6
A Centrist Manifesto

6.1 Striking a Balance

The main goal of this chapter is to take stock of what we have reviewed so far, and to assess the broader theoretical implications. We have covered the first three of the five main issues introduced in Chapter 1: the neural correlates of consciousness (NCC), the relationship between attention and consciousness, and the functions of consciousness. The remaining two issues concern whether animals and robots can be conscious. To address them we need to make some theoretical generalizations. Fortunately, the previous chapters provide some constraints for what an adequate theory should look like.

Overall, neither the global nor local view works well. Instead the evidence points to a synthesis. The key findings from the previous chapters will be summarized here.

Throughout, I will also introduce some new empirical and theoretical considerations, especially in Sections 6.6–6.9.

6.2 Troubles for Global Theories

The global view faces several empirical challenges. The first is that when experimental confounders are controlled for, the activations in the prefrontal and parietal cortices are not as widespread as we once thought (Chapter 2). In some cases null findings were overinterpreted: when reports were not required, there were actually still clearly observable activity in the prefrontal and parietal cortices; this depends on the measurement methods. But it is fair to say that the activity becomes more subtle under these conditions. Likewise, controlling for task-performance capacity reduces activity in these regions. These findings suggest that global broadcast may largely support task performance and reports, rather than subjective experience per se.

Congruent with this interpretation is the fact that information represented in the workspace does not always come with subjective experience. As discussed in Chapter 5, Dehaene and colleagues have themselves been studying

In Consciousness We Trust. Hakwan Lau, Oxford University Press. © Hakwan Lau 2022.
DOI: 10.1093/oso/9780198856771.003.0007

nonconscious working memory by applying visual masks to the relevant stimuli. But even for typical, unmasked stimuli, there aren't always strong perceptual experiences during the working-memory delay. For example, if one has to memorize a visual pattern for 10 seconds, one may invoke visual imagery of the pattern during the delay. But this visual imagery is not generally confused with the subjective experience of conscious perception. More importantly, some people never experience vivid visual imagery—a condition known as aphantasia (Zeman, Dewar, and Della Sala 2015). And yet they seem to have no trouble doing these working-memory tasks. According to global theories, during the working-memory delay, the information is maintained in part by workspace mechanisms. Why does the content of working memory not "leak out" into consciousness, so that we subjectively perceive it as if the stimulus is right in front of us?

And then other nonconscious stimuli also seem to reach the prefrontal cortex and influence higher cognitive functions, as shown in subliminal priming studies (Chapter 5). Just why don't these stimuli lead to conscious experience, given that they have reached the workspace?

Similarly, the global view may have troubles accounting for certain perceptual phenomena such as blindsight. If the explanation is that such information fails to reach the workspace, one needs to account for how they can lead to task performance at sometimes above 80% correct (in, e.g., two-choice discriminations). What may be the mechanisms that allow a signal to influence behavior so strongly, and yet bypass the workspace?

Dehaene and colleagues may answer that there could be a nonconscious channel operating *in parallel*, which accounts for blindsight and other nonconscious behavior. But when this parallel model was formally compared against *hierarchical* models, the latter performed better (Chapter 3 Section 3.8). This is congruent with the fact that conscious and nonconscious processing both depend on the very same early sensory areas (Chapter 2). There are no exclusive early sensory pathways for conscious perception only. Instead, the difference between conscious and nonconscious processing seems to depend on some specific late-stage process (the hierarchical model). It is unlikely that such a late-stage process is the global workspace itself. That's because the workspace is meant to be functionally important. If the difference in late-stage process is associated with such a big functional difference, we would expect consciousness to come with very many functional advantages. And yet it doesn't. To the extent there are such advantages, the most likely candidates seem to be metacognition, and to a lesser extent, inhibitory control (Chapter 5). But global broadcast functions are sometimes thought to be independent from metacognition (Dehaene, Lau, and Kouider 2017).

Not only are there cases where functionally strong perceptual signals aren't conscious, the opposite scenario also seems possible. One example is peripheral vision, or subjective perception in the unattended background (Chapter 4). There, the information seems not to be processed well in the central workspace. People are poor at reporting the details, or they miss salient events altogether. And yet, subjectively, there seems to be an inflated sense of experience; the unattended background looks more vivid than expected, given how poorly the details are actually represented.

Finally, lesion studies also put some pressure on the global view. Prefrontal and parietal lesions do not seem to abolish broadcast. There are important caveats to keep in mind as we interpret lesion data; they do not straightforwardly tell us whether certain brain areas are "necessary" (see Sections 2.3 and 3.2). However, despite this caveat, it is true that prefrontal lesions impair metacognition and response inhibition. But these patients continue to be able to perform many other perceptual and cognitive tasks at a high level, as if the global broadcast mechanism remains unaffected.

These are not the only problems for the global view. As we'll see in the next chapters, the view also predicts that very simple computer programs and robots may be conscious. That may be considered implausible by some. But for now, the evidence based on human data may suffice.

6.3 Rejoinders?

The global theorist can perhaps bluntly deny the relevance of lesion studies. Perhaps the lack of impact on workspace functions is because of the resilience of the frontoparietal network. When one part is damaged another can take over (as discussed in Chapter 3). It is also true that inhibitory control can be considered a higher cognitive function related to workspace mechanisms.

They can also point to the paucity of decisive evidence against the parallel channels model. Maniscalco and Lau (2016) was just a single study, and its conclusion awaits to be confirmed by more studies directly comparing models, with different datasets. Likewise, they can write-off subliminal priming studies because most of those effects are small. Perhaps the global workspace is needed to exercise those cognitive functions fully.

