Chapter X. Visual consciousness

Supplementary contents at **http://bit.ly/2FHXycS**

As discussed in the last two chapters, there has been significant progress in computer vision. Machines are becoming quite proficient at a wide variety of visual tasks. Teenagers are not surprised by a phone that can recognize their faces. Self-driving cars are a matter of daily real-world discussions. Having cameras in the house that can detect a person's mood is probably not too far off. 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. Let us assume that we cannot distinguish the answers given by the machine from the answers that a human would give, that is, assume that machines can pass the Turing test for vision, as defined in **Chapter IX**. Would we claim that such a machine can *see*? Would such a machine have *visual consciousness*?

Most laypeople would still answer "no" to this question. They would 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 claim that machines can beat the world champion in Chess or Go, but they do not "understand" the game. They would point out that humans are different, humans can *experience* the image, have *feelings* about the image, can laugh at the image, or be scared by its contents; the image evokes sensations and specific quality. Humans have a sense of *qualia* about the image.

Qualia is an intriguing term introduced by philosophers; the dictionary defines qualia as "… the internal and subjective component of sense perceptions, arising from stimulation of the senses by phenomena." This definition does not seem to be particularly helpful in discerning whether our extraordinary visual machine, which can pass the Turing test for vision, does or does not have consciousness. Nevertheless, this vague definition will have to suffice for now, until we have better ones that are directly based on a rigorous understanding of how qualia can be mapped to neuronal circuit function. The Turing test is defined strictly in terms of questions and answers, that is, in terms of behavior. Such observable behaviors do not necessarily reflect what humans or machines experience when exposed to a given image. It would be useful to have an operational definition, with a Turing test analogous to the one introduced in the previous chapter, for visual consciousness. Having such a Turing test may help us discern whether a machine can display consciousness or not, and can also help define which animal species are conscious.

To make progress towards a definition of consciousness and qualia, 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 that respond when illumination changes in specific locations within the visual field, we marveled at neurons that are selectively activated by different types of shapes, we discussed how tolerant neurons are for changes in the stimulus properties, we were intrigued by neurons that can respond to imagined things that do not directly reflect what is in the outside world such as illusory contours, we discovered neurons that respond in the absence of a visual stimulus in a correlate of the mysterious process of visual imagery. Ascending through the visual hierarchy, there is an increasing degree of similarity between neuronal response properties and behavioral recognition capabilities. Along the way, we have perhaps forgotten about a profound aspect of our visual experience, namely, the subjective feeling of seeing and experiencing the visual world. How does neuronal activity give rise to those subjective feelings? What are the biological mechanisms responsible for qualia?

Coming up with concrete definitions in the arena of consciousness might be a bit premature. Several investigators have attempted to draw distinctions between consciousness, awareness, qualia, and subjective percepts. For example, the philosopher David Chalmers has proposed to reserve the term awareness to denote the reportable and accessible contents of consciousness while the other terms are linked to direct experience irrespective of reports. Here I will use all of these terms indistinguishably. Likely, mixing these terms is not a wise idea, and future work will help us sharpen our understanding of the nuances of conscious perception. For the moment, rather than attempting a precise definition, we will examine concrete experiments that aim to elucidate the biological mechanisms that correlate with conscious perception. Within the context of those experiments, the questions are well defined by mapping percepts to behavioral reports. There are also “no-report” parallels of those experiments where we imagine that the percepts are identical except for the behavioral motor outputs.

The question of subjective awareness in the context of visual perception is part of the grander theme of consciousness. Visual consciousness is but one example of the type of sensations that our brain has to represent. Visual consciousness may be particularly dominant with respect to other sensations for primate species, but there are still other aspects of conscious experience that do not depend on vision. Other sensations include auditory consciousness, the feeling of pain, love, volition, and hunger. 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 two decades, there has been increased interest in using modern Neuroscience techniques to further our understanding of the circuits and mechanisms by which neurons represent and

distinguish conscious content. Here we focus on those experiments and theories within the framework of visual processing.

X.1. A non-exhaustive list of possible answers

A mechanistic explanation of visual consciousness should ultimately be expressed in terms of the fundamental physical structures that support qualia, that is, neurons and their interactions. However, it is perplexing to imagine how physical systems can have subjective awareness. 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. There does not seem to be any chapter in our Physics textbooks for anything closely resembling consciousness. Physics textbooks do not have a chapter about Genetics either, but we can trace a path from atoms to molecules, to the rich chemistry of carbon molecules, to the structure of DNA, and onto Genetics. We lack even a sketch of such a path in the case of consciousness.

Multiple answers have been proposed over the years in an attempt to explain how a physical system can give rise to consciousness. We will not be able to do justice or discuss all of those proposals in detail here. Instead, we coarsely classify those ideas and list some of the main answers that scholars have proposed through the ages.

(1) “Religious”, “dualistic”, and “non-physical” answers. These are non-scientific explanations that often invoke the need for a soul, a homunculus, an engine, or some form of communication between physical systems and other non-physical entities. Often, a distinction is made between the brain, a physical substrate, and the “mind,” an ethereal concept that may or may not connect with the brain, depending on whom you ask. Several variants of these explanations abound, including passages in the Bible, the writings of Plato, Aristotle, Thomas Aquinas, Rene Descartes, Karl Popper, Sigmund Freud, and even top-notch neuroscientists such as John Eccles. For simplicity, I am taking the liberty of lumping every form of dualism into the same cluster, which I refer to as “religious” / “dualistic” answers. However, it should be noted that there are important differences among these different thinkers; certainly, not all of them embraced dualism because of religious reasons. To make matters more complicated, some religious people do *not* support dualism. I am merely pointing out that any explanation that is not based on Physics, and by extension on brain science, necessitates some extra “magic juice.” This magic juice has been called a soul, a mind, or a homunculus.

The dualism between the brain and the “mind” pervades our vocabulary. We speak of “minding the gap,” “keeping an open mind,” or “changing your mind.” Furthermore, even top-notch neuroscientists who do not necessarily embrace dualism still use strange dualistic descriptions, as in “… the brain knows our decisions before we do …” or “Our brain doesn’t tell us everything it knows. And

sometimes it goes further and actively mislead us." It is hard to eradicate the long and dark shadow of a Cartesian dichotomy between the mind and the brain.

(2) The "mysterian" answer. Proponents of this idea, including giants of the caliber of Thomas Nagel, Frank Jackson, and David Chalmers, argue that science simply cannot fully explain 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," or "science relies on objective measurements and consciousness requires a subjective aspect." 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. Even more problematic is the fact that this answer is not easily falsifiable without first solving the problem of consciousness, thereby making it a circular proposition.

