to make an auditory analogy, it is probably somewhat more like the scream of a cat rather than the sound of waves in the ocean.

That is to say, in describing the qualitative character of a conscious experience, the best we can do is often to relate and compare it with other experiences we've had. In fact, perhaps this is not just a matter of communication.

In Consciousness We Trust. Hakwan Lau, Oxford University Press. © Hakwan Lau 2022.
DOI: 10.1093/oso/9780198856771.003.0010

When we think about what an experience is like, we can't help but think in these terms too. As soon as we imagine what the experience is like, comparing it with other experiences seems intuitive. Even when we don't explicitly make these comparisons, perhaps such comparisons are already made implicitly. In fact, it may be difficult to imagine having a conscious experience in ways that does not involve such implicit comparisons. When we see red, we see it as looking rather different from blue. Red looks the way it does because it looks somewhat closer to orange, to pink, to brown, than to blue. Red looks the way it does because it looks *redder* than everything else.

We can call this a structural, relational, or holistic view of perceptual experiences. The idea is that a conscious experience cannot really be defined in isolation. If a creature is only ever to see a single flash of light in its entire life and evolutionary history, there is perhaps just no fact of the matter whether the flash will look red or green. It will just look like an indistinct, nondescript flash—if it looks like anything at all. The qualitative experience of seeing specific colors only comes about because there are different colors that we can see and subjectively distinguish from each other.

This view has been well-articulated by philosophers. In recent years, authors like Austen Clark (2000) and David Rosenthal (2010) have developed detailed versions of this line of thinking. The central idea is that we can determine the qualitative character of an experience based on the position of the relevant stimulus on a *mental quality space*. Such a space is a theoretical posit, construed such that stimuli that are subjectively similar are placed close to each other on the space. In other words, each point on the space represents a stimulus, and the distance between two points reflects the discriminability between two stimuli.

Note that this pairwise discriminability here is defined functionally, as in psychophysics. So two stimuli are more discriminable from each other if the subject is more able to distinguish them behaviorally. As such, the quality space allows us to characterize the subjective quality of conscious experiences in functional terms, without circularity. The distinctive phenomenal quality in consciously perceiving a stimulus is determined by how that stimulus is functionally distinguishable from other stimuli for the subject (Figure 9.1).

In the chapter I will flesh out how this mental quality space view fits well with the theory of perceptual reality monitoring (PRM) introduced in Chapter 7. So taken together we have an empirically grounded theory of consciousness that can account for the phenomenal character of subjective experience too.

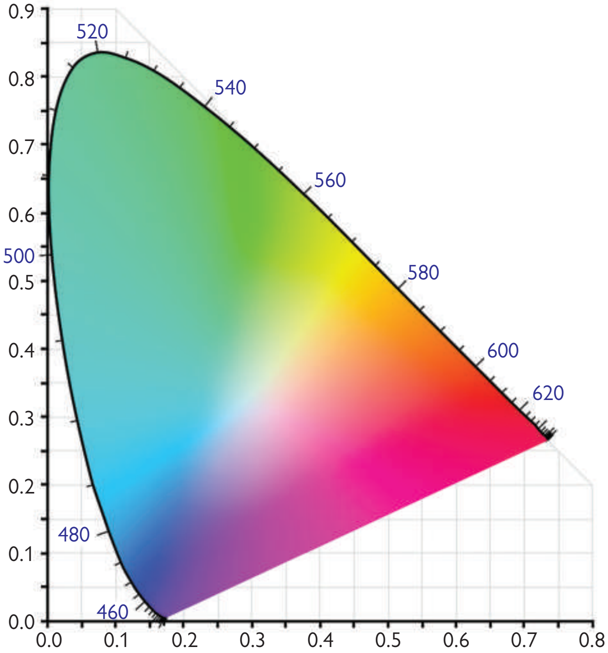

Figure 9.1 A hypothetical mental quality space for color perception for a particular individual; the distance between two points here concerns the individual's ability to distinguish between the two stimuli, rather than the objective physical similarity between the stimuli, so this space is different for different individuals. The subjective similarity between two colors can be explained by the degree of overlap between the relevant neuronal coalitions. Colors represented by more distinct populations are functionally easier to be distinguished.

9.3 "Knowing" the Quality Space

Given that we *know* what it is like to have a certain conscious experience, we presumably represent the relevant information in our brains somehow. But just because the qualitative or phenomenal character of subjective experience is best described relationally, it does not mean that we have to do it in terms of the positions on a quality space. As I have done above, we can also do it in propositional forms (i.e., something like sentences: seeing red is somewhat like orange, a little like pink, or brown, nothing like blue, nothing like silver, *and so on*).

But to *exactly* spell out the full proposition including all the relevant relations with *all the possible stimuli* would be way too clumsy, to the point that it is virtually impossible. Perhaps this explains why the quality of subjective experience seems so hard to articulate; philosophers sometimes say that it is just *ineffable*. Representing this subjective quality in terms of the location of a space is a handy alternative. It encapsulates a lot of information conveniently. But such spatio-geometrical information is only useful if we have some grasp of that space. It's no good to talk about a coordinate on a map if we don't even know what the map is.

How do we represent the mental quality space? This may seem rather challenging. The space allows for all the possible stimuli that one can perceive. Furthermore, one needs to know the discriminability of *any* two stimuli. Because the discriminability concerns how well oneself can perform the discrimination, that's a lot of self-knowledge involved.

Fortunately, there may be a plausible shortcut. The behavioral discriminability between two stimuli ultimately comes from the neural representations themselves. Suppose two colors activate very similar patterns of neural activity. Let's say for the neurons excited by these two colors, 98% of them fire at similar levels for both. By just knowing this fact, the brain should be able to infer that the two colors are not very discriminable. On the other hand, if there is a third stimulus that activates an entirely different neuronal population, the brain should "know" that it should be very discriminable from the two colors. The difficulty in behaviorally distinguishing two stimuli comes from the fact that the relevant sensory neural codes are similar. Therefore, if the brain "knows" what stimuli are represented by certain activity patterns, it should be able to tell how discriminable the stimuli are, by comparing the corresponding activity patterns.

