# Chapter X. Visual consciousness

There has been major progress in computer vision and machines are becoming quite proficient at multiple visual tasks. Teenagers are not surprised by a phone that can recognize their face and having cameras in your house that can look at you and detect your mood is probably not too far off. We have argued that there have been major strides towards developing machines that can recognize objects using algorithms that are inspired by biological circuits. Now imagine a world where we have machines that can visually interpret the world the way we do. To be more precise, imagine a world where we have machines that can flexibly answer a seemingly infinite number of questions on a given image and that you cannot distinguish the answers from those a human would give. Would we claim that such a machine can see? Would such a machine have visual consciousness? Most people would still answer no to this question. They would probably argue that such a machine is nothing more, and nothing less, than a very sophisticated algorithm capable of extracting a relevant answer from a collection of pixels. They would point out that humans are different, that humans can have *feelings* about the image, that humans can laugh at the image, or be scared by its contents, that humans have a sense of *qualia*. Qualia is an intriguing term introduced by philosophers; the dictionary defines it as "… the internal and subjective component of sense perceptions, arising from stimulation of the senses by phenomena". The definition does not seem to be particularly helpful to help us discern whether our extraordinary visual machine has consciousness or not.

Maybe it is time to go back into the brain. We have accompanied and witnessed the adventures of information processing along the ventral visual stream, starting with photons impinging on the retina all the way to the remarkable responses of neurons in inferior temporal cortex. Throughout this cascade of processes, we found neurons with increasing degree of similarity to our recognition capabilities. Along the way, we have perhaps forgotten about a major aspect of our visual experience, namely, the subjective feeling of seeing and experiencing the visual world. How does neuronal activity give rise to conscious experience? What are the biological mechanisms responsible for qualia?

The question of subjective awareness in the context of visual perception is part of the grander theme of consciousness. The age-old question of how a physical system can give rise to consciousness has been debated by philosophers, clinicians and scientists for millennia. Over the last decade, there has been increased interest in using modern Neuroscience techniques to further our understanding of the circuits and mechanisms by which neurons may represent and distinguish conscious content (Crick, 1994; Koch, 2005).

## 10.1. A non-exhaustive list of possible answers

It makes sense to assume that individual atoms do not possess or give rise to qualia. Connecting Physical realism to the world of experience is perhaps one of the hardest questions of all time. Multiple answers have been proposed over the years in an attempt to explain how a physical system can give rise to consciousness. I will not have time to do justice or discuss them in detail here. Instead, I would like to group them and list some of the main answers that scholars have proposed.

(1) “Religious” answers. These are non-scientific explanations that often invoke the need for a soul, a homunculus, or some form of communication between physical systems and other non-physical entities. Several variants of these explanations abound including passages in the Bible, the writings of Plato, Aristotle, Thomas Aquinas, Karl Popper, Sigmund Freud and even top-notch scientists such as John Eccles.

(2) The “mysterian” approach. Proponents of this approach argue that science simply cannot understand consciousness. There are several variations of this idea including statements such as “a system cannot understand itself”, or “the answer is just too complex for our simple brains to grasp”. This defeatist approach does not seem to be particularly useful. In the absence of any compelling proof that science cannot solve the problem, it seems better to try and fail rather than not try at all.

(3) Consciousness as an illusion. Some philosophers have argued that there is no real phenomenon such as consciousness. The feeling of consciousness is just an illusion (Dennett, 1991). But what an extraordinary illusion it is! We have made extraordinary progress understanding the neural basis for multiple illusions. For example, when we perceive illusory contours, we know that there is no magic, there are actual neurons that respond vigorously to those contours and explicitly represent the lines that we see (von der Heydt et al., 1984). It would be particularly exciting to be able to provide a similar mechanistic explanation for the neural basis of conscious sensations.

(4) Consciousness as an epiphenomenon. A related version of consciousness as an illusion is the notion that consciousness is an epiphenomenon. As soon as multiple neurons and complex networks are connected, the feeling of consciousness arises but it does not serve any purpose (in the same way that a computer may heat up but this heat does not really serve any computational purpose).