(3) Consciousness as an illusion. Some philosophers like Daniel Dennett have argued that there is no such thing as consciousness. Therefore, there is nothing that warrants an explanation in terms of brain circuits. According to this view, consciousness is not a real phenomenon; the feeling of consciousness is just an illusion. But what an extraordinary illusion it is! We have made extraordinary progress in understanding the neural basis for multiple visual 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 (**Chapter V**). We even have computational models that suggest how the neuronal responses to illusory contours may come about through the integration via horizontal connections of signals from other neurons responding to real contours. It would be particularly exciting to be able to provide a similar mechanistic explanation for the neural basis of conscious sensations, regardless of whether these sensations are called illusions or not.

(4) Consciousness as an epiphenomenon. A related version of consciousness as an illusion is the notion that consciousness is an epiphenomenon. This proposal maintains that consciousness has no causal power, that is, that consciousness cannot cause any changes in the physical state of the system. As soon as multiple neurons and complex networks are connected, the feeling of consciousness arises. According to this viewpoint, this feeling does not serve any purpose. An analogy that is often used to illustrate this proposal is the following: a computer may heat up while it is doing its job, but this heat does not serve any purpose in and of itself; it is merely a side consequence of the machinery used to perform the actual computations. However, in this case, we also understand quite well where this heat comes from in terms of physical laws. It would be equally exciting to provide a mechanistic explanation for the neural basis of the "conscious heat," regardless of whether it serves a purpose or not.

(5) Consciousness and new laws of Physics. Others like the brilliant mathematician and physicist Roger Penrose argue that we need new, as yet

undiscovered, laws of Physics to explain consciousness. The argument is that current laws are insufficient in some way. This proposition may very well end up being true. However, at least historically, new laws have been discovered by trying to describe experimental results with existing laws and failing to do so. Even better is actually showing that existing laws lead to wrong predictions that are inconsistent with empirical findings. Stating *a priori* that new laws are necessary seems to skip an essential step in scientific inquiry. There are interesting philosophical and practical questions about when enough evidence accumulates to suggest that the current paradigm is wrong. The field has been thinking about how to explain consciousness based on the activity of neural circuits for about two decades now; this does not seem to be enough time to declare that current laws of Physics fail to explain the phenomenon. Penrose and others might be right, but we respectfully ask them to give us more time to try to solve the problem using nothing more and nothing less than the powerful artillery of current Physics.

In stark contrast with the above approaches, several neuroscientists have become interested in the arguably more straightforward notion that consciousness arises from specific interactions within neuronal circuits that are defined by known neurobiological principles. Consciousness is a real observation intrinsic to an organism; like any other observation, consciousness deserves a mechanistic explanation. There is no need to invoke magic juice, or to impose new laws of Physics. Consciousness might well be considered to be an illusion in the sense that all of our percepts are constructs fabricated by the brain. Moreover, it seems premature to question whether consciousness has causals power or not, given that we are still taking the first preliminary steps towards defining consciousness in terms of brain science principles. According to this framework, we already have the key ingredients towards explaining consciousness. Which circuits, when, and how neuronal activity orchestrates consciousness, remains to be determined through scientific investigation without invoking new laws or non-physical engines. We assume that consciousness can and should be explained in neurobiological terms, and that there is no limit to our capability of arriving at the answer. We still do not understand many aspects of brain function. In fact, I would argue that we still do not understand *most* aspects of brain function. If I had to guess and place the history of Neuroscience in comparison with the history of research in Physics, I would argue that Neuroscience is still in a pre-Newtonian state. However, this delightful level of ignorance does not imply that we should give up and invoke the explanations above for all the observations related to brain function that we still cannot grasp.

The neuroscientific approach to studying consciousness involves several working assumptions:

(1) We *are* conscious. Consciousness is not an epiphenomenon. There is a sensation produced by visual inputs, which is reliable, reproducible, and even mostly universal across humans. Therefore, consciousness deserves an

explanation like any other empirical observation, like the tides, the position of the moon, the firing patterns of retinal ganglion cells, or the perception of illusory contours.

(2) Other animals are also conscious. This assumption enables us to probe consciousness in non-human animals. It seems too early to draw the line and unequivocally dictate which animals do show consciousness and which ones do not. It seems prudent to assume that bacteria do not have any form of visual consciousness, even those that can capture light to perform photosynthesis. Beyond bacteria, it is hard to tell for other species. Plants can also capture light and perform photosynthesis, along with many other exciting processes; however, the working assumption of an explanation based on neural circuits would also rule them out of the consciousness discussion. Once we understand the neuronal mechanisms that constitute consciousness, we might figure out that some species, say the fruit fly as an example, may show all the ingredients required for visual consciousness. Alternatively, we may come to understand that the fruit fly's visually triggered behaviors are purely automatic reflexes that involve no conscious sensation at all. Right now, it is too early to tell, and we should keep our brains open (not our minds open because that would be dualistic!), and we should be willing to be surprised by the scientific answers.

(3) We focus in this chapter on visual consciousness. There are several advantages to studying visual consciousness: we know more about the neuroanatomy and neurophysiology of the visual system than about other domains (**Chapters II**, **V**, **VI**), we have image-computable models (**Chapters VII**, **VIII**), and we can rigorously control stimulus timing and content while measuring behavior (**Chapter III**). Other investigators have begun studying consciousness in other domains outside of visual processing as well. We expect that we will be able to generalize what we learn from vision to other sensations (e.g., pain, smell, self-awareness). The study of visual computations in the brain has inspired progress in many other domains of Neuroscience, including other sensory modalities, but also research on learning, memories, decision-making, and other processes. Therefore, we hope that once we make progress towards elucidating the neuronal mechanisms that represent visual consciousness, the results might transfer to other aspects of consciousness as well.

The focus on visual consciousness leaves out many fascinating aspects of consciousness. Some of these topics include dreams, lucid dreaming, out of body experiences, hallucinations, meditation, sleepwalking, hypnosis, the notion of qualia, and feelings. We do not mean to imply these are uninteresting or irrelevant topics. Many courageous scientists are investigating some of these other aspects of consciousness as well.

(4) We need an explicit and mechanistic representation. Only a restricted set of brain parts will correlate with the contents of consciousness. It is not sufficient to state that consciousness is in the brain. We would like to have quantitative models of visual consciousness, similar in spirit and perhaps even similar in

format and architecture, to the types of models discussed in **Chapters VII-IX**. We hope that these models will enable us to predict how conscious sensations impact neuronal activity and to read out conscious perception from neuronal activity.