Of course, here I put the word *know* in quotes, meaning that I use it in a subpersonal, implicit sense. The prefrontal cortex is not a person, and if it "knows" something, it does not mean that the person having that prefrontal cortex knows it explicitly. It may not be something that we can articulate easily without further reflections. But the information is there. It is implicitly known.

Recall that according to PRM, a specific higher-order mechanism (akin to a discriminator in generative adversarial networks (GANs) discussed in Chapter 6) determines whether some sensory activity should give rise to subjective experience. The mechanism *refers* to first-order sensory states, in a way that was likened to using indexes or addresses (Section 7.6). But what are these addresses? In the mammalian brain, sensory neuronal representations are spatially laid out. Different sensory representations are often supported by different neurons, sometimes in distinct brain regions, rather than by different firing patterns within the same set of neurons. This is why in Chapter 3 Section 3.11 we likened the sensory cortices to a piano.

For the prefrontal cortex to be able to target specific first-order states, for the purpose of attention, cognitive control, etc., it must "know" this spatial layout of sensory neurons somehow. It needs to "know" where the top-down signals should go. But once this layout is "known," one can already infer the discriminability between two stimuli fairly easily, at least by approximation. A pair of stimuli are not so discriminable if their spatial addresses are very similar.

An analogy may help: suppose I tell you that there are two words I'm concerned with, and they are both on page 148 of the *Oxford English Dictionary*. By giving you these rough "addresses" (the page number), you can already guess that the two words are quite likely to start with the same letter. In fact, if you know the dictionary well, you may even know exactly what that letter is. And then I can be more precise about the address: let's say one word is at line 12 of the first column, and the other word is at line 15 of the same column. Given that, one can infer that the two words likely share the first few letters. They may even look similar at a glance. All of this can be derived because we know how the dictionary is organized.

So similarly, I argue that higher-order mechanisms in the prefrontal cortex also implicitly "know" the mental quality space, at least approximately. That is because they have to "know" the spatial organization of the sensory neurons, as well as what these neurons represent, in order to allow the relevant *top-down* processes to function. When the perceptual reality monitor "decides" that some neurons representing color red are correctly representing the world right now, by *referring* to these neurons correctly, the mechanism already has the information that the color is nothing like blue or silver. It may be somewhat more like brown, orange, or pink, or maybe purple. But definitely nothing like the taste of vanilla ice cream—at least for me.

9.4 Metacognitive Benefits

The last point about vanilla ice cream is perhaps more than just a silly remark. It may indeed depend on the person. To me, color red is just nothing like ice cream of any flavor at all. But some may feel it is somewhat more like the taste of strawberry ice cream rather than chocolate ice cream. This sort of thing is ultimately somewhat subjective. Some people are more imaginative than others. Yet others experience atypically strong links between stimuli from different sensory modalities—a condition known as synesthesia (Baron-Cohen and Harrison 1997).

Going back to colors, two patches of red may also be more distinguishable for some people than others. Some people may think crimson and scarlet are clearly distinct. In a formal psychophysical test they may do well even at very brief presentations, followed by a mask, for example. Someone like myself may be rather poor at doing that.

This is to say, there are considerable individual differences in how experiences may be subjectively similar to one another, as measured by one's ability to distinguish between them in pairwise comparisons. So, for the higher-order mechanisms to implicitly represent the mental quality space, rich metacognitive information specifically *about oneself* has to be encapsulated. We can unpack this a bit by thinking explicitly about the mental quality space too.

Suppose we have the space laid out in front of us, as in Figure 9.1. This is not to be confused with the typical color space, defined in terms of physical dimensions such as *hue*, *saturation*, and *brightness*. This is a space specific to each person, based on the individual's neural codes, or the ability to distinguish between any two stimuli in the space. Suppose I quickly flash a color patch to the subject and ask if it was crimson or scarlet. Let's say the subject sees the patch as closer to crimson rather than scarlet, and answers accordingly. How confident should the subject be? Regarding this, the mental quality space contains very useful information. If crimson is very far from scarlet, the person should feel more confident. Likewise, based on the space, the person should know that a two-choice discrimination between scarlet and blue would be easier still *because* blue is so far away on the mental quality space.

Perhaps this accounts for why the mechanisms for PRM are also involved in metacognition too (Sections 5.9, 6.8–6.9, and 7.5), at least in part. That is to say, to "know" the mental quality space is to have some kind of metacognitive, self-knowledge.

9.5 Analog Representations

The discussion in Section 9.4 depends critically on the assumption that two stimuli are more easily confused with one another when the relevant neural activity patterns share similar spatial addresses. For a neural pattern, if we add a small amount of neuronal noise to it, the percept should not change very much. The result should be quite similar to the original percept. A radical change in content is only possible if we change the neural activity pattern to a larger degree.

This point may sound so trivial that it just seems tautological. But in fact, this is only true of pictorial or analog representations. For symbolic or sentence-like representations when you change just a single letter in a sentence sometimes the entire message changes. For example, “I will pay my landlord in time” versus “I will play my landlord in time.”

So what are analog or pictorial representations exactly? There are very many definitions, concerning distinct key features of these representations (Beck 2015). But for our purposes, it suffices to focus on the issue of noise tolerance described in the preceding paragraphs. We can say that analog representations tolerate noise more gracefully, as compared to symbolic or sentence-like representations. And then we can say that pictorial representations are a particular type of analog representations, where the content more gracefully tolerates *spatial* noise—that is, noise that changes the spatial properties of the physical representations. For example, if a neuronal firing pattern shifts spatially to a neighboring location of very close proximity, we often expect the content not to change so drastically, as compared to moving the same firing pattern to another group of neurons many synapses away in the cortex. In this sense, the sensory codes in the mammalian brains are both pictorial and analog. (Sometimes I will say that pictorial representations are spatially analog.)

Recall that according to PRM, consciousness is in a sense the interface between perception and cognition. It selects the deserving first-order states for direct impact on higher-cognitive processing at the symbolic level. On this level, thoughts and beliefs are expressed in formats somewhat akin to sentences. As such, noisy sensory inputs are better filtered out, as they may cause dramatic and unpredictable errors on this higher level—they can cause multiplicative errors. Real percepts and our own imagination must be also delineated clearly as such early on, as they have vastly different implications for reasoning.