(5) Consciousness and new laws of Physics. Others (e.g. Roger Penrose) argue that we need new (as yet undiscovered) laws of Physics to explain consciousness.

In stark contrast with the above approaches, several neuroscientists have become interested in the arguably simpler notion that consciousness arises from the specific function of neuronal circuits. Which circuits, when and how remains to be determined through scientific investigation without invoking new laws of physics, and without invoking souls. We assume that consciousness can and should be explained in neurobiological terms, and that there is no limit to our capability towards arriving at the answer. We still do not understand many aspects of brain function (e.g., we do not understand what changes in neural circuits give rise to Autism), but that does not mean that we need to invoke the explanations above for all the brain phenomenology that we still cannot grasp.

The neuroscientific approach to studying consciousness involves several working assumptions:
(1) We are conscious. Consciousness is not an epiphenomenon. Therefore, consciousness deserves an explanation like any other aspect of brain function.
(2) Other animals are also conscious. This assumption enables us to probe for consciousness in animal models. It seems too early to draw the line and unequivocally dictate which animals do show consciousness and which ones do not.
(3) We start with simple questions that we can try to study rigorously. We start with vision. Hopefully, we will be able to extrapolate some of what we learn from vision to other sensations (e.g. pain, smell, self-awareness)
(4) We need an explicit representation. Only parts of the brain will correlate with the contents of consciousness. We search the *neuronal correlates of consciousness* (NCC).

The strategic decision to start by investigating a rather reduced domain, the neuronal correlates underlying visual awareness, clearly leaves many fascinating topics out. Some of these topics include dreams, lucid dreaming, out of body experiences, hallucinations, meditation, sleep walking, hypnosis, self awareness, the so-called notion of qualia and feelings. This does not imply that these are not interesting and relevant topics; it merely reflects a strategic decision of how to approach a difficult scientific question.

## 10.2. The search for the NCC, the neuronal correlates of consciousness

The NCC (neuronal correlates of consciousness) is defined as a *minimal* set of neuronal events and mechanisms that are jointly *sufficient* for a *specific conscious percept* (Crick and Koch, 1990, 2003).

It is critical to define some of these terms. The NCC is defined as a *minimal* set. A solution such as “the whole healthy human brain can experience consciousness” is not very informative. The neural mechanism should be “sufficient” (not just necessary) to represent a conscious percept. This clause leaves out so-called “enabling” factors such as the heart or the cholinergic systems arising in the brainstem. We are seeking for the correlates for “specific conscious percepts" such as seeing a face (as opposed to generic aspects such as being conscious/unconscious).

It is quite clear that not all brain activity is directly linked to conscious perception at any given point. To clarify, this does not mean that those brain processes are not important or interesting. For example, significant resources and neurons are devoted to controlling breathing, posture, walking, etc. With some exceptions, most of the time we are not aware of such processes.

A particularly striking documentation of relatively sophisticated brain processing that does not reach awareness is given by a patient studied by Goodale and Miller (Goodale and Milner, 1992). This patient had severe damage along the ventral visual stream and the dorsal stream was relatively unimpaired. The patient could not recognize shapes but could still act on those shapes with relatively sophisticated precision. For example, the patient could not report the orientation of a slit but could place an envelope in the slit rather accurately. The search for the NCC concerns investigating which neuronal processes and mechanisms correlate with conscious content and which ones do not.

## 10.3. In search of an explicit representation

Upon seeing an object, neurons in the retinae are activated. In fact, stimulating each of the photoreceptors in the same pattern and magnitude evoked by a given object should elicit the object's percept. Does this imply that the retinal photoreceptors constitute the desired NCC? Not quite. Those neurons in the retina activate neurons in the LGN, which in turn activate primary visual cortex, which in turn transmit the information to higher areas within ventral visual cortex. Several lines of evidence suggest that the activity in early visual areas from the retina to primary visual cortex cannot be the locus of the NCC (Crick and Koch, 1995). One striking example is what happens when you are watching TV. The TV has a certain refresh rate, that is, it shows a number of frames per second, say 60 frames per second. Retinal ganglion cells and neurons in primary visual cortex fire vigorously following those rapid changes in the visual input. Yet, our perception is essentially oblivious to what is happening there. A critical aspect of the NCC is that the representation of visual information must be "explicit". If neurons are representing information that we are not aware of, then those neurons cannot be quite part of the NCC, in the same way that there are neurons that control how you walk, yet you are typically not aware of their activity.