X.2. The search for the NCC, the neuronal correlates of consciousness

The NCC (neuronal correlates of consciousness, ***Figure X-1***) is defined as a *minimal* set of neuronal events and mechanisms that are jointly *sufficient* for a *specific conscious percept*. 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 mechanisms 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 the correlates for the specific content of conscious percepts such as seeing a face, as opposed to generic aspects such as being conscious/unconscious.

INSERT Figure X-1 AROUND HERE
Figure X-1. The neural correlates of consciousness (NCC).
Any percept must be associated with a minimal and explicit representation. For example, if we were to record the activity of neurons that have a receptive field located at the intersections of the squares in this famous illusion, we would expect the NCC to be active if and only if the subject perceives a black spot at that intersection at any given time.

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

A particularly striking documentation of sophisticated brain processing that does not reach awareness is given by a patient studied by Melvyn Goodale and David Milner, described in **Chapter IV**. This patient had severe damage along the ventral visual stream, while the dorsal stream was relatively unimpaired. The patient could not recognize shapes and had no awareness about 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 elucidating which neuronal processes correlate with conscious content and which ones do not.

X.3. The representation of conscious content must be explicit

Upon seeing an object, neurons in the retinae are activated (**Chapter II**). In fact, stimulating each of the retina's photoreceptors in precisely the same

pattern and magnitude evoked by a given object should elicit a percept of that object. 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 neurons in 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 is unlikely to be the locus of the NCC. One striking example is what happens when we watch TV. The TV monitor has a certain refresh rate, that is, it shows multiple frames per second, say 60 frames per second. Retinal ganglion cells and neurons in primary visual cortex fire vigorously because of those rapid changes in the visual input, following the screen refresh rate, transiently increasing the firing rate in response to every flash of a new frame. However, our perception is virtually oblivious to this frame-by-frame information; we perceive continuous motion without any flickering unless the refresh rate is very low. In other words, there are RGC responses that do not reach conscious perception. Conversely, the contents of perception may include signals that are not directly reflected by RGCs. A striking example is the blind spot (**Chapter II**). Covering one eye, there is a region of the visual field for which there are simply no photoreceptors in the eye. However, we do not see an empty or black scotoma in that region. Brains fill in the scene despite the absence of information coming from the retina in the blind spot. We are also rarely aware of blinks, even though the whole world becomes dark momentarily for the RGCs.

A critical aspect of the NCC is that the representation of visual information must be "explicit." If there are neurons representing information that we are not aware of at a given time, then those neurons cannot be part of the NCC at that moment in time. As noted earlier, some neurons control our breathing and how we walk, yet we are typically not aware of their activity. In the same fashion, our percepts do not directly correlate with neuronal activity in the retina.

What exactly is an *explicit representation,* and how would we ever know if we find one? After all, information from RGCs is obviously required for vision. What makes their representation implicit as opposed to explicit? One way to define an explicit representation is that it should be possible to decode the information via a one-layer neural network (**Chapters VI-VII**). In the simplest case, a perceptron should be able to decode the information: if we have a population of neurons with activities $x_1, x_2, \ldots, x_n$, then the perceptron classifier can be expressed as $g(w_1 x_1 + w_2 x_2 + \ldots + w_n x_n)$ where g is a non-linear function like a threshold. An explicit representation may still depend on joint activity within a population of neurons, the emphasis being on whether it is readily decodable, as opposed to the type of implicit information as present in the retina.

If we see a chair, then that chair is represented by the activity of RGCs, but we cannot read out the presence or absence of a chair from the retina using

a single-layer network. Analogously, 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 the pixels of the digital photograph. Similarly, the retina does not hold an explicit representation of our percepts.

An explicit representation of the visual perception contents at any given time should not follow the refresh rate of the monitor, should be able to fill in the missing information in the blind spot, and should be subject to visual illusions in the same way that perception dictates. For example, consider the Kanizsa triangle (**Chapter I**): the perception of an edge when there is none suggests that there should be neurons that represent that subjective edge. Neurons in the retina do not respond to such illusory contours, but neurons in cortical area V2 do (**Chapter V**).

X.4. Experimental approaches to study visual consciousness

The Kanizsa triangle example and other visual illusions suggest a promising path to investigate the neuronal correlates of visual consciousness by determining which neuronal processes coincide with subjective perception. A particularly fruitful experimental approach has been to focus on situations where the same visual stimulus can lead to visual awareness only sometimes, but not always (***Figure X-2***).

One example is to consider perception near discrimination thresholds. For example, a stimulus may be rendered hard to detect by decreasing its contrast. If the contrast is high enough, then subjects can detect the stimulus most of the time. If the contrast is too low, then subjects fail to detect the stimulus most of the time. There is an intermediate regime near threshold where subjects can sometimes see the stimulus, and other times they cannot, as assessed by behavioral measurements. The same physical stimulus sometimes leads to perception, but sometimes it does not. Let us assume that we can ensure that we are presenting the exact same stimulus, also that the eyes are fixating on the same location, and that there are no other changes. Under these conditions, it seems safe to assume that the neuronal responses in the retina would be similar in those trials when the stimulus is perceived and when it is not. However, something must change somewhere in the brain to lead subjects to report that they see the stimulus in some trials. We can investigate where, when, and how neuronal responses along visual cortex correlate with the subjective percept.

A similar situation can be reached in backward masking experiments where a stimulus is flashed for a brief amount of time, followed by a rapid noise mask (***Figure X-2D, Section III-6***). If the duration is too long, then subjects can easily see the stimulus. If the duration is too short, then subjects never see the stimulus. There is an intermediate regime, with durations on the order of 25

milliseconds, where subjects report seeing the stimuli only in some but not all trials.

INSERT Figure X-1 AROUND HERE

Figure X-2. Example tasks used to probe the NCC.

***A**. Motion induced blindness. When the blue dots move about, the yellow circles intermittently disappear from perception. **B**. Mooney images. It is generally difficult to interpret the images in the top row. Exposure to the grayscale counterparts (bottom row) immediately renders the Mooney images interpretable. **C**. During visual search, subjects will often fixate on the target object and continue searching without realizing it. **D**. Backward masking can render a stimulus invisible.*

Another example is the interpretation of images that are hard to recognize, like Mooney images. These images are black and white impoverished renderings that are difficult to interpret at first glance. A famous example is the Dalmatian dog illusion. A few more examples are shown in ***Figure X-2B***. Consider the example on the top left in ***Figure X-2B***. At first glance, the image appears to contain multiple arbitrarily shaped black spots randomly scattered throughout. Yet, the image contains a rhino in a natural scene. If someone traces the rhino's contour, or upon observing the grayscale counterpart to this image (bottom left in ***Figure X-2B***), observers can readily recognize the animal and also interpret the rest of the scene. The same image, and assuming the same fixation location, can lead to interpreting it as noise or a rhino. We conjecture that the neural representation of the image at the level of the retina would be indistinguishable between the noise and rhino interpretations. However, there must be a representation of the rhino, perhaps in inferior temporal cortex neurons (**Chapter VI**), and this representation should be activated if and only if the observer can correctly interpret the image.