Despite the discussion in Chapter 8, some local theorists will remain skeptical that subjective experiences have anything to do with this kind of higher-cognitive reasoning. Perhaps this type of processing reminds them too much of good old-fashioned Artificial Intelligence. These symbolic-level rule-based operations may make some impressive chatbots. But it is not clear what it has to do with the real issue here: something so “raw” as *what it is like* to have certain conscious experiences.

But PRM does not say that consciousness happens *at* that higher-cognitive level. The point is just to highlight the possible causal connections, in order to avoid prematurely writing them off from the outset. Like local theorists and biopsychists (Chapter 6 Section 6.6), we too think the first-order states are important. In fact, I would go so far to agree that not only is the content of these states important, the nature of the physical representations themselves probably also matters (at least in most cases).

But as scientists, why would we stop at this realization? We wouldn't just identify these first-order states as the brute "correlates" of consciousness, as if no more can be said about them. Instead, we should understand what makes these states so special. Perhaps they are only "special" because they are functionally unlike sentences or symbolic representations. Perhaps they are "special" only because of their known anatomical and physiological properties. That is, because of their analog nature, and the way these first-order states are spatially organized in the sensory cortices, a high-order mechanism capable of "addressing" specific first-order states also contains the statistical information needed to tell what these states are *like.*

9.6 Revisiting Labeled Lines & Sparse Codes

It may not be so surprising that neural coding in the brain is mostly analog. The basic signal a neuron sends to other neurons, the action potential, or "spike," may seem like an on/off signal. It either happens or it doesn't. But typically, a single spike doesn't mean much. It's the spike rate or pattern that matters (Ainsworth et al. 2012). These are in turn rather analog-like: small fluctuations lead to small changes in content. And neuronal firing does fluctuate a lot. So, being analog may be a nice feature.

This analog nature applies to pretty much all neurons throughout the brain. But it is specifically in the sensory cortices that we find this highly spatially organized layout, where the content of a representation can be easily identified with a neuronal "address." Take the mammalian primary visual cortex for example (Tusa, Palmer, and Rosenquist 1978; Tootell et al. 1988; Engel, Glover, and Wandell 1997). The retinotopic organization means that for any two neurons, if they are close to each other in cortical space, the retinal locations for which stimulation would trigger their firing (i.e., the receptive fields of the neurons) are also more likely to be close to each other. Correspondingly, it means a small change in the spatial neuronal address should lead to a small change in the spatial content of the relevant percept.

Importantly, the coding of sensory neurons also seems relatively specific (Rose and Blakemore 1974; Hubel, Wiesel, and Stryker 1978; Tootell et al. 1988; Ben-Yishai, Bar-Or, and Sompolinsky 1995). For a neuron in the early visual cortex, the receptive field is often just a fraction of the entire retinal space; it seems to mainly "care" about stimuli placed with a small spatial location on the retina. Featural tuning is likewise often narrow. For a neuron, it may only care about line segments of a particular (narrow range of) orientation. It may not respond to motion, color, or sound, for example. Recall from Chapter 3 Section 3.11, sometimes we say that neurons are like "labeled lines" (Gross 2002), as in a neuron can be given a label indicating what it signals: "cats presented in this spatial location," or "45-degree right-tilted lines in that spatial location." With respect to the qualitative content, it is as if this spatially specific label is all that really matters. For the same neurons, increase in firing often just indicates that the luminance contrast increases; the signal is stronger. To change the content qualitatively, we need to activate different neurons. And often, the further away the neurons are, the more different is the resulting content.

Very much related to the concept of "labeled lines," sensory neurons are also sometimes said to adopt a "sparse" code (Olshausen and Field 2004). This means that at a given time, only very few neurons are involved in signaling the presence of a specific stimulus; most other neurons just firing at baseline level. Again, this means that there is a very clear spatial layout. To tell what the stimulus is, one only needs to know which neurons are active—or in other words, "where" is the activity.

That is why so much information can be read out or "decoded" from the fine-grained voxel patterns in functional magnetic resonance imaging (fMRI), especially within the sensory cortices, even though the signals measured are sluggish, reflecting minimal neuronal dynamics. This is also one of main reasons why decoding information from the prefrontal cortex is hard, leading some authors to mistakenly think of fMRI or electroencephalogram activity there as not reflecting specific content. Such methodological difficulty is just as expected because coding in the prefrontal cortex is known to be far less sparse (Rigotti et al. 2013; Fusi, Miller, and Rigotti 2016; Lindsay et al. 2017).

9.7 Fruit Flies

One may argue that vision, like hearing, is intrinsically spatial. Perhaps that's why neural coding is spatial for these modalities? When we see or hear a stimulus, we typically know where it comes from, or where it falls within our sensory space. Do the principles discussed in Section 9.6 apply also to sensory modalities like olfaction? Turns out, there is likewise a similar sparse coding scheme (Vosshall, Wong, and Axel 2000).

In particular, in the well-studied fruit fly olfactory system, there seems to be an active mechanism of *sparsification*. That is, at the sensory receptor level, there are just about 50 types of receptors. Together they form a combinatorial "dense" code—as in the opposite of "sparse," meaning that one single odor may activate many types of receptors, to varying degrees. But the receptors are projected (somewhat randomly) to roughly as many as 2000 Kenyon cells in the mushroom body. Feedback from a later stage of processing means that all but the top 5% of Kenyon cells with the highest-firing are suppressed. This results in a very sparse code at the Kenyon cells level, as if by "design."

What's the advantage of this kind of sparse coding scheme? Computer scientists have been inspired by how common this architecture is in the nervous system, and found that this helps to solve some challenging real-world computational problems. For example, with this sparse code, similar odors are often represented by common Kenyon cells, at least partially. Because only a few Kenyon cells are activated for a specific odor, any overlap in the Kenyon cell's firing pattern between two odors means that the two odors are likely similar, or possibly the same. It was suggested that this can help to overcome the challenging problem of *similarity search*. That is, given an odor, it may be useful to find out what other odors are similar. This can, for example, help us make generalizations for learning; similar odors may represent similar food values. Interestingly, a formal analysis shows that a computational architecture akin to the fruit flies olfactory sparse code system can outperform previous algorithms invented by computer scientists (Dasgupta, Stevens, and Navlakha 2017).