But I'm not sure how a global theorist can satisfactorily address the apparent double dissociation constituted by the cases of working memory and peripheral vision. By double dissociation, I'm referring to the fact that information represented in the workspace is sometimes nonconscious (e.g., visual working memory, especially in aphantasics), and that conscious experiences

sometimes outstrip information represented in the workspace (as in cases of peripheral or unattended perception).

Regarding peripheral vision or perception in the unattended background, perhaps the global theorist can say that subjective experience isn't really inflated. They can insist that the experience is actually just as sparse as the represented content, only that we are mistaken about the experience. In Chapter 4 we discussed that the subjective appearance of richness may not be universal. But at least some subjects feel that they see more than they have access to, at least under some conditions. For the globalist reply to work, we need to deny that they are right about their own experiences. We need to say that subjects are only aware of what their global workspace can represent, but not more, regardless of what they think. This leads to a conceptual question: to what extent can we really be mistaken about our conscious experiences? If we honestly feel that the peripheral perception is rich, how wrong can we be? Or consider this alternative: if we honestly feel a blinding headache, what does it matter if someone says we are mistaken?

These problems should be considered in the context of the frontal and parietal activations found in studies of consciousness. The current domination of the global view has much to do with the supposedly widespread and robust nature of these activations. But in the light of the new findings of much more subtle activity, when confounders were controlled for, it may be time to revisit whether our initial enthusiasm for the global view is justified. To the extent that some fronto-parietal mechanisms are critical for consciousness, it may be much more specific than a widespread broadcast mechanism. If we detach the notion of consciousness from such a general functional network, the problem of the double dissociation discussed may be easier to address. Perhaps global broadcast does happen often for consciously perceived stimuli, but only as a typical downstream consequence rather than as a constitutive mechanism.

6.4 Troubles for Local Theories

The local view likewise cannot account for the present evidence. While local theorists argue that the activity in the prefrontal and parietal cortices was reduced when report and attention were controlled for, the activity was not completely gone. In particular, this seems to depend heavily on the measurement method. Using invasive methods at high resolution, the measured activity remained strong even under these controls (Chapter 2). This suggests that the activity involved in conscious perception goes beyond the local sensory circuits.

Furthermore, it is unclear if the local activity itself survives the control of similar confounders. When task-performance capacity was controlled for, null results were also obtained in the visual cortex (Section 2.11). As in the case of the prefrontal cortex, we should not overinterpret individual null results. But it does raise the question of whether activity in the visual cortex really just drives basic visual processing for potential task performance, or subjective experience per se. This worry is highlighted by the phenomenon of nonconscious binocular rivalry. Invisible stimuli leading to such rivalry seem to activate the visual cortex just as visible stimuli do (Section 2.10). So, early sensory activity alone is not always associated with subjective experience.

Similarly, regarding the case of unattended or peripheral perception, the localist argument against the global view may not work well or may even backfire. Local theorists appeal to the fact that subjective experience seems rich, at least to some subjects. Because prefrontal mechanisms are supposed to have limited processing capacity, they are thought to be "overflown" by the richness of experience. But there are two problems with this argument. The first is that the role of the prefrontal cortex may not be to "duplicate" the sensory information. Rather, it may just monitor and redirect information in the sensory cortices, using something akin to indexing mechanisms. If so, the putative limited capacity may not be an issue.

The second problem is more directly challenging for the local view itself: it is unclear if the early sensory cortices represent perceptual information detailed enough to account for the subjective richness. In particular, although the level of reported richness may vary across people, it seems relatively stable within a person. However, early sensory activity is modulated strongly by attention. Even holding the spatial focus constant, a mere change in the task is enough to robustly change early sensory activity. And yet the subjective experience of richness does not seem to change nearly as much (Chapter 4).

Overall, there seems to be many instances in which the content reflected by early visual activity just does not seem to match with the content of conscious perception (Section 2.5). This is especially a problem if we focus on a specific local view, according to which the NCC involves a specific visual area (e.g., (feedback to) V1). The different sensory areas all exist for important functional reasons. Depending on the stimuli and context, is it likely that perception will capitalize on all possible sensory resources under different situations. At times, the perceptual phenomenology is probably too complex to be described by just content at one level (Section 4.12).

As with the global view, some local theories also tend to implicitly adopt the parallel channels model. In order to account for strong nonconscious perceptual processes (e.g., in blindsight), some alternative pathways presumably

need to be invoked. But as we mentioned earlier in Section 6.2, the parallel channels model is not well-supported empirically, compared to the hierarchical model.

One way to avoid this problem may be to think of feedforward processes (e.g., from V1 to MT) as nonconscious, such that feedback (e.g., from MT back to V1) may constitute the later stage process within a hierarchy. This would give us a hierarchical interpretation of Lamme's recurrency theory. However, the evidence in support of the role of feedback processes in conscious perception is confounded by task-performance capacity and attention. That is, when feedback processes were supposedly disrupted, not only was subjective experience abolished, processing capacity was also much weakened (Manita et al. 2015; Peters et al. 2017a). As such, it is unclear if feedforward processes alone can lead to nonconscious perceptual signals which are as strong as those supported by recurrency. This makes the hierarchical interpretation problematic because the late stage in the model is not supposed to contribute directly to task performance itself.