A daily example takes place during visual search. Imagine that we are looking for our car keys on top of a cluttered desk, or looking for Waldo in ***Figure X-2C***. The eyes scan the desk for several seconds through multiple saccades. Sometimes we will directly fixate on the car keys, yet we will not be aware that our eyes landed on the keys, and will continue searching. Eventually, our eyes fixate on the keys, *and* we become aware that you found them. Here is a case of two fixations, let us assume for the sake of simplicity in the same location, with the same visual stimulus, one with and one without awareness.

A similar situation arises during the phenomena of *inattentional blindness* and *change blindness*. During inattentional blindness, observers fail to notice a fully visible object, presumably because attention is engaged elsewhere. A notable demonstration of this phenomenon is the well-known video where there are two teams, a black and a white team, passing around two basketballs. Subjects are asked to count the number of passes between members of one team. Unbeknown to the subjects (and I apologize beforehand if I am spoiling the effect for the reader), a man disguised as a gorilla slowly walks through the middle of the scene. Remarkably, about half of the subjects utterly fail to notice the gorilla.

Without a doubt, the information about the gorilla reaches the retinal ganglion cells, and probably also up to primary visual cortex, maybe even higher areas within visual cortex as well. However, many subjects are utterly oblivious to the presence of the gorilla. In the related case of change blindness, subjects fail to notice that something has been altered in a display. One instantiation involves flashing an image repeatedly with a brief blank interval in between. In alternate flashes, there is a substantial change in the image; for example, the color of the trousers of one person may change. Even though subjects can freely move their eyes to scrutinize the display, it is often quite tricky and frustrating to spot the change, which may require tens of seconds to detect.

A particular type of visual illusion that has been influential in the study of visual consciousness is 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 a phenomenon known as *binocular rivalry* (***Figure X-3***). Under normal circumstances, the information that the right and left eyes convey is highly correlated. What the right eye and left eye see is not identical: the small differences between the input from the right and left eye provide strong cues to obtain three-dimensional information. What would happen if we show two completely different stimuli to the right and left eyes? Under these conditions, observers perceive either one stimulus *or* the other one, alternating between the two in a seemingly random fashion, a rivalry between the inputs from the two eyes.

Extensive psychophysical investigations have provided a wealth of information about the conditions that lead to perceptual dominance of one or the other visual stimulus, what can or cannot be done with the information that is being suppressed, and the dynamics underlying perceptual alterations. 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.

A simple demonstration of binocular rivalry can be elicited by rolling a piece of paper and looking through it with one eye. With both eyes open and holding the piece of paper with one hand, it is possible for one eye to be focusing on objects far away and for the other eye to focus on the hand in front of you. The percept mysteriously alternates between the hand, and those objects far away seen through an apparent hole in your hand.

INSERT Figure X-3 AROUND HERE

Figure X-3. Binocular rivalry.

A*. A stimulus (Gioconda) is shown to one eye, and a different stimulus (Sunflowers) is shown to the other eye.* ***B****. Perception typically alternates between the two possible percepts, with transient periods of piecemeal rivalry where the two stimuli are merged.*

The duration of dominance for each of the two stimuli follows a gamma distribution, and percepts shift involuntarily from one stimulus to the other, sometimes passing through a mixed percept known as piecemeal rivalry. It is as if the brain were wired to understand that there cannot be two different objects in the same location at the same time. Those two objects compete for perception; one of them wins momentarily, but the fierce competition continues, and eventually, the other object takes over. While the name and the presentation format would seem to suggest a competition between monocular channels, several pieces of evidence suggest that the competition also takes place at a higher level, between representations of the two objects: (1) it is possible to elicit *monocular* rivalry, a weaker phenomenon, where competition between two possible interpretations of the input takes place even though inputs are presented only to one eye via superposition; (2) the stimuli can be arranged such that half of the object information is presented one eye and half to the other eye; instead of experiencing alternations between the two half percepts, rivalry occurs between the two complete objects, which requires putting together information from the two eyes; (3) astute experiments where the stimuli are rapidly shifted from one eye to the other further reveal that the competition can happen at the level of the object representations themselves rather than between the two eyes.

There exist several variations of binocular rivalry. *Flash suppression* refers to a situation where a stimulus, say the Gioconda, is shown monocularly, say to the right eye. Immediately following, the Gioconda stays on the right eye, but a new stimulus, say sunflowers, are flashed onto the other eye. Under these conditions, the new stimulus, the sunflowers, dominate perception, and the old stimulus, the Gioconda, is completely suppressed. If the two stimuli remain on the screen, one shown to each eye, eventually, binocular rivalry ensues, and perception begins to alternate between the two. An interesting variation is the phenomenon of *continuous flash suppression,* where the Gioconda stays on the right eye while a series of stimuli are continuously flashed to the left eye. Under these conditions, subjects perceive the continuous stream of flashed stimuli, and the Gioconda can remain perceptually invisible for several minutes.

As in the other examples, we expect that the activity of RGCs will be oblivious to the internal perceptual alternations in switching between one interpretation of the image and the other one during binocular rivalry. On the other hand, the NCC should directly correlate with perceptual changes.

X.5. Neurophysiological correlates of visual consciousness during binocular rivalry

The phenomenon of binocular rivalry has been prominently studied at the neurophysiological level. Investigators search for the neuronal changes that correlate with the subjective transitions between the input to one or the other eye. An interesting property of binocular rivalry is that the phenomenon can be

triggered using essentially any stimulus shape. Binocular rivalry can take place by presenting a horizontal grating to the right eye and a vertical grating to the left eye, or a picture of a face to the right eye and a picture of a grating to the left eye. Armed with the ability to interrogate neuronal responses along the ventral visual cortex (**Chapters V**-**VI**), we can ask whether neurons that are activated by those stimuli follow the subjective perceptual reports or not.

Nikos Logothetis and collaborators have studied this question extensively throughout the visual cortex. They employed a variety of astute strategies to train monkeys to report their percepts during the perceptual alternations. For example, periods of binocular presentation were randomly intermixed with periods of monocular presentation that can be used as controls to ensure that the monkey is reporting the percepts correctly.