But what exactly is an *explicit representation* and how would we ever know if we have one? After all, information from the retinal ganglion cells is obviously required for vision. What makes their representation implicit and not explicit? One way to define an explicit representation is that it should be possible to decode the information via a simple linear classifier. If our perception indicates that we are seeing a chair, the chair is represented by the activity of retinal ganglion cells, but we cannot linearly read out the presence or absence of a chair from the retina. Similarly, a computer may hold a representation of the information for the chair in a digital photograph. However, as we have discussed in the previous chapters, decoding such information requires a cascade of multiple computations. Information about objects is not explicitly represented in a digital photograph. Similarly, the retina does not hold an explicit representation of objects.

Several visual illusions acutely point out the need for explicit representations. Consider the Kanizsa triangle illustrated in Chapter 1. We perceive strong edges defining the triangle even in parts of the image where there is no visual information (i.e. there is no real edge). Such a perception of an edge implies that there should be neurons that represent that subjective edge. Neurons in the retina do not respond to such illusory contours.

## 10.4. An experimental approach to study visual consciousness

The discussion of visual illusions suggests a promising path to investigate the neuronal correlates of visual consciousness by investigating which neuronal processes coincide with subjective perception. A particular type of visual illusion that has been quite fruitful in this regard involves the use of bistable percepts. A famous example of a bistable percept is the Necker cube. The same visual input can be seen in two different configurations. In the case of the Necker cube, it is possible to voluntarily switch between the two possible interpretations of the same input.

Such volitional control is not possible in the case of binocular rivalry. Under normal circumstances, the information that the right and left eyes convey is highly correlated[1]. What would happen if you show two completely different stimuli to the right and left eyes? Under these conditions, we perceive either one or the other stimulus in a seemingly random fashion, a phenomenon called binocular rivalry (Blake and Logothetis, 2002). Extensive psychophysical investigation has provided a wealth of information about the conditions that lead to perceptual dominance of visual stimuli, what can or cannot be done with the information that is being suppressed and the dynamics underlying perceptual alterations (Blake, 1989; Blake and Logothetis, 2002). What is particularly interesting about this phenomenon is that, to a reasonably good first approximation, the visual input is constant and yet subjective perception alternates between two possible interpretations of the visual world. Investigators then ask: what are the neuronal changes that correlate with these subjective transitions? Several studies have shown that only a small fraction of neurons in early visual areas follow the subjective changes whereas most neurons in higher visual areas are strongly modulated by the immediate contents of visual awareness (Leopold and Logothetis, 1999).

[1] It is not identical, though. The small differences between the input from the right eye and left eye provide strong cues to obtain 3D information. 3D movies specifically

This type of experiments may pave the road to an initial understanding of certain circuits and neuronal activity changes that correlate with subjective perception. What would constitute evidence of understanding the NCC? We argue that four conditions should be met. (1) We should be able to model and predict neuronal responses given a perceptual state. (2) Conversely, we should be able to predict perceptual states from neuronal responses. (3) We should be able to elicit a percept by activating the corresponding neuronal patterns (e.g. via electrical stimulation). (4) We should be able to inactivate or repress a perceptual state by modifying the neuronal activity patterns. There still seems to be a long way to understand the neuronal correlates of visual consciousness by meeting these four conditions. Yet, nowadays, these questions and themes have become a major area of research and we may be surprised to observe major progress in the field in the years to come.