Let's assume, for the sake of argument, that there can be strong nonconscious perceptual processes without feedback. One problem is that this version of a local view makes rather implausible predictions about the functions of consciousness. As suggested by van Gaal and Lamme, the nonconscious feedforward signal can propagate into the prefrontal cortex to exercise higher cognitive functions (Section 5.5). This may seem compatible with studies of subliminal priming. However, according to this local view, these signals can be very strong. But subliminal priming effects are invariably weak. The local theorist will have to predict that at least under some conditions, there can be nonconscious effects on these higher cognitive functions that are just as strong and robust as conscious cases. Given current findings, this seems improbable. I am aware of no reports so far on *fully preserved* higher cognitive functions when conscious perception of the relevant stimuli is abolished (via blocking of feedback processing). Overall, it seems much more plausible that feedback processing in the sensory cortices is important for perception in general, but not specifically for subjective experience per se.

Finally, local theorists often criticize the global view based on lesion cases, but their own view is in fact just as problematic in this context, if not more. It is true that in blindsight, the abolishment of visual awareness happens after lesions to the primary cortex. However, blindsight is most clearly established for static stimuli. By magnetically stimulating the remaining extrastriate areas, researchers successfully induced subjective experience of motion in a blindsight patient, where the corresponding V1 was absent (Section 2.4). There have also been reports about patients experiencing hallucinations after

damage to early visual areas (Lau and Brown 2019). Vivid visual mental imagery is also possible after V1 damage (Bridge et al. 2012). In subjects from the general population, V1 itself also seems relatively deactivated in dreams (Braun et al. 1998).

6.5 Contrivance

The localists can insist that strong subliminal modulation of higher cognitive functions is in fact possible. Maybe the problem is just that we have not *yet* figured out how to truly selectively abolish feedback to early sensory areas, without compromising the overall perceptual signal, in order to induce such strong nonconscious percepts. In the absence of positive evidence, I'm not so sure how plausible a hierarchical interpretation of their theory is. But perhaps, like the global theorists, they can also deny the current evidence against the parallel channel model, on the grounds that more studies are needed.

Still, the burden of proof should be on the local theorists to demonstrate that their early sensory correlates are not merely driven by the confounder of performance capacity. As it currently stands, the confounder seems to be the most likely explanation. That is, early sensory activity drives perception, but not subjective experience per se.

As to prefrontal activity surviving various confounds, the local theorists can insist that such activity is weak. They can also argue that evidence for their causal involvement in consciousness (e.g., from lesion and stimulation) is not so strong. But the analysis of the NCC is a logical matter. It shouldn't be based on the *impression* of what looks more obvious given our current methods. We should not treat weak effects as nonexistent, especially if we have independent reasons to expect such effects to be weak, given the anatomy and physiology of the prefrontal cortex (Section 3.11).

However, I concede that prefrontal involvement in subjective perceptual experience is in fact subtle. So the localists may be right that the global position is not sound. Subjective experience may not always constitutively depend on something as involved as global broadcast. But it does not mean that the localists are right in writing off the prefrontal cortex entirely either.

Another difficult challenge for local theorists is the issue of content mismatch (Section 2.5). Again, like the global theorists, the localists can perhaps insist that whatever content is reflected by early sensory activity, that is in fact the content of conscious perception. Any discrepancy may be due to the subjects' mistaking what they truly perceive. But some of these

illusions are so vivid and easy for all of us to see (Section 2.5). To say we can be so consistently wrong about our own experiences just seems rather unconvincing.

Above all, it should be clear that the localist NCC candidate of feedback to V1 isn't very promising (Hupé et al. 2001; Huang et al. 2020). There are clear cases of occurrence of subjective experience in the absence of V1 activity. To say that we are mistaken about some details of our perceptual content is one thing; it's a different matter to say that subjects regularly misconstrues themselves as having perceptual experiences when there are actually none (according to one's assumed notion of NCC). Therefore, to the extent a local view is plausible *at all*, the likely NCC candidate is probably extrastriate local activity, rather than feedback to V1.

6.6 Two Opposing Dogmas

The above leaves us with a conundrum. Philosophers like Ned Block sometimes favor the local recurrency view, I think largely because "recurrency" gives a flavor that it is a special kind of biological activity. They are hoping to look for that unique physical substrate, to be identified with the inexplicable characters of subjective experience (or to otherwise explain them away somehow).

Given that V1 is unlikely to be truly part of the NCC, in order to retain the notion of "recurrency," perhaps one can hold that visual awareness constitutively depends on interactions between the prefrontal cortex and extrastriate areas (Huang et al. 2020). This view may not be so implausible. However, this will not strictly be a *local* view.

More importantly, the problem is that there is actually nothing magical about recurrent activity. Cortical areas are bidirectionally connected. Recurrency is likely just the normal way different areas work together effectively. Many artificial neural networks also employ feedback architectures.

So, if the local NCC is just extrastriate activity rather than interregional recurrency, what is so special about such activity? For truly local theorists, the downstream impact (e.g., to the prefrontal cortex) of such activity shouldn't matter. What is so special about these signals within an area, which may not even end up having any downstream impact at all?

One extreme answer is that there is nothing special. Any physical objects that can signal or represent information are conscious to some extent (Chalmers 1996; Roelofs 2019). Because even a single photon can arguably

"represent" some minimal amount of information, the view implies that pretty much all physical things can be conscious. This rather far-fetched view, called *panpsychism*, remains mostly a matter of philosophical conjecture. As such, I hesitate to even mention it here. However, although the view lacks serious scientific attention, it has generated considerable public excitement in recent years. In Chapter 8 I will discuss some disadvantages of this position. It would most likely disallow us from making useful connections with much of the rest of academia. And then in Chapter 9 we will finally revisit this "problem" again.