The investigators recorded the activity of visually selective neurons that would respond more strongly to a given stimulus A compared to another stimulus B (similar to the examples shown in **Chapter VI**). Next, the investigators presented A to one eye and B to the other eye (***Figure X-4***). For example, the investigators recorded the activity of a neuron in inferior temporal cortex that responded more strongly to a picture of an orangutan than to a picture of an abstract pattern during monocular presentation or during binocular presentation when the same stimulus was presented to both eyes. Remarkably, when the orangutan and abstract pattern were presented during a binocular rivalry experiment, the dynamic changes in the neuronal firing rate correlated strongly with the monkey's perceptual reports: if the monkey indicated perceiving the orangutan, the neuron would show a high firing rate whereas whenever the monkey indicated perceiving the abstract pattern, the neuron would show a low firing rate. The changes in firing rate preceded the perceptual reports by a few hundred milliseconds, consistent with the idea that the neuronal responses reflect a perceptual change, and that it takes time to elicit the required motor output to provide a perceptual report. The vast majority of neurons in inferior temporal cortex showed this behavior whereby their activity correlated with the subjective perceptual reports.

The activity of neurons in the human medial temporal lobe also shows such correlations with perception. In all of these experiments both in monkeys and humans, neuronal responses may precede the behavioral report of perceptual transitions by a few hundred milliseconds. At least partly, this may indicate that we do not have very accurate ways of measuring the exact timing of the perceptual transition, and the behavioral reports may be delayed. Yet, intriguingly, human medial temporal lobe neurons may become activated well before perceptual transitions, even up to 1,000 milliseconds, and in frontal areas, some neurons were activated even earlier than that. It seems unlikely that such long delays could be ascribed purely to delayed behavioral reports. Therefore these neurons could be involved in as yet poorly understood preconscious mechanisms that eventually culminate in perceptual transitions.

INSERT Figure X-4 AROUND HERE

Figure X-4. Schematic of a neuron that follows the percept during binocular rivalry. ***A****. During monocular presentation, the neuron shows a stronger response to the Gioconda than to the Sunflowers.* ***B****. During binocular rivalry, the neuron shows a stronger response whenever the subject reports perceiving the Gioconda.*

In contrast to the correlations observed for neurons in ITC and the medial temporal lobe, the activity of neurons in V1 typically did *not* follow the subjective report. Primary visual cortex neurons indicated the physical presence of their preferred stimuli, and in most cases, their activity was oblivious to the perceptual reports indicated by the monkey. Intermediate visual areas like V4 and area MT showed results that were in between those in V1 and those in ITC. In other words, there is a progression in the proportion of neurons that correlate with the subjective report as we ascend through the visual hierarchy.

The exact proportion of neurons that correlate with perceptual transitions in a given area may depend on the experimental conditions. For example, an elegant study showed that in area MT, changing the stimulus and the context could lead to different neurons showing firing rates changes concomitant with changes in awareness. In other words, the NCC may not be static, but rather may dynamically depend on the task and conditions.

One concern about these experiments is that we need to obtain a behavioral response from the subjects to figure out what the subjective percept is. Are the neuronal responses indicative of the conscious percepts, or do they reflect the decision and motor signals involved in reporting perception? Several experimental variations have been devised to address these concerns by capitalizing on ingenious ways of reading out what the percept is without a behavioral report. In these so-called no-report paradigms, either pupil size or other independent signatures of one or the other stimulus are used to deduce the perceptual transitions without an overt behavioral report. The results from the no-report paradigms appear to corroborate the results from the earlier studies, showing neuronal correlates of subjective percepts, particularly along the highest echelons of visual cortex.

Another question that has been raised about the interpretation of studies that aim to track correlates of conscious perception is whether neuronal responses reflect changes in consciousness or changes in attention. Under most circumstances, attention and consciousness are strongly correlated, and we are conscious of whatever we are attending to. However, it is possible to design experiments where attention and consciousness are dissociated. These experiments show that subjects can consciously perceive an object or scene in the absence of top-down attentional mechanisms. Additionally, subjects can also pay attention to objects that are perceptually invisible.

X.6. Desiderata for the NCC

Experiments with bistable percepts like binocular rivalry have paved the road towards an initial understanding that changes in specific neuronal activity patterns correlate with transitions in subjective perception. At the same time, there are many other neurons in the brain that continue to fulfill their chores independently of the moment-to-moment contents of consciousness.

What would constitute evidence of finding the NCC? In parallel to the discussion of computational models in **Chapters VII-IX**, we seek a quantitative description of subjective perception. In **Chapter VIII**, we argued that a complete computational account of vision should be able to predict the neuronal responses to *any* arbitrary image, and also predict the behavioral responses in *any* visual task in response to any image. Extending this definition to the domain of visual consciousness, four conditions should be met for a complete account of the NCC for vision:

(1) We should be able to quantitatively predict neuronal responses given a perceptual state. For example, during binocular rivalry, we should be able to predict neuronal activity for neurons in different brain areas, given the perceptual state of the subject.

(2) Conversely, we should be able to predict perceptual states from neuronal responses. By recording the activity of populations of neurons (the specific neuronal types, circuits, and areas for the NCC), we want to tell what the subject is consciously perceiving at any given time.

(3) We should be able to elicit a specific percept by activating the corresponding neuronal patterns (e.g., via electrical stimulation, **Chapter IV**). These neuronal patterns could be in one brain area or multiple brain areas. The resulting percept should be specific (e.g., a woman sitting in an outside park next to a tree), as opposed to merely eliciting phosphenes of light by activating clusters of neurons in primary visual cortex at once. Furthermore, in a binocular rivalry experiment, stimulation of the NCC should be able to shift the perceptual state of the subject. This extended notion of the NCC postulates that activation of those specific neural circuits is directly and causally connected to the perceptual state. Therefore, even if the subject is asleep, activating the NCC should trigger a dream or a hallucination of that specific perceptual state.

(4) We should be able to inactivate or repress a perceptual state by modifying the neuronal activity patterns. In a binocular rivalry experiment, we could ensure that subjects do not perceive one of the stimuli by inactivating the corresponding NCC. Again, because the NCC is directly and causally responsible for perception, in principle, we could show a picture of a woman sitting in a park next to a tree, and the subject would not perceive any of that if the corresponding NCC is inactivated. This manipulation should be specific to the specific contents signaled by the NCC (e.g., closing the eyes to reduce activity in all neurons in the visual system would not constitute a test of this requirement).

Needless to say, we are still a long way from understanding the neuronal correlates of visual consciousness by meeting these four conditions. Nevertheless, nowadays, these questions have become a major area of research, and we may be surprised to observe exciting progress in the field in the years to come.