Another study showed that this scheme of sparsified coding can help an agent determine whether a stimulus is novel, that is, not having been encountered before (Dasgupta et al. 2018). Again, formal analysis shows that the sparsified coding scheme found in fruit flies can perform better than previous computer algorithms.

The reader may wonder if this means that the humble fruit fly is conscious, in the sense of knowing what an odor is subjectively *like*. The answer, at least according to PRM, is probably not. It is unclear if fruit flies have perceptual reality monitors (i.e., discriminators in a GANs-like framework). The view put forward here is not that just having analog or pictorial sensory representations is enough for consciousness. Having these representations allow subjective experiences to *potentially* occur with their distinct qualitative characters. But subjective experiences arise *only* when these representations are appropriately addressed by a perceptual reality monitor "knowing" the spatial layout of these addresses.

Also, although the Kenyon cells in fruit flies show sparse coding, the spatial organization may seem to be not as systematic or structured as in the human sensory cortices. At a first glance, the projections from the olfactory receptors to the Kenyon cells look totally random. That is rather unlike the situation in the mammalian visual and auditory cortices, where a clear and systematic spatial structure is preserved.

This lack of distinct spatial structure in coding in specific stages of olfactory processing is not just in the fruit flies, but present also in mammals (Kay 2011). However, we should distinguish between representations that are *physically* spatial analog, and representations that are *functionally* spatial analog. Olfactory coding may not show a clear spatial organization from the perspective of an experimenter holding a microscope. But functionally, it could be fundamentally similar to coding in the mammalian visual system. That is to say, let's say we hypothetically scramble around the neurons in the visual system of a mammal, while keeping the connections between the neurons the same. This should not fundamentally change the basic computational properties of the circuit. It would just muddle physical structure from an external perspective. But functionally, the internal analog structure is determined by how things are wired, which would remain just the same.

So, we can think of the olfactory system as just being physically scrambled. But functionally, the same spatial analog structure is known to exist; similar odors share similar neuronal codes (Endo, Tsuchimoto, and Kazama 2020; Pashkovski et al. 2020). To the extent that both olfaction and vision give rise to qualitative experiences in humans—there is something it is like to smell as much as to see—it seems that what is really important is this *functional* aspect of the spatial analog nature of representations. In neural network models, we sometimes call this representational property "smoothness" (Jin et al. 2020; Rosca et al. 2020).

9.8 Mantis Shrimps

To drive home the point that it is functions rather than sheer physical spatial layout that is really essential, let's consider the example of color vision in the mantis shrimp (Thoen et al. 2014). These crustaceans have over a dozen photoreceptor types, many more than we do. Surprisingly, however, mantis shrimps are not very good at fine-grained color discrimination. That is not because the photoreceptors themselves lack precision. Rather, it has more to do with the ways things are wired up the nervous system. In humans and other mammals capable of color vision, signals from the different photoreceptors are mixed together in an "opponency" scheme, meaning that it is often the relative difference in activation levels between the signaling channels that really matters (Schiller, Logothetis, and Charles 1990). However, in the mantis shrimp, it is as if each photoreceptor type has its own signaling channel, behaving rather independently from the others. Therefore, the mantis shrimp may be able to recognize individual colors detected by the different photoreceptor types. But these receptors don't coordinate together to form a continuous spectrum for fine-grained discrimination.

In other words, we can think of the color vision system in the mantis shrimp as having an extreme "labeled line" structure. Importantly, these labeled lines are relatively independent. A sensory neuron mainly takes input just from one type of photoreceptor and is therefore clear what color wavelength it signals. So in a sense, there is a spatial addressing system too. But it is not a pictorial system. It is not *spatially* analog. If anything, it is spatially discrete and symbolic. Different neurons just signal different color wavelengths, and they don't together form a continuous population code like we do in the human brain.

As such, it is reasonable to expect that when a mantis shrimp detects a color, it cannot tell how subjectively similar it is to the other colors that it can also detect with other photoreceptor channels. There are just different colors. The mantis shrimp cannot (in principle) spontaneously come up with another color similar to the detected one. And to the extent the colors represented by all photoreceptor types have been detected before, the mantis shrimp cannot ever detect a new color and consider it novel. The colors signaled by the number of receptor type channels are all that a mantis shrimp can detect. It does not have the architecture to represent and recognize a new "mix" of the fundamental signals, as a new color.

Together with the example of the olfactory system in fruit flies, hopefully this helps to make the contrast, to indicate what is special about our sensory cortex. Its physical structure generally matters for consciousness. But it only matters for computational reasons: it affords a *functionally* spatial analog (i.e., "smooth" and pictorial) address system. There is no magical "biopsychic" force involved.

9.9 Putting It All Together

In summary, according to PRM, we become aware of the content of a certain first-order (sensory) representation, when a discriminator-like mechanism "decides" that this content is suitable for further downstream cognitive processing (Chapter 7). If it is decided that the first-order activity probably reflects noise, the information will not be actively routed anywhere further. There will be no corresponding subjective experience. Otherwise, depending on whether the first-order information is deemed to reflect the state of the world right now, or some memory of the past, or some imagination of the future, and the like, it would be routed for making an appropriate impact on high-level cognition correspondingly. Global broadcast is one *potential* downstream consequence, which would facilitate some degree of cognitive control. But more important is that the information will be routed for making an appropriate impact on a narrative system capable of causal reasoning. Information processing at this level is symbolic rather than analog. This is one reason why unreliable, noisy signals are best filtered out early.

In Chapter 8, we did not speculate in detail what may be the brain mechanisms for this putative narrative system. Some preliminary evidence suggests that specific areas of the prefrontal cortex may be involved (LeDoux and Lau 2020), but the hippocampus is likely also important. The hippocampus contains cognitive maps (Tolman 1948; O'keefe and Nadel 1978), and is important for autobiographical, episodic memories (Tulving and Markowitsch 1998; Burgess, Maguire, and O'Keefe 2002). The interaction between the prefrontal cortex and the hippocampus is known to be important for the encoding of episodic memories (Eichenbaum 2017).