A more scientifically acceptable position may be biopsychism (Godfrey-Smith 2016). On a strong version, all and only all biological organisms are conscious. On a weaker version, only biological organisms are conscious; some aren't. Let us focus on this latter, weaker version of biopsychism. According to this view, perhaps the need to self-regulate metabolic activity is an important ingredient for consciousness, but downstream impact on information processing is not. That is, if we replace those extrastriate spiking activities with silicon chips and electrodes exactly mimicking the outgoing signals, the subjective experience may well be gone (or at least drastically different). The correct substrate has to be "biological."

I suspect that some cognitive neuroscientists will find this biopsychic view puzzling, if not downright absurd. The modern approach to studying the brain is to think of it as an organ for information processing. We analogize brains with computers. Brain mechanisms exist because they *function* in a certain way. But all the same, some of the most prominent local theorists do endorse biopsychism. Arguably, this is the logical end point of local theories: effective information processing ultimately depends on the global context of the entire system. If local theorists want to have nothing to do with that, their view is unavoidably at odds with the very premise of cognitive neuroscience. They are essentially skeptics of the information processing approach to understanding the mind.

One motivation for biopsychism may be that the obvious alternatives may not seem so appealing. Some local theorists may argue against *functionalism*, the idea that subjective experience is determined by the roles played by the relevant substrates in informational processing. Specifically, one view that has been systematically criticized is what we can call *long-arm* functionalism (Block 1990). According to the view, consciousness is determined by what is targeted or represented by the system *in the environment*. For example, some brain activity is about an apple in front of us because it tracks and is about the apple. To the extent that the activity tracks the apple, and that it signals the rest of the system to refer to *that* apple, the

activity plays the relevant "long-arm" functional roles. Long-arm functionalism isn't so attractive because it might be thought to equate consciousness with behavioral functions. It is all about how the brain or a physical system can relate to objects out there, and act accordingly. Arguably, nonconscious perceptual processes can also play these long-arm functional roles. Also, the same represented apple may cause different subjective experiences, depending on contexts, perspectives, etc.

To take it to the "globalist" extreme, some versions of functionalism may further postulate that language is needed to relate to these objects in the environment in a truly meaningful way. To be conscious of the objects, we may need to be able to form explicit thoughts about them, and to potentially articulate such thoughts to others and ourselves. This kind of view is often criticized for overintellectualizing consciousness. That is, it misconstrues the sheer occurrence of simple subjective experience as something that requires sophisticated forms of cognition and intelligence.

Fortunately, not all versions of functionalism are of the long-arm type, nor do they require overintellectualization. In fact, most modern cognitive neuroscientists don't endorse such views. All they hold is that information processing is all that matters; the hardware, that is the physical substrate, doesn't really matter, *to the extent that the relevant algorithms are correctly implemented*. As such, the "long-arm" functional roles are only a small part of the story. How these neural representations function *internally*, that is, how they impact downstream cognition, also matters. We can call this view internal functionalism. But sometimes I will just call it functionalism, because long-arm functionalism is a just strawman in this specific context.

This broader notion of (internal) functionalism can sit comfortably between the two extreme dogmas: that consciousness is either something to be equated with sophisticated cognition, or something physically characterized such that it may not do much cognitively at all. Neither is right. But the functional role played by conscious neural representations may not be to broadcast globally to the entire system, in order to lead to *more* cognition in general. It is likely something more specific and modest. So this would avoid much of the problems faced by the global view. But still, whatever functions it involves, it would likely have some specific impact on downstream cognition and behavior. That would explain the subtle involvement of the prefrontal cortex in studies of conscious perception.

Nor do I write-off the concerns of biopsychists entirely; perhaps something about the physical substrates does matter too. I will try to account for this latter intuition in Chapter 9, within a functionalist framework.

6.7 Varieties of Centrism

What may be a functionalist account of consciousness that does not equate consciousness with global broadcast? There are many variants that would more or less qualify (Brown, Lau, and LeDoux 2019). For example, I am sympathetic to Axel Cleeremans' self-organizing metarepresentational account (Cleeremans et al. 2020). According to the view, subjective experience arises when the brain learns to represent its own sensory states for the purpose of hierarchical control of perceptual information and action. That is, the information in the early sensory states are being redescribed at a later stage, for self-monitoring purposes.

This view shares some similarity with a broad class of philosophical theories known as higher-order views. One influential version is David Rosenthal's higher-order thought theory (2005), which we will discuss briefly in Chapter 7. One key characteristic of this family of views is that the relevant higher-order mechanism is meant to be much more specific than global broadcast. To the extent that it is functionally important, it may only serve a rather narrow range of purposes. It may not strengthen the relevant internal perceptual signal or make it more stable, sustained, or complex. In fact, Rosenthal himself argued that consciousness may come with little or no added utility. To some, this may sound extreme and implausible. I myself also do not share this "minimalist" view on the functions of consciousness. But in contrast with global views, we can see why this is a relatively moderate, "centrist" position. Like a globalist position this is a functionalist account, and yet like a local view it does not assume that consciousness is just the same as more powerful and effective cognition in general. This would fit much better than the evidence reviewed in Chapter 5.

Of note is the fact that Richard Brown is also a higher-order theorist (LeDoux and Brown 2017), and yet he is more sympathetic to biopsychism than to functionalism (Brown 2012). According to such a view, perhaps the relevant self-monitoring (i.e., higher-order) mechanisms need to be implemented biologically. This highlights the fact that within this space between the two theoretical extremes, there are really many options available. To be a "centrist" is to resist the global and local extremes. But this does not on its own commit the theorist to functionalism or biopsychism; although most "centrists," including myself, are functionalists (in the broad, internal sense described in the Section 6.6). Like I said in the previous section, we will address some biopsychic intuitions again in Chapter 9.