## 10.5. Integrated information theory

The last decade has seen the development of an elegant theoretical framework that deserves discussion, the integrated information theory (IIT), by Giulio Tononi. In an oversimplified form, the basic intuition behind IIT is that conscious experience represents information and that this representation is unique. According to IIT, a dynamical system of interconnected parts is characterized by a metric, connoted by $\Phi$, which has a lower value when the system can be described by smaller relatively independent subsystems. The larger $\Phi$, the more integrated information the system has. The theory postulates that conscious experience is proportional to $\Phi$ *(Tononi, 2005; Seth et al., 2011; Tegmark, 2014; Tononi and Koch, 2015; Tononi et al., 2016)*. The definition of $\Phi$ comprises two steps: (i) perform an imaginary partition of the system and compute $\phi$, a measure of how much the two parts affect each other (i.e., how well we can predict the evolution of the system based on the conditional transition probabilities); and (ii) define $\Phi$ as the "cruelest" such partition that minimizes $\phi$. Elegantly, the theory provides specific mathematical definitions to calculate these quantities *(Tegmark, 2015; Tegmark, 2016; Tononi et al., 2016)*. The definitions of $\Phi$ by Tononi's group *(Tononi et al., 2016)* and variations by others such as Barrett and Seth *(Seth et al., 2011)* can all be incorporated in this general formalism.

A major challenge in testing the IIT framework has been that, for real systems, the above equations are prohibitively challenging to compute. For a given partition, the computational time grows exponentially with the size of the system. However, Tegmark recently developed an approximation to calculate $\Phi$ using graph theory (Tegmark, 2016), bringing the calculations to a polynomial dependency on the system size, and making this algorithm readily applicable to the large scale of physiological recordings in this proposal.

The theory is notably elegant, starting from axioms and proposing concrete quantitative definitions, which sets it apart from other discussions about consciousness, which are merely qualitative. At the same time, the theory suggests many counterintuitive predictions. Any object, even your cellular phone, has a certain $\Phi$ value. One may expect that inanimate objects or plants should have $\Phi=0$, but this is not what the theory states. Those objects may have low values of $\Phi$, but not zero. Additionally, it is in principle possible to create artificial systems with high $\Phi$ values, yet it seems unlikely that such systems would show consciousness. Ultimately, it will be quite interesting to test the theory empirically.

## 10.6. References

Blake R (1989) A neural theory of binocular rivalry. Psychological Review 96:145-167.

Blake R, Logothetis N (2002) Visual competition. Nature Reviews Neuroscience 3:13-21.

Crick F (1994) The astonishing hypothesis. New York: Simon & Schuster.

Crick F, Koch C (1990) Towards a neruobiological theory of conciousness. Seminars in the Neurosciences 2:263-275.

Crick F, Koch C (1995) Are we aware of neural activity in primary visual cortex ? Nature 375:121-123.

Crick F, Koch C (2003) A framework for consciousness. Nature neuroscience 6:119-126.

Dennett D (1991) Consciousness explained. Boston: Little and Brown.

Goodale M, Milner A (1992) Separate visual pathways for perception and action. TRENDS IN NEUROSCIENCES 15:20-25.

Koch C (2005) The quest for consciousness, 1 Edition. Los Angeles: Roberts & Company Publishers.

Leopold DA, Logothetis NK (1999) Multistable phenomena: changing views in perception. Trends in Cognitive Sciences 3:254-264.

Seth AK, Barrett AB, Barnett L (2011) Causal density and integrated information as measures of conscious level. Philos Trans A Math Phys Eng Sci 369:3748-3767.

Tegmark M (2014) Conciousness as a state of matter. Arxiv 14011219.

Tegmark M (2015) Improved measures of integrated information

Tegmark M (2016) Improved Measures of Integrated Information. PLoS Comput Biol 12:e1005123.

Tononi G (2005) Consciousness, information integration, and the brain. Prog Brain Res 150:109-126.

Tononi G, Koch C (2015) Consciousness: here, there and everywhere? Philos Trans R Soc Lond B Biol Sci 370.

Tononi G, Boly M, Massimini M, Koch C (2016) Integrated information theory: from consciousness to its physical substrate. Nat Rev Neurosci 17:450-461.

von der Heydt R, Peterhans E, Baumgartner G (1984) Illusory contours and cortical neuron responses. Science 224:1260-1262.