Whereas the specific regions in the sensory cortices represent various perceptual features, in a subjective memory episode, these different features need to be organized coherently together in terms of spatial and temporal references. The prefrontal and parietal cortices are closely connected, and both seem to be involved in spatial and temporal processing (Bueti and Walsh 2009; Peer et al. 2015; Marcos and Genovesio 2017). But the determination of temporal context likely depends more critically on the prefrontal circuits (Knight and Grabowecky 2000). To the extent that the prefrontal cortex is involved in spatial processing too, it probably more directly supports spatial processing with an egocentric (i.e., self-oriented) rather than an allocentric (i.e., world-oriented) frame of reference (Kesner, Farnsworth, and DiMattia 1989; Ma, Tian, and Wilson 2003). In generating our own self-narratives, it is important to distinguish between what is here and now, and the past or the future, from the point of view of oneself.

In other words, the various different brain regions likely work in concert in support of autobiographical, subjective narrative processing. Within the hippocampus, it is known that the storage of long-term memories does not take the form of a frame-by-frame detailed video-like recording. Instead, one enduring idea is that some kind of index system is used for efficient storage (Teyler and Rudy 2007; Tanaka and McHugh 2018). With these indexes one can retrieve the sensory details from the representations throughout the cortex. This suggests that the "addressing" system proposed earlier for the prefrontal mechanisms to refer to early sensory activity may not be unique.

Perhaps it makes sense for downstream areas to all refer to these same addresses or indexes when they communicate with each other about the relevant sensory content.

For an analogy, this is a bit like the way we pass hyperlinks for internet webpages in emails, without duplicating the detailed content—except that here the "hyperlinks" themselves are structured enough that we know similar links will take us to subjectively similar contents.

In a sense, this means that the different downstream brain regions communicate with an internal phenomenal "language": When the prefrontal cortex signals to the hippocampus that "this" sensory activity reflects the state of the world right now, the hippocampus "knows" that what should be encoded into our narratives is a sensory stimulus that *looks like* something, and yet unlike something else, for example. This may be how the qualitative nature of subjective experience comes about. The primary function of reality monitoring is for routing sensory information, to direct such information toward appropriate downstream symbolic-level processing. But, in doing so, the system implicitly knows what the stimulus in question is subjectively *like*. Because this "knowledge" is implicit, it may be difficult for the subject to articulate it. But it is part of the language through which our different brain mechanisms communicate.

Overall, this extended version of PRM is congruent with what we speculated back in Chapter 5 Section 5.11, about the functions of consciousness. In blindsight, or other forms of nonconscious perception, one should expect certain functions to be either impossible, or at least frequently compromised. That's because when a perceptual process does not lead to subjective qualitative experiences, chances are that the PRM is missing or malfunctioning, or it somehow does not address the relevant perceptual signal as correctly reflecting the state of the world right now. This would imply compromised spontaneous belief formation, and/or metacognition. It is unlikely that later on one would recall the experience vividly from memory. If the PRM is generally compromised *because* the overall prefrontal mechanisms responsible for top-down addressing of sensory signals are malfunctioning, we expect deficits in sensory inhibition and attention too. Alternatively, it could be that the PRM is doing fine, but the perceptual signals are themselves not spatially analog in nature. In that case one would be conscious of the relevant information in the sense of having access, but the relevant "experience" would not be qualitative. Accordingly, these signals would not support the ways our brains perform efficient similarity searches, and novelty detection. So, according to the PRM theory, full-blown qualitative consciousness is causally associated with these specific functions.

9.10 Robot Consciousness Revisited

We can flesh out the implications of this functional account in terms of a concrete example. Let us consider again the Hard Enough Problem (Section 7.13). If the analysis is correct so far, all the ingredients needed for subjective qualitative experience to occur can be simulated in artificial computational systems.

So let us think about the simple robot introduced in Section 7.13 again. Let us equip the robot with smooth, pictorial representations similar to ours on the first-order level. So now when a false alarm occurs at the "bodily damage" sensor located at a fingertip, not only will the robot find that disturbing to its ongoing cognitive reasoning. Not only will there be this stubborn assertoric force, that something is wrong at the fingertip even when all evidence suggests otherwise, the robot will also be able to think about what this sensory state is *like*—it can think about, from its own point of view, what other sensory states are similar to this state, so much so that oneself may mistake those other states as the current state. For example, the current state may be virtually indistinguishable from a similar signal in an adjacent sensor, but it is different from signals coming from another fingertip. The signals from another set of sensors detecting high pressure often co-activate with the current signal, and at high pressure intensity these signals can be confused with one another. But the current signal is nothing like gentle strokes, which are detected by another set of sensors giving very distinct signals. The robot will know all this through "introspection" alone.

Suppose the robot has cameras supporting visual capacities similar to ours too. If we ask the robot what red is like, it may reply that it is somewhat like brown, a bit like purple, pink, but nothing like blue. In particular the robot will not be answering this as if it is a general knowledge question, about the physical similarity between colors. It will answer based on how these other colors may be mistaken as red under suboptimal conditions, by its own cameras and visual processing. It can tell you whether scarlet is more like cherry or crimson, as they are sensed from its own point of view, at the current moment.

Also, when presented with a stimulus that it hasn't seen before, it will spontaneously report: "Wow I haven't seen this before." When asked to describe what it is like, it may say "It is a color patch. Something like right between red and yellow. Probably this is what other people call 'orange.' "

Can we imagine a blindsight patient ever behaving like this? If not, are we really so fundamentally different from this robot?

Ultimately, some may still find it hard to accept that this robot has anything like our subjective experiences. Intuitions vary from person to person. But are we at least somewhat on the right track, or not even close? To the die-hard skeptics, perhaps we should be reminded that our intuitions about the nature of consciousness may well turn out to be illusory—it's not that we can be mistaken about the very existence of consciousness in any form, but we may well grossly mischaracterize its fundamental nature (Frankish 2016). And importantly, remember that the Hard Enough Problem is a relative matter. What more plausible alternatives do we have? An inactive set of logic gates? A piece of brain tissue on a petri dish? A simple network capable of global broadcast?