X.7. Integrated information theory

INSERT Figure X-5 **AROUND HERE**

Figure X-5. Axioms of integrated information theory (IIT).
*IIT proposes **five fundamental** axioms about the nature of conscious experience: (1) intrinsic experience; (2) composition; (3) information; (4) integration; and (5) exclusion (adapted from Tononi and Koch, 2015).*

The previous sections have focused on empirical measurements trying to elucidate which specific neuronal activity patterns correlate with subjective percepts or not. These empirical observations have given rise to accounts about the relative order in which different areas may be activated during conscious perception. The relative order of activation of different neural circuits during perceptual transitions is summarized in the idea of a global workspace that takes sensory information and spreads this information "globally," or at least to multiple other brain regions. Some investigators have proposed that the spreading to other brain regions ignites changes in subjective perception.

In parallel to the empirical observations about neural activity patterns that accompany visual consciousness, the last decade has seen the development of an elegant, ambitious, and controversial 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. This framework nicely starts with a set of five axioms (***Figure X-5***) and quantitatively derives a definition of information and integration. These five axioms state that: (1) consciousness exists as a unique internal experience (*intrinsic existence*), (2) conscious experience is composed of multiple phenomenological elements (*composition*), (3) consciousness is specific (*information*), (4) conscious experience is unified and irreducible (*integration*), and (5) the content of consciousness is circumscribed in space and time (*exclusion*). The theory then derives postulates from these axioms to establish the necessary conditions for a system to show these aspects of experience.

According to IIT, a dynamical system of interconnected parts is characterized by a metric, connoted by Φ (phi), which has a higher value when the system *cannot* be described by smaller, relatively independent, subsystems. The larger Φ, the more integrated information the system has. The theory postulates that conscious experience is proportional to Φ. The definition of Φ comprises two steps: (1) perform an imaginary partition of the system and compute ϕ, 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 (2) define Φ as the "cruelest" such partition that minimizes ϕ. Elegantly, the theory provides specific mathematical definitions to calculate these quantities, given the dynamic transitions in a system of interconnected parts like a neuronal circuit.

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. Max Tegmark and others recently developed an approximation to calculate Φ using graph theory, bringing the calculations to a polynomial dependency on the system size and making this algorithm readily applicable to the large scale of physiological recordings.

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 makes many counterintuitive predictions. Any object, the cellular phone, or even the chair we are sitting on has a certain Φ value. One may expect that inanimate objects or bacteria should have $\Phi=0$, but this is not what the theory states. Those objects may have low values of Φ, perhaps even negligibly small, but not zero. Intuitively, one would like any theory to indicate that a chair has no consciousness, not that it has a small amount of consciousness. Perhaps this is more of a semantic concern that can be remedied by imposing a threshold on Φ.

Another bewildering aspect of IIT is that it is, in principle, possible to create relatively "simple" artificial systems with high Φ values (for the aficionados, an example is the so-called *Vandermonde* matrices). However, it seems counterintuitive that such artificial systems would show consciousness. Of course, the introspective observation that these predictions are counterintuitive does not make them wrong. There are plenty of examples in science where counterintuitive predictions have led the way to exciting new discoveries. Science should be guided by experimentally testable predictions and the empirical results, not by our taste or intuitions.

Ultimately, it will be interesting to test the integrated information theory empirically. Regardless of whether this theoretical framework is entirely right, whether it will require revisions and refinements like all other theories in science, or even if it is entirely wrong, it is the very first time that a quantitative theory has been proposed to account for one of the most elusive mysteries of human existence, consciousness.

X.8. Summary

- Consciousness has been discussed for millennia by thinkers from a wide variety of different fields, yet only recently has it become an important topic of investigation for rigorous neuroscience theorists and experimentalists.

- Experimental efforts have focused on searching for minimal and jointly sufficient neuronal correlates of consciousness, the NCC.

- Several experimental paradigms, where the input is constant, yet perception changes over time, have been developed to study visual consciousness. These experiments include backward masking, attentional manipulations, visual search, and bistable percepts such as binocular rivalry.

- During binocular rivalry, neuronal responses in the highest parts of visual cortex correlate with the dynamical changes in the contents of consciousness.

- A full description of the NCC would require a quantitative computational model that can predict neuronal responses given the perceptual state, and that can also predict the perceptual state given the neuronal responses. Activating or suppressing the NCC should elicit or silence specific perceptual states.

- Integrated information theory (IIT) is the first quantitative theoretical framework that aims to explain how consciousness emerges from a dynamical system with interconnected parts.

X.9. References

- Crick F (1994) The astonishing hypothesis. New York: Simon & Schuster.
- 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.
- Tononi G (2005) Consciousness, information integration, and the brain. Prog Brain Res 150:109-126.
- Tononi G, Boly M, Massimini M, Koch C (2016) Integrated information theory: from consciousness to its physical substrate. Nat Rev Neurosci 17:450-461.