Instead of doing a detailed comparison between all possible "centrist" positions, I will instead highlight two sets of considerations, one empirical and one

theoretical, inspired by current models of artificial intelligence. Together, they put further constraints on what a plausible centrist theory should look like.

6.8 Metacognition & Detection

In Chapters 3 and 4, we pointed out that metacognition seems relevant to subjective experience. In a sense, we could consider this a brute fact: we identified areas of the prefrontal cortex as important for consciousness, mostly based on findings from neuroimaging (Chapter 2). When these areas were targeted by magnetic stimulation and chemical inactivation, or if they were lesioned, we found that metacognition was also affected (Chapter 3). In particular, at least in some studies, this was specific to perceptual rather than mnemonic metacognition.

However, that two functions employ a common brain region may not mean much; we only have so many brain regions at this coarse-grained level. But conceptually there may be a deeper link between metacognition and consciousness too. If one consciously perceives something, it seems to make little sense to say that one has zero confidence about any aspect of the percept. Of course, sometimes we *see* something without being able to recognize what the object is. But at least we should be fairly sure that something is *seen*, not *heard*. If one is truly unsure what modality the percept involves, perhaps it is doubtful whether there is any subjective experience at all.

That is to say, having a subjective experience seems to involve *detecting* the presence of some signal in the relevant sensory modality (Fleming 2020). This is not to suggest they are one and the same. But incidentally, as we mentioned in Section 3.9, lesions to the prefrontal cortex can impair behavior in some detection tasks too. In peripheral or unattended vision, we also know that subjects use a relatively liberal detection criterion—a phenomenon I called "inflation." In Chapter 4, I speculated that the underlying mechanism may be in the prefrontal cortex, which we know is important for perceptual metacognition.

Why are metacognition and detection linked? Theoretically, they do not have to be. If the task is discrimination between two stimulus alternatives (whether an apple or an orange is presented), one can just compare the two relevant signals against each other. Whether they are both strong or weak does not strictly matter; we only need to know the direction and magnitude of the *difference* between the two signals. If the difference is large, we rate high confidence; if the difference is small, we rate low confidence. But empirically we know that human subjects often do not do that. Unless they are overtrained

on a specific task, when asked to rate confidence, they often resort to using the detectability of the stimulus as a heuristic (Maniscalco, Peters, and Lau 2016).

For example, Megan Peters and Aurelio Cortese have conducted neuroimaging studies to test this. I was involved in both of their studies, but they were conducted independently using different methods. In both cases, we found that the internal brain signals contributing to confidence ratings were basically just the amount of total detectable signals for the stimuli, rather than the difference between the relevant signals (Cortese et al. 2016; Peters et al. 2017c).

Why do human subjects use such a strategy, which seems highly suboptimal? Essentially, they are neglecting useful information. The exact mechanism remains an open topic for current research (Miyoshi and Lau 2020). But in any case, it points to a strong empirical link between detection and metacognition. Somehow, when asked to give confidence ratings, people resort to heuristics based largely on detectability of the stimuli. Arguably, both metacognition and detection are also conceptually related to subjective experiences. In Chapter 5 we also highlighted metacognition as one of the possible key functions of consciousness, maybe the most plausible one so far, given the limited evidence we have. It would be a nice feature of a theory of consciousness to be able to say something about how they are all tied together.

6.9 Predictive Coding & Generative Adversarial Networks

In recent years, *predictive coding* has become a trendy phrase. This has led some authors to speculate that it may have something to do with consciousness. But sometimes the phrase just refers to the fact that the brain can generate or modulate internal sensory representations in ways not entirely driven by external input (Cao 2020). In this sense, the notion is nothing new. With a few exceptions (Gibson 1968), most modern psychologists agree that perception involves top-down mechanisms of some sort.

More specifically, one idea is that external sensory inputs may be assessed with respect to the endogenously generated expectation. If our expectations are not violated, we learn nothing new; not much is worth signaling. In contrast, "prediction errors" are very much worth signaling downstream, as they carry novel information. But even this more specific notion of predictive coding is actually rather general; some have long considered it a "standard" view on how the brain works. This is not to say there is no controversy around how it works (Aitchison and Lengyel 2017). But there has also been numerous

reports showing that stimuli not consciously perceived can drive this type of predictive process as well, at least to some extent (Iijima and Sakai 2014; Chang et al. 2016; Parras et al. 2017; Meijs et al. 2018; Nourski et al. 2018; Rowe, Tsuchiya, and Garrido 2020). So, it is not clear if "predictive coding" has anything specific to do with consciousness per se.

One interesting fact about predictive coding is that it is not entirely trivial to implement using artificial neural networks. Until recently, most pattern recognition networks adopted a feedforward-only architecture. This is not because computer scientists do not recognize the benefits of predictive coding, of which there are many. The problem is that building a network model to accurately produce top-down generations requires a lot of training time and data.

One solution has been extremely influential within the artificial intelligence community. In generative adversarial networks (GANs), a "generator" could take high level conceptual inputs (e.g., the notion of a "cat") and come up with a corresponding pictorial representation (Goodfellow, Bengio, and Courville 2016). A "discriminator" learns to distinguish between these pictorial outputs from the "generator," and actual images in the world. Both components are easy to build using current technology. In particular, the "discriminator" is akin to a simple pattern classification network. Just like the way a simple facial recognition algorithm can categorize faces as male or female, the "discriminator" makes simple binary decisions on some image inputs to classify input as "real" or "self-generated" (Figure 6.1).