This is not a book of philosophy, and I'm not a philosopher. But varieties of the structural-relational view have been defended in detail elsewhere, against classic philosophical challenges (Clark 2000; Rosenthal 2010). As in philosophical discussions, the debates continue. Some will always insist that there are unresolved issues, and they may not be entirely wrong about that. But the question remains: what are the alternatives? Are these alternatives going to lead to more meaningful progress?

9.11 Metaphysical Alternatives

Authors who insist that a functional account will never be satisfactory sometimes look for what they call "fundamental" theories of consciousness. Perhaps subjective experiences can only be understood at the foundational level of physical reality.

In the introduction, Sections I.4–I.6, I have already expressed my misgivings about such "physics-centric" approaches. Not all physicists are unreasonable, of course, but in trying to derive "first principles," some researchers end up ignoring basic empirical facts about our brains and psychological functions. Typically, as one gets past the unnecessarily complicated math, what passes on as "axioms" and theoretical principles often turn out to be nothing but unexamined philosophical assumptions.

But these views have also been proposed and defended within philosophy. For example, panpsychism is the view that all physical entities are conscious in some sense (Section 6.6); either they have conscious experiences themselves or they are part of a larger entity having conscious experiences.

The last point is related to what is called the combination problem (Roelofs 2019). So, some panpsychists say that even a single photon is conscious. But the content of my consciousness reflects a relatively unified perspective, not many independent streams of experiences by very many photons, molecules, and so on. So when I become conscious, all these tiny things inside my brain, each are otherwise conscious (according to the view), must somehow combine together to form *my* consciousness. But how does the content of my consciousness relate to the experiences by these photons before they are combined into mine? How does this combination work exactly? Why doesn't everything combine together so that the entire universe may also be conscious?

To the last question, some authors actually give a positive answer (!). In turn, panpsychism has generally been harshly criticized within philosophy. However, in recent years, there are signs that the view is gaining popularity; much effort has been put in to promote the view to the general public. Although scientists generally dismiss the view, some authors try to make it sound like that the view is being taken seriously by (some) neuroscientists.

There are indeed a few influential authors who take such views. But the nature of this topic of inquiry is that you will find all sorts of people advocating for pretty much any kind of view, however radical or improbable. Science should not be a matter of following the subjective opinions of some "influencers." We need evidence. We need logical arguments, not some highly speculative ideas hidden behind abstruse equations or populist authority.

And I'm not sure how this kind of fundamental theory can account for why consciousness matters at all. When I have a conscious experience, I can think and talk about what it is like. When I am in pain, I tend to really want to get rid of it. Just how does some fundamental property of some physical matter account for these functional and behavioral facts? And then there are all those issues mentioned in Chapter 8. A theory of consciousness does not necessarily have to give a detailed account of these higher-cognitive aspects of the mind and rationality. But the problem here is that it is not clear what these fundamental theories can meaningfully say, even in principle, about any such possible connections.

In fact, even for the Hard Problem itself, it's just as unclear if these theories help very much. Just how does stipulating that photons or the entire universe are conscious help to explain consciousness *as we know it*? Even if photons were conscious, supposedly their "experiences" aren't anything like the redness of red, the smell of roses, the sharp pain in a finger, for example.

And even if photons are conscious, just *why* are they so? Can we not imagine a universe in which photons are *not* conscious? So even if they were indeed conscious, are we to accept this as just a brute fact? Why isn't it just as plausible a brute fact that creatures who truthfully think they are conscious are, and photons just aren't?

9.12 End Game?

The last questions I raised in Section 9.11 are unlikely to discourage philosophers holding panpsychist views. To some of them, perhaps their version of brute fact—that everything is somewhat conscious just because that's the way it is—is just somehow more elegant and parsimonious than whatever else we intuitively think. But as in physics, we have also seen how a dogmatic preference of subjective theoretical beauty has led science astray (Hossenfelder 2018).

Amid all the heady metaphysical speculations, it may also be easy to forget why society values science as such. As Richard Feynman famously said, "science is like sex: sometimes something useful comes out, but that is not the reason we are doing it" (Feynman and Leighton 2001). But given all the unresolved issues discussed in the last chapter, sometimes I do wonder: What exactly are we doing here, as a discipline? It may be all good and noble for one to pursue scientific "truths" for their own sake, but our answers to questions as weighty as consciousness will inevitably have practical and ethical consequences. Far from being oblivious to the issues, panpsychists often themselves discuss these potential ethical consequences. It should not be controversial that we owe it to more than our own aesthetics and scholarly ambitions to get things right.

In this context, the following quote may be telling. In defending another metaphysical view, idealism (according to which physical things don't really ever exist outside of our mental lives), Dave Chalmers (2019) wrote: "I do not claim that idealism is plausible. *No position on the mind–body problem is plausible* ... So even though idealism is implausible, there is a non-negligible probability that it is true" (italics mine).

In a way, I agree: there is indeed some non-zero probability, for pretty much anything. But overemphasizing this seems to go against the general spirit of science in practice. Scientists learn to live with imperfect theories. They say all theories are wrong, but some are more useful than others. We sometimes call these "working hypotheses" in order not to commit ourselves to thinking that they are absolutely right or complete. We should ever be open to other possibilities. But we don't invoke and promote more radical views just because the current views aren't perceived to be perfect—unless the said radical views have clear advantages over the more boring, default views.

Perhaps this difference between the disciplines is just as it should be. Philosophers are meant to explore relatively far-fetched ideas. They think ahead for us. It is in their interest to emphasize and defend the value of what they do for a living. And sometimes I agree with them too.

But as science progresses, we expect our views to mature. As more concrete phenomena are accurately predicted, more practical applications generated, eventually our theories should become more easily accepted. So long as we stay focused, we shall get there one day. Historically, this is how difficult problems are typically "solved" in the basic sciences. But this only works if the group of experts who evaluate and permit this progress are in some ways neutral. It may be harder to convince experts of the plausibility of an existing view if their own career interests depend on the very impression of an unresolved scientific mystery.