Figure X-1

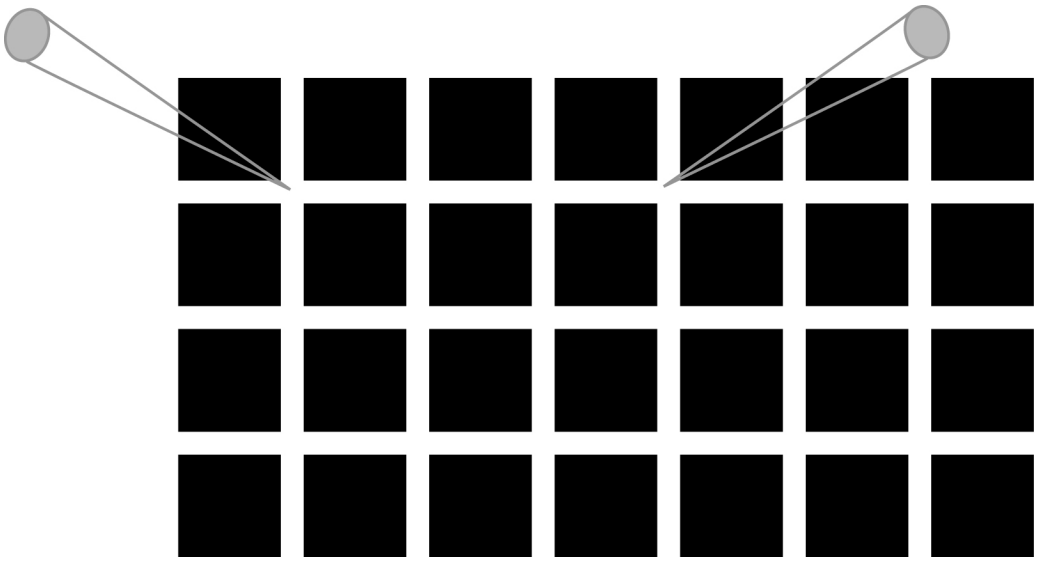


Figure X-2

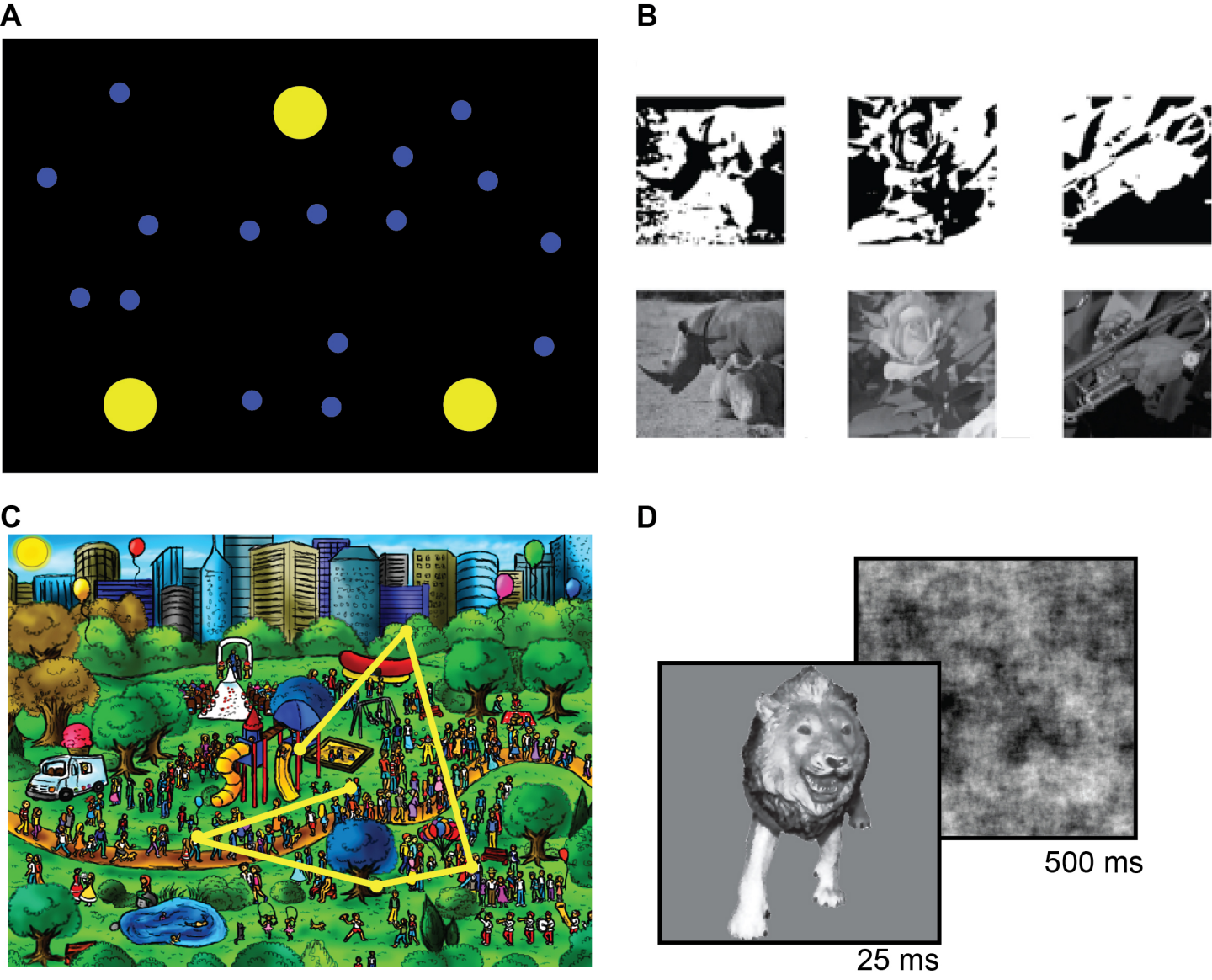


Figure X-3

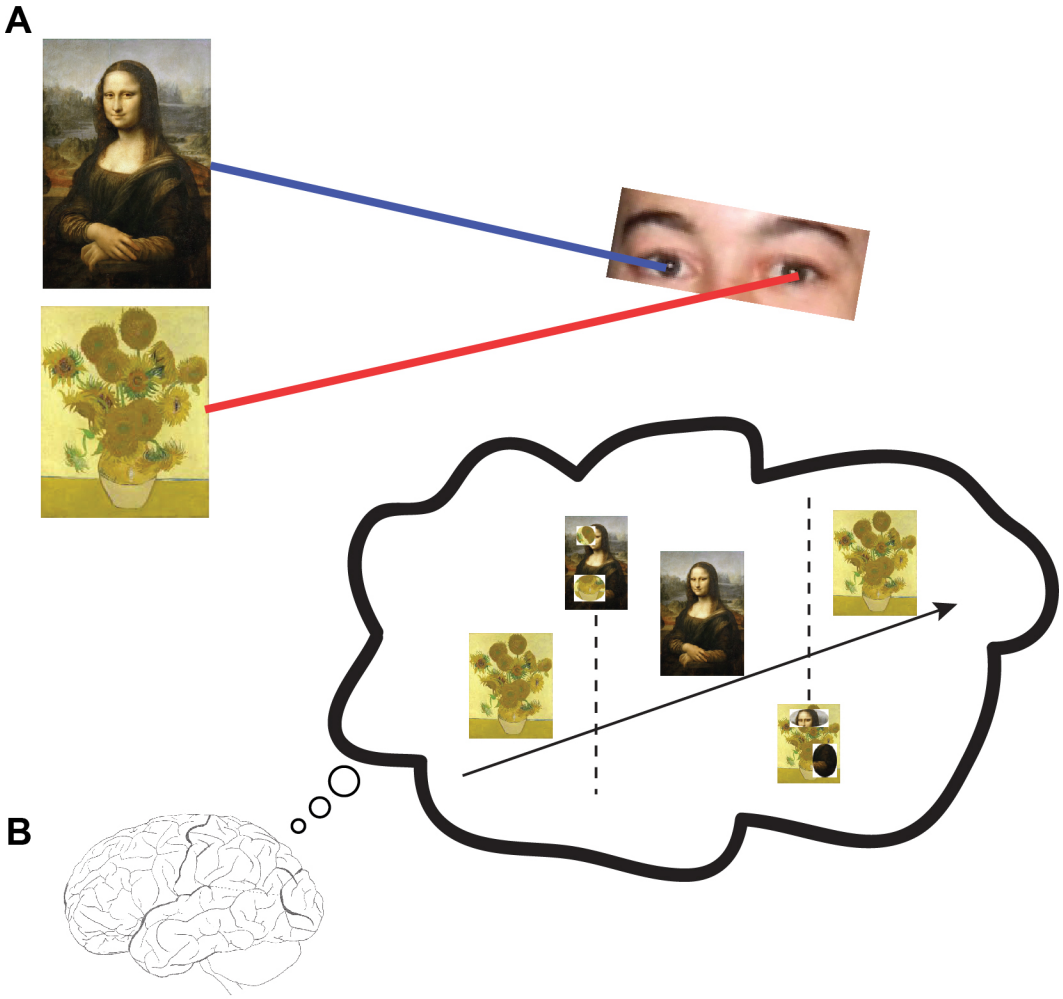


Figure X-4

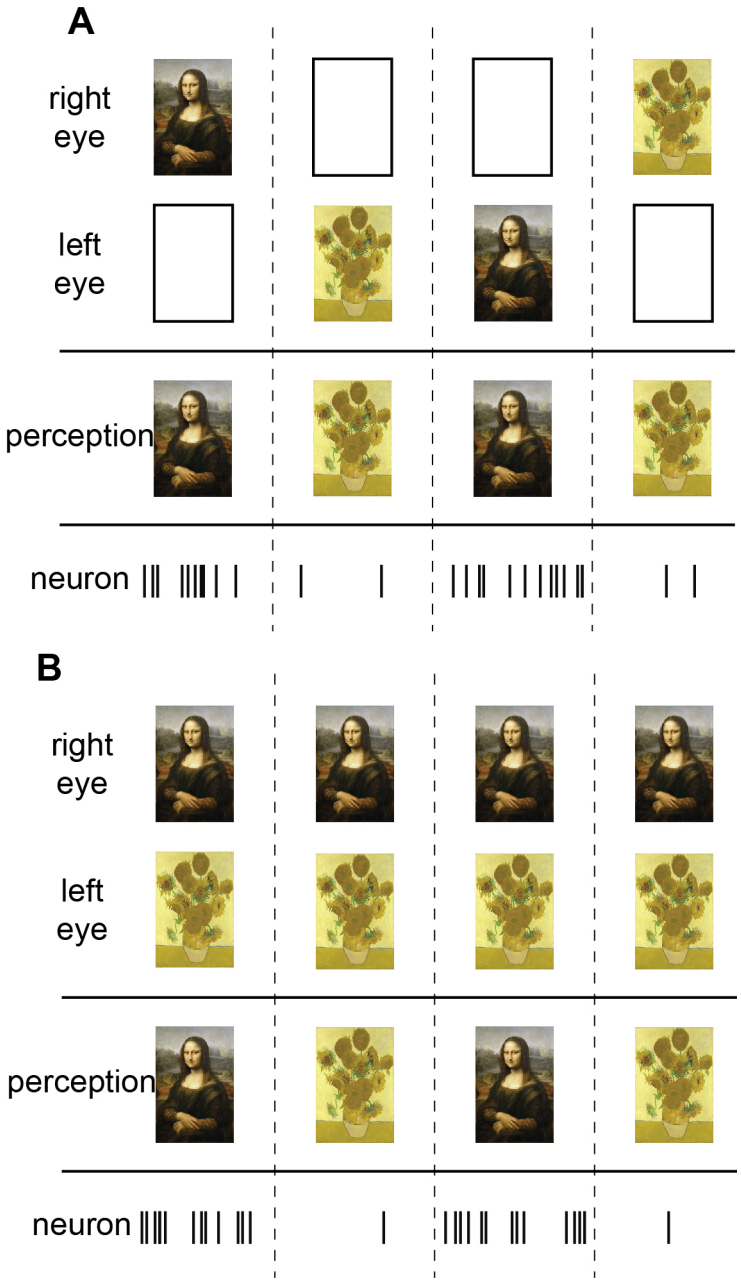