What is interesting is that when we pit the two networks against each other, both networks quickly learn to do their jobs well. That is, we can think of the generator's job is to create "forgeries" (i.e., internally generated pictures that are close enough to the real ones). If it succeeds in fooling the discriminator, it counts as a win. Likewise, the discriminator's "goal" is to catch such forgeries. With these simple and competitive goals set up, both networks can be trained efficiently.

This GANs architecture may be related to consciousness in several ways. First, it has been suggested on theoretical grounds that a discriminator-like mechanism may reside within the prefrontal cortex (Gershman 2019; Lau 2019). In support of this claim, there has been physiological evidence (Mendoza-Halliday and Martinez-Trujillo 2017) showing that neurons in the dorsolateral prefrontal cortex can distinguish between external perceptual content and endogenous generation of the same content (i.e., maintenance of the same information during working-memory delay). This is interesting because holding an image in working memory tends to induce somewhat similar activity in the early visual areas as normal perception (Harrison and

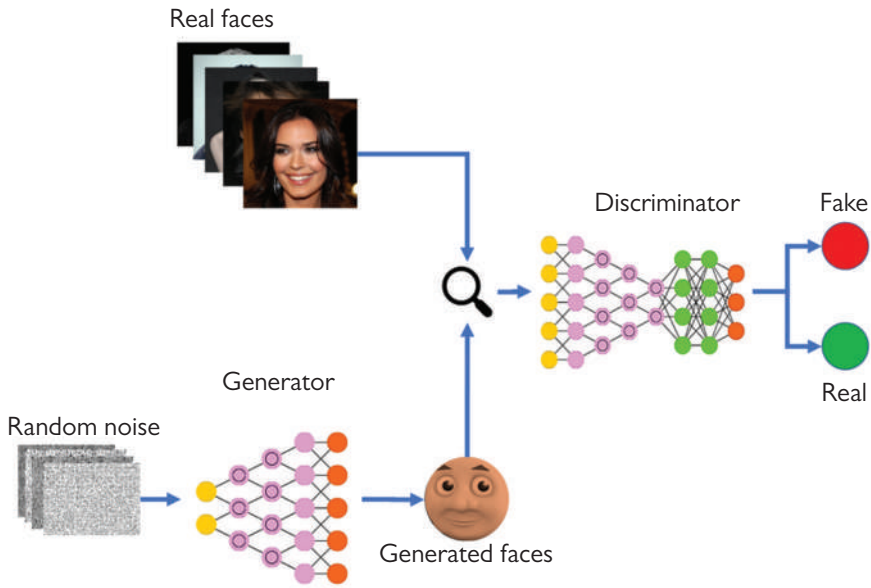

Figure 6.1 GANs architecture

Tong 2009). So, the difference in conscious experience between the two conditions, working memory and perception, may be captured by a discriminator-like mechanism in the prefrontal cortex.

Relatedly, a postdoc in my lab, Taylor Webb, has recently trained a GANs model to perform a simple perceptual task. He found that the discriminator can be "repurposed" to perform metacognitive functions (i.e., generation of confidence ratings) with minimal impact to the discriminator's performance. On the assumption that metacognition and consciousness are linked, perhaps a discriminator-like mechanism can contribute to both too. Although this work is yet to be published, the finding may not be surprising to those familiar to GANs. The discriminator naturally contains rich statistical information about the relevant stimuli, for otherwise it would not be able to do its job.

In Chapter 7, we will put forward a theory describing how human consciousness may critically depend on a discriminator-like mechanism within a GANs-like architecture.

6.10 Interlude: Some Loose Ends

Because this chapter in part summarizes findings discussed in previous chapters, I will not provide a detailed recap here. In brief, we have argued against

both the global and local views. We have outlined what developing a plausible synthesis involves. We need to resist the localist temptation, to fall into extreme positions such as panpsychism. To provide a meaningful mechanistic explanation, we need some form of functionalist account. But we must also avoid equating consciousness with basic cognitive functions such as global control or broadcast. As far as subjective experiences are concerned, such a view is not really compatible with available data.

Instead, we want a view that can account for the relatively subtle activity in the prefrontal cortex during conscious perception, as well as the functional advantages provided—which is far more specific than a globalist may expect. This should in turn elucidate how inflation works, as well as to explain the links between conscious perception, metacognition, and detection. Ideally, this should also take into consideration how the relevant prefrontal mechanisms may relate to current models of predictive coding. Specifically, we should try to take into account the electrophysiological evidence that the prefrontal cortex seems to play a role in discriminating between self-generated and externally triggered perceptual signals.

This agenda is based on my review of the empirical literature so far. Like in any review, I cannot claim that I'm entirely unbiased. All I can say is that my various arguments are based on different evidence and considerations. To the extent that these separate arguments point to a converging conclusion, it is, I hope, not so easy to argue against all of them at once. However, there is indeed one crucial sticking point. My experience in debating with critics is that once this point is accepted, all the counter-arguments will fall like dominoes. But it also means that if this point is challenged, a lot would be at stake.

What is this crucial sticking point? My overall take on the empirical literature depends critically on the notion of task-performance capacity confounders. Performance *capacity* is not performance itself. Even when there is no task, such as in some binocular rivalry experiments, there is this same hidden problem of difference in sheer internal processing signal strength (Chapter 2). Some may say: but this *is* consciousness. My reply is mostly based on the counterexample of blindsight: sometimes performance comes with no corresponding subjective experience.