The last point should not be mistaken for a cynical argument. There need not be any intellectually dishonest ulterior motives on anyone's part. The aforementioned potential lack of neutrality can very well emerge at a group level, as certain types of individual academics with totally sincere dispositions are selected and promoted by the system consistently. Also, given their career status, perhaps it makes total sense *for them* to go for the most radical approach possible. And I am certainly not singling out philosophers; scientists participate in this same evaluative process too, and many also thrive in hypes rather than lasting progress. But all the same, if the development of the science of consciousness itself is hindered by this process, the Hard Problem may well perpetuate as a self-fulfilling prophecy (Lau and Michel 2019).

9.13 Coda: Science & Its Players

When I was young, I thought I would one day write a book to solve the Hard Problem. I still have not given up on the problem itself. But by now I recognize that it was a deluded idea. At the moment, the best that I can offer is summarized in this final chapter, an extended version of PRM which includes some elements of quality space theory. Little of that is original, and some readers will no doubt remain unconvinced.

But worse still, I now realize that perhaps no amount of books will ever be enough to solve the problem. To solve a problem we need to first recognize its nature. Certain problems require collective action. As Thomas Kuhn (1962) famously put it, in science, there are ultimately matters that "can never be unequivocally settled by logic and experiment alone." To deny the sociohistorical aspects of science is to "dehumanize" its players.

I joined the field when the late Francis Crick was still an active champion of the problem of consciousness. In the early 2000s up till his passing, there was a general sense of optimism and promise: theoretical progress shall be built upon the solid foundation of empirical caution. I'd like to think this book represents a small step inspired by that tradition. Much of the details given here will likely turn out to be wrong. But it may be good enough if I shall turn out to be wrong rather than not even wrong. I have to count on future researchers more capable than myself to correct my many errors.

In recent years, the direction of the field has become somewhat diffused. Scientific disagreements have become increasingly difficult to resolve with conventional methods. Some may think that recent trends like "open science" may help. But this overlooks some unique features of the field. There are often deep theoretical differences between research groups. But above all, I also suspect that the bigger problem may well be the *system* itself.

Since the 1990s, our field is unique in that media visibility matters disproportionately. It has become a perfectly viable strategy to ignore peer opinions and critics within the academic expert group. To advance one's own career and ideas, one would do just as well to focus on impressing editors and private donors, via personal connections and populist appeal. One may think these problems are common in other disciplines too. But our field is unique in how rampant they are, as standard mechanisms of scientific evaluation and open competition are stifled by the relative lack of public funding and tenure-track jobs, especially in the United States.

Our larger-than-life media presence means that when newcomers approach the topic, they often don't feel the need to consult the existing literature very carefully. To some, this is just an exciting new playground best suited for trying out something risky and "different." To be fair to them, following the literature can also be difficult, as known empirical falsehoods are often repeated in high profile journals, sometimes by "authoritative" figures. And yet, as a developing field we count on these newcomers to take our subject matter seriously.

Besides media hype, philosophical opinion has also become a significant factor influencing science funding. Of course, I value interdisciplinary exchanges highly. We need philosophers as critics, as well as their constructive conceptual analysis. But if they serve also as gatekeepers who control which scientists' views are promoted as gaining momentum in prominent venues, the balance can become problematic. Certain views favored by philosophers may be intellectually interesting, but they aren't necessarily conducive to scientific progress.

Accordingly, in the past few years alone, we have seen tens of millions of US dollars of private funding poured into the field, with a particular focus on the more "theoretically ambitious" projects. For our small field, the effects will no doubt be felt for decades to come.

As in any system, there are pros and cons. The future will tell whether these are for good or for ill. My only wish is that we recognize these deep structural issues rather than deny their existence.

I have also met highly influential colleagues who vehemently defend the ways things are currently done, over more "traditional" scientific models. The argument is that our science is intrinsically special, and, thereby, it requires unconventional methods with more open mindsets.

The reader may not be surprised that I do not agree with these colleagues. People will naturally and understandably defend the very system from which their careers have benefited. I myself have also done alright in the existing system. But to my mind, the field will ever remain "special," in not necessarily very good ways, if this is how we choose to continue to comport ourselves. To have any chance of inching toward our fabled "end game," we have to think about the institutional contexts that allow good science to happen, that allow *others* to succeed. We want the field to be represented by people who are fair. We need a literature that we can *trust*. But I shall refrain from discussing these issues more than I already have. After all, this is a book about science. I've been told many times by colleagues that, as scientists, it is better for us to *focus on the science* rather than to "politicize" it.

I have often wondered about the last point. Perhaps an analogy would be: as citizens, we should also *focus on living* and leave politics to the politicians. Again, there may be pros and cons for different approaches. But I submit: this may be the Truly Hard Problem of consciousness.

References

Ainsworth M, Lee S, Cunningham MO et al. Rates and rhythms: A synergistic view of frequency and temporal coding in neuronal networks. *Neuron* 2012;**75**:572–583.

Baron-Cohen SE, Harrison JE (eds). *Synaesthesia: Classic and Contemporary Readings*. Blackwell, 1997.

Beck J. Analogue magnitude representations: A philosophical introduction. *Br J Philos Sci* 2015;**66**:829–855.

Ben-Yishai R, Bar-Or RL, Sompolinsky H. Theory of orientation tuning in visual cortex. *Proc Natl Acad Sci U S A* 1995;**92**:3844–3848.

Bueti D, Walsh V. The parietal cortex and the representation of time, space, number and other magnitudes. *Philos Trans R Soc Lond B Biol Sci* 2009;**364**:1831–1840.

Burgess N, Maguire EA, O'Keefe J. The human hippocampus and spatial and episodic memory. *Neuron* 2002;**35**:625–641.

Chalmers D. Idealism and the mind-body problem. In: W Seager (ed), *The Routledge Handbook of Panpsychism*. Routledge, 2019; 353–373.

Clark A. *A Theory of Sentience*. Clarendon Press, 2000.