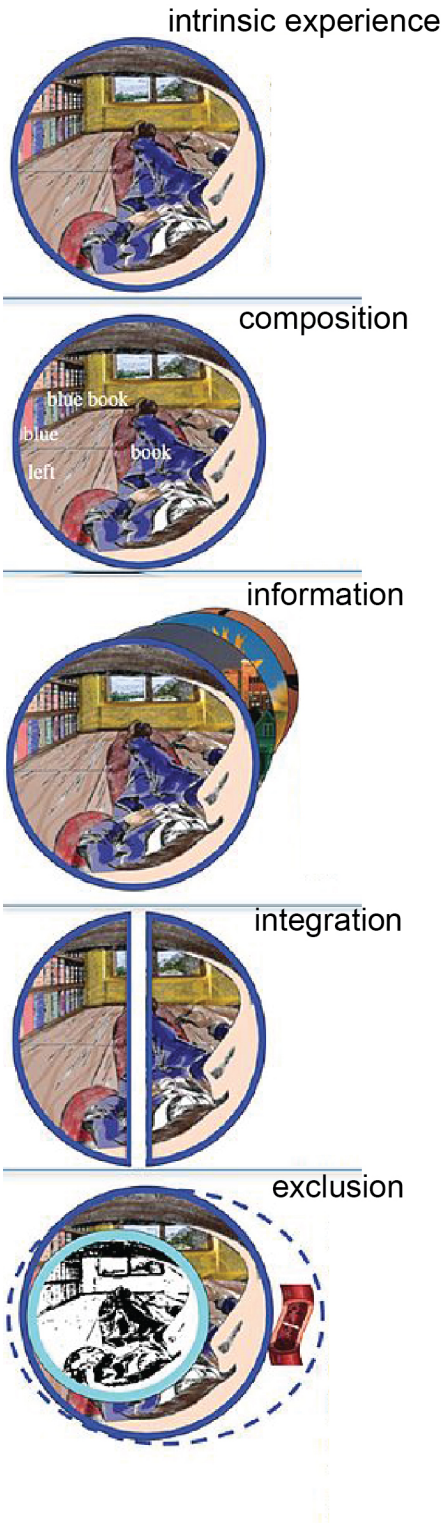

A
right eye
left eye
perception
neuron
B
right eye
left eye
perception
neuron

Figure X-5


intrinsic experience
composition
blue book
blue
book
left
information
integration
exclusion