Understanding this potential dissociation between task-performance capacity and subjective experience may be the cornerstone for the entire thesis here. But some have challenged whether blindsight truly exists (Phillips 2020). There may be subjective experience unreported by the patients, because the experience is so impoverished and unusual. I agree that some "diagnosed" patients may not have "true" blindsight. Even for those who do, it may not be present under all conditions. But for some well-studied patients, for

certain stimuli (e.g., static, low-contrast gratings) very meticulous controls have been performed (Cowey 2010). Given the nature of patient studies, this is in fact one of the better-established phenomena within neuropsychology. Accordingly, Matthias Michel and I have responded to Ian Phillips's recent arguments against the existence of blindsight in detail (Michel and Lau 2021).

The important key point is that blindsight does not need to occur very often. If it happens convincingly on rare occasions, the phenomenon establishes the conceptual and empirical *possibility* of the dissociation between task-performance capacity and subjective experience.

Perhaps Phillips' arguments should be considered within a broader context too, in which he has argued against the possibility of nonconscious perception in general (Peters et al. 2017b). This debate largely hinges on how *perception* is defined. Specifically, our concern here is not whether there can be nonconscious perception that takes place *at the personal level*. If one does not consciously see something, there's perhaps an argument to be made that one does not, as an agent, "perceive" the relevant object, in some meaningful sense. But to deny any kind of nonconscious perceptual *process* would entail that all perceptually relevant processes in the brain are by definition conscious, which seems highly implausible.

Accordingly, my argument for task-performance capacity confounders rests on just this fact: that some meaningful information processing can take place nonconsciously. This kind of dissociation between subjective experience and information processing is also found in other classic cases in neuropsychology (e.g., amnesia and split-brain patients; LeDoux, Michel, and Lau 2020). So, consciousness just isn't as simple as effective information processing of any kind. With this, we establish the need to control for this confounder. This realization renders most current popular claims about the NCC problematic and forces us to accept a centrist position.

But there is another problem: how do we control for the confounder? I suggest that we can match performance levels in some tasks. But then the question arises as to what tasks are most relevant. For example, in Lau and Passingham (2006), the stimuli were a square and a rhombus (i.e., a square-tilted by 45 degrees). The subjects discriminated between the two. It may seem fair that we match performance levels in this discrimination task. But why not a detection task (of identifying stimulus present or absence)? When task performance in discrimination is matched, are we sure detection performance is also matched? If not, do we not leave open a detection-task-performance confounder?

This is a deep and complicated issue. The processing sensitivity for a stimulus can be assessed in different "dimensions" (e.g., detectability, and

discriminability based on various features). Ideally, we match basic performance in all these different tasks. To the extent that the relationship between sensitivities in these dimensions don't change, matching it on one task is sufficient. But the relationship may not be empirically constant, which can lead to some serious complications. That said, matching performance in one obviously relevant task dimension is in any case far better than no match whatsoever. So, this is perhaps the bottom line: just because a confounder is difficult to deal with doesn't mean that we should give up dealing with it altogether. We do it as best as we can.

A different issue also concerns the NCC. My view may come across as giving the prefrontal cortex too much attention. In the beginning of Chapter 3, I justified this focus. Unlike local theorists, our intention is not to write-off other areas as irrelevant. Many other areas, including subcortical regions, may also be highly important. My point here is only that the prefrontal cortex may be one of the many important pieces of the puzzle.

That said, from here, the theory I will introduce will be grounded in what we have reviewed so far. As such, it may neglect the possible roles played by some other brain regions and circuits, including, for example, the claustrum and the thalamus (including, especially, the pulvinar). Perhaps to build a theory that usefully captures the key factors, some oversimplification is unavoidable. But I do acknowledge this limitation. I hope to improve things in the future.

With this, I have given you most of the key facts (except for a final piece of consideration described in Sections 9.5–9.8, that would help explain the qualitative nature of subjective experiences afforded by mammalian brains). Together they form an overall landscape on which a theory can be built. The readers who are skeptical of others' theoretical speculations may take these facts for their own purposes. In Chapter 7 we will introduce a theory. This view is a "centrist" position because it relates consciousness to broader notions of rational cognition (Chapter 8), and at the same time it also addresses the philosophical problems of "qualia," especially the biopsychic intuitions (Chapter 9).

References

Aitchison L, Lengyel M. With or without you: Predictive coding and Bayesian inference in the brain. *Curr Opin Neurobiol* 2017;**46**:219–227.

Block N. Inverted Earth. *Philos Perspect* 1990;**4**:53–79.

Braun AR, Balkin TJ, Wesensten NJ et al. Dissociated pattern of activity in visual cortices and their projections during human rapid eye movement sleep. *Science* 1998;**279**:91–95.

Bridge H, Harrold S, Holmes EA et al. Vivid visual mental imagery in the absence of the primary visual cortex. *J Neurol* 2012;**259**:1062–1070.

Brown R. The brain and its states. In: S Edelman, T Fekete, N Zach (eds), *Being in Time: Dynamical Models of Phenomenal Experience*. 2012, John Benjamins, 211–230.

Brown R, Lau H, LeDoux JE. Understanding the higher-order approach to consciousness. *Trends Cogn Sci* 2019;**23**:754–768.

Cao R. New labels for old ideas: Predictive processing and the interpretation of neural signals. *Rev Philos Psychol* 2020;**11**:517–546.

Chalmers DJ. *The Conscious Mind: In Search of a Fundamental Theory*. Oxford Paperbacks, 1996.

Chang R, Baria AT, Flounders MW et al. Unconsciously elicited perceptual prior. *Neurosci Conscious* 2016;**2016**. https://doi.org/10.1093/nc/niw008.