Dasgupta S, Sheehan TC, Stevens CF et al. A neural data structure for novelty detection. *Proc Natl Acad Sci U S A* 2018;**115**:13093–13098.

Dasgupta S, Stevens CF, Navlakha S. A neural algorithm for a fundamental computing problem. *Science* 2017;**358**:793–796.

Eichenbaum H. Prefrontal-hippocampal interactions in episodic memory. *Nat Rev Neurosci* 2017;**18**:547–558.

Endo K, Tsuchimoto Y, Kazama H. Synthesis of conserved odor object representations in a random, divergent-convergent network. *Neuron* 2020;**108**:367–381.e5.

Engel SA, Glover GH, Wandell BA. Retinotopic organization in human visual cortex and the spatial precision of functional MRI. *Cereb Cortex* 1997;**7**:181–192.

Feynman RP, Leighton R. *"What Do You Care What Other People Think?": Further Adventures of a Curious Character*. WW Norton & Company, 2001.

Frankish K. Illusionism as a theory of consciousness. *J Consciousness Studies* 2016;**23**:11–39.

Fusi S, Miller EK, Rigotti M. Why neurons mix: High dimensionality for higher cognition. *Curr Opin Neurobiol* 2016;**37**:66–74.

Gross CG. Genealogy of the "grandmother cell." *Neuroscientist* 2002;**8**:512–518.

Hossenfelder S. *Lost in Math: How Beauty Leads Physics Astray*. Hachette UK, 2018.

Hubel DH, Wiesel TN, Stryker MP. Anatomical demonstration of orientation columns in macaque monkey. *J Comp Neurol* 1978;**177**:361–380.

Jin P, Lu L, Tang Y et al. Quantifying the generalization error in deep learning in terms of data distribution and neural network smoothness. *Neural Netw* 2020;**130**:85–99.

Kay LM. Olfactory coding: Random scents make sense. *Curr Biol* 2011;**21**:R928–R929.

Kesner RP, Farnsworth G, DiMattia BV. Double dissociation of egocentric and allocentric space following medial prefrontal and parietal cortex lesions in the rat. *Behav Neurosci* 1989;**103**:956–961.

Knight RT, Grabowecky M. Prefrontal cortex, time and consciousness. *The New Cognitive Neurosciences*. (2nd ed), Cambridge, MA: The MIT Press, 2000; pp. 1319–1339.

Kuhn TS. *The Structure of Scientific Revolutions: 50th Anniversary Edition*. University of Chicago Press, 1962.

Lau H, Michel M. A socio-historical take on the meta-problem of consciousness. *Journal of Consciousness Studies* 2019;**26**:136–147.

LeDoux JE, Lau H. Seeing consciousness through the lens of memory. *Curr Biol* 2020;**30**:R1018–R1022.

Lindsay GW, Rigotti M, Warden MR et al. Hebbian learning in a random network captures selectivity properties of the prefrontal cortex. *Journal of Neuroscience* 2017;**37**:11021–11036.

Marcos E, Genovesio A. Interference between space and time estimations: From behavior to neurons. *Front Neurosci* 2017;**11**:631.

Ma Y-Y, Tian BP, Wilson FAW. Dissociation of egocentric and allocentric spatial processing in prefrontal cortex. *Neuroreport* 2003;**14**:1737–1741.

Nagel T. What is it like to be a bat? *Philos Rev* 1974;**83**:435–450.

O'keefe J, Nadel L. *The Hippocampus as a Cognitive Map*. Clarendon Press, 1978.

Olshausen BA, Field DJ. Sparse coding of sensory inputs. *Curr Opin Neurobiol* 2004;**14**:481–487.

Pashkovski SL, Iurilli G, Brann D et al. Structure and flexibility in cortical representations of odour space. *Nature* 2020;583:253–258.

Peer M, Salomon R, Goldberg I et al. Brain system for mental orientation in space, time, and person. *Proc Natl Acad Sci U S A* 2015;**112**:11072–11077.

Rigotti M, Barak O, Warden MR et al. The importance of mixed selectivity in complex cognitive tasks. *Nature* 2013;**497**:585–590.

Roelofs L. *Combining Minds: How to Think about Composite Subjectivity*. Oxford University Press, 2019.

Rosca M, Weber T, Gretton A. and Mohamed S. A case for new neural network smoothness constraints. Proceedings on "I Can't Believe It's Not Better!" at NeurIPS Workshops, in Proceedings of Machine Learning Research 2020;137:21–32. Available from https://proceedings.mlr.press/v137/rosca20a.html.

Rose D, Blakemore C. An analysis of orientation selectivity in the cat's visual cortex. *Exp Brain Res* 1974;**20**:1–17.

Rosenthal D. How to think about mental qualities. *Philosophical Issues* 2010;**20**:368–393.

Schiller PH, Logothetis NK, Charles ER. Role of the color-opponent and broad-band channels in vision. *Vis Neurosci* 1990;**5**:321–346.

eTanaka KZ, McHugh TJ. The hippocampal engram as a memory index. *J Exp Neurosci* 2018;**12**:1179069518815942.

Teyler TJ, Rudy JW. The hippocampal indexing theory and episodic memory: Updating the index. *Hippocampus* 2007;**17**:1158–1169.

Thoen HH, How MJ, Chiou T-H et al. A different form of color vision in mantis shrimp. *Science* 2014;**343**:411–413.

Tolman EC. Cognitive maps in rats and men. *Psychol Rev* 1948;**55**:189–208.

Tootell RB, Switkes E, Silverman MS et al. Functional anatomy of macaque striate cortex. II. Retinotopic organization. *J Neurosci* 1988;**8**:1531–1568.

Tulving E, Markowitsch HJ. Episodic and declarative memory: Role of the hippocampus. *Hippocampus* 1998;**8**:198–204.

Tusa RJ, Palmer LA, Rosenquist AC. The retinotopic organization of area 17 (striate cortex) in the cat. *J Comp Neurol* 1978;**177**:213–235.

Vosshall LB, Wong AM, Axel R. An olfactory sensory map in the fly brain. *Cell* 2000;**102**:147–159.