Cleeremans A, Achoui D, Beauny A et al. Learning to be conscious. *Trends Cogn Sci* 2020;**24**:112–123.

Cortese A, Amano K, Koizumi A et al. Multivoxel neurofeedback selectively modulates confidence without changing perceptual performance. *Nat Commun* 2016;**7**:13669.

Cowey A. The blindsight saga. *Exp Brain Res* 2010;**200**:3–24.

Dehaene S, Lau H, Kouider S. What is consciousness, and could machines have it? *Science* 2017;**358**:486–492.

Fleming SM. Awareness as inference in a higher-order state space. *Neurosci Conscious* 2020;**2020**:niz020.

Gershman SJ. The generative adversarial brain. Front. Artif. Intel. *Appl* 2019;**2**:18.

Gibson JJ. *The Senses Considered as Perceptual Systems* [With illustrations]. L Carmichael (ed). Boston: Houghton Mifflin, 1968.

Godfrey-Smith P. Mind, matter, and metabolism. *J Philos* 2016;**113**:481–506.

Goodfellow I, Bengio Y, Courville A. *Deep Learning*. MIT Press, 2016.

Harrison SA, Tong F. Decoding reveals the contents of visual working memory in early visual areas. *Nature* 2009;**458**:632–635.

Huang L, Wang L, Shen W et al. A source for awareness-dependent figure-ground segregation in human prefrontal cortex. *Proc Natl Acad Sci U S A* 2020;**117**:30836–30847.

Hupé JM, James AC, Girard P et al. Response modulations by static texture surround in area V1 of the macaque monkey do not depend on feedback connections from V2. *J Neurophysiol* 2001;**85**:146–163.

Iijima K, Sakai KL. Subliminal enhancement of predictive effects during syntactic processing in the left inferior frontal gyrus: An MEG study. *Front Syst Neurosci* 2014;**8**:217.

Lau H. Consciousness, metacognition, & perceptual reality monitoring. 2019. Preprint. https://doi.org/10.31234/osf.io/ckbyf.

Lau H, Brown R. The emperor's new phenomenology? The empirical case for conscious experiences without first-order representations. In: A Pautz, D Stoljar (eds), *Blockheads! Essays on Ned Block's Philosophy of Mind and Consciousness*. MIT Press, 2019, 171.

Lau HC, Passingham RE. Relative blindsight in normal observers and the neural correlate of visual consciousness. *Proc Natl Acad Sci U S A* 2006;**103**:18763–18768.

LeDoux JE, Brown R. A higher-order theory of emotional consciousness. *Proc Natl Acad Sci U S A* 2017;**114**:E2016–E2025.

LeDoux JE, Michel M, Lau H. A little history goes a long way toward understanding why we study consciousness the way we do today. *Proc Natl Acad Sci U S A* 2020;**117**:6976–6984.

Maniscalco B, Lau H. The signal processing architecture underlying subjective reports of sensory awareness. *Neurosci Conscious* 2016;**2016**. https://doi.org/10.1093/nc/niw002.

Maniscalco B, Peters MAK, Lau H. Heuristic use of perceptual evidence leads to dissociation between performance and metacognitive sensitivity. *Atten Percept Psychophys* 2016;**78**:923–937.

Manita S, Suzuki T, Homma C et al. A top-down cortical circuit for accurate sensory perception. *Neuron* 2015;**86**:1304–1316.

Meijs EL, Slagter HA, de Lange FP et al. Dynamic interactions between top–down expectations and conscious awareness. *J Neurosci* 2018;**38**:2318–2327.

Mendoza-Halliday D, Martinez-Trujillo JC. Neuronal population coding of perceived and memorized visual features in the lateral prefrontal cortex. *Nat Commun* 2017;**8**:15471.

Michel M, Lau H. Is blindsight possible under signal detection theory? Comment on Phillips (2021). *Psychol Rev* 2021;**128**:585–591.

Miyoshi K, Lau H. A decision-congruent heuristic gives superior metacognitive sensitivity under realistic variance assumptions. *Psychol Rev* 2020;127:655–671.

Nourski KV, Steinschneider M, Rhone AE et al. Auditory predictive coding across awareness states under anesthesia: An intracranial electrophysiology study. *J Neurosci* 2018;**38**:8441–8452.

Parras GG, Nieto-Diego J, Carbajal GV et al. Neurons along the auditory pathway exhibit a hierarchical organization of prediction error. *Nat Commun* 2017;**8**:2148.

Peters MAK, Fesi J, Amendi N et al. Transcranial magnetic stimulation to visual cortex induces suboptimal introspection. *Cortex* 2017a;**93**:119–132.

Peters MAK, Kentridge RW, Phillips I et al. Does unconscious perception really exist? Continuing the ASSC20 debate. *Neurosci Conscious* 2017b;**2017**:nix015.

Peters MAK, Thesen T, Ko YD et al. Perceptual confidence neglects decision-incongruent evidence in the brain. *Nat Hum Behav* 2017c;**1**. https://doi.org/10.1038/s41562-017-0139.

Phillips I. Blindsight is qualitatively degraded conscious vision. *Psychol Rev* 2020.128;**3**;558–584. https://doi.org/10.1037/rev0000254.

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

Rosenthal, D. Consciousness and mind. Clarendon Press, 2005..

Rowe EG, Tsuchiya N, Garrido MI. Detecting (un)seen change: The neural underpinnings of (un)conscious prediction errors. *Front Syst Neurosci* 2020;**14**:81.

Zeman A, Dewar M, Della Sala S. Lives without imagery: Congenital aphantasia. *Cortex* 2015;**73**:378–380.