The Radical Plasticity Thesis: Consciousness is something that the brain learns to do

Abstract
Here, I explore the idea that consciousness is something that the brain learns to do rather than an intrinsic property of certain neural states and not others. Starting from the idea that neural activity is inherently unconscious, the question thus becomes: How does the brain learn to be conscious? I suggest that consciousness arises as a result of the brain's continuous attempts at predicting not only the consequences of its actions on the world and on other agents, but also the consequences of activity in one cerebral region on activity in other regions. By this account, the brain continuously and unconsciously learns to redescribe its own activity to itself, so developing systems of meta-representations that characterise and qualify the target first-order representations. Such learned redescriptions, enriched by the emotional value associated with them, form the basis of conscious experience. Learning and plasticity are thus central to consciousness, to the extent that experiences only occur in experiencers that have learned to *know* they possess certain first-order states and that have learned to *care* more about certain states than about others. This is what I call the "Radical Plasticity Thesis". In a sense thus, this is the enactive perspective, but turned both inwards and (further) outwards. Consciousness involves "signal detection on the mind"; the conscious mind is the brain's (non-conceptual, implicit) theory about itself. I illustrate these ideas through neural network models that simulate the relationships between performance and awareness in different tasks.

Keywords: Prediction, consciousness, neural networks, metacognition, enaction.

1 Introduction

Prediction is a ubiquitous computational principle in the brain (Bar, 2009; Clark, 2013). As Clark puts it: "Brains, it has recently been argued, are essentially prediction machines". Clark fleshes out this claim by highlighting the lineage between the early ideas of Helmoltz (1860/1962) and the much more recent ideas associated with contemporary connectionist models (McClelland & Rumelhart, 1986; Rumelhart, Hinton, & Williams, 1986), in particular the so-called generative models described by Hinton (Dayan, Hinton, & Neal, 1995; Hinton, 2007) and by Friston and colleagues (Friston, 2006, 2010). There is considerable evidence for predictive mechanisms in the human brain (Bar, 2009). This idea, in fact, forms the core of the Bayesian perspective on information processing and is at the heart of Friston's free energy principle (Friston,

2006), according to which the brain continuously attempts to minimize "surprise" or conflict by anticipating its own future activity based on learned priors.

Helmholz first proposed the idea that perception involves a form of prediction-driven inference through which the mind attempts to reconstruct the sensory causes of bodily effects. Perception, in this view, is thus an active process of elaborating the best possible representations of the input based on both the sensory evidence and relevant prior knowledge (the "priors") rather than a mere bottom-up process. The work of Hinton, Friston and colleagues elaborated on this view by showing how it is possible to conceive learning rules able to shape top-down connection weights to as to minimize "prediction error", that is, the difference between expected and observed inputs. In such models, the top-down flow of information thus attempts to "explain the sensory input away", leaving only information about the residual errors (the "prediction error") to flow upwards in a hierarchy of interconnected layers. This is the core principle of "predictive coding" (Rao & Ballard, 1999), through which hierarchical systems can simultaneously learn about their inputs and about the best internal models of these same inputs. As Clark (2013) puts it, this dampens the distinction between perception and belief to the point that they appear almost identical to each other : "To perceive the world just is to use what you know to explain away the sensory signal across multiple spatial and temporal scales. The process of perception is thus inseparable from rational (broadly Bayesian) processes of belief fixation, and context (top down) effects are felt at every intermediate level of processing. As thought, sensing, and movement here unfold, we discover no stable or well-specified interface or interfaces between cognition and perception. Believing and perceiving, although conceptually distinct, emerge as deeply mechanically intertwined." (p. 29). Here, I will attempt to delineate the deep connections between these ideas and a novel theory of consciousness in which prediction-driven learning and plasticity mechanisms play a central role.

2 Consciousness as a prediction-driven redescription process

A central aspect of the entire hierarchical predictive coding approach is the emphasis it puts on learning mechanisms. In other works (Cleeremans, 2008, 2011), I have defended the idea that consciousness is itself the result of learning. From this perspective, agents *become* conscious in virtue of redescribing their own activity to themselves. In this respect, it is important to note that learning can create as well as eliminate contents from phenomenal experience. Thus, tasting wine for the first time is a wholly different experience than that of an oenologist (Smith, 2006), whose phenomenology has been enriched through expertise. But expertise can also eliminate phenomenal contents from awareness, as in the 'find the F's" illusion, whereby observers are asked to count the number of instances of the letter "F" in a display. Observers often fail to reach the correct answer because reading expertise has eliminated function words from awareness. There are many other examples of such "predictive attenuation" mechanisms: Tickling one's self is far less effective than being tickled (Blakemore, Frith, & Wolpert, 1999), for when we tickle ourselves (but not

when we are tickled) our brain can predict the consequences of our actions. Cognitive development also highlights how some changes go unheeded (i.e., the fact that our action and perceptual systems remain adapted despite our limbs growing spectacularly during the first few years) whereas other changes have profound phenomenal consequences (i.e., learning to read). Thus, learning shapes conscious experience and how conscious experiences shapes learning. Consciousness, in this light, is a profoundly dynamical process through which our experience of the world is constantly shaped both by incoming sensory inputs and by our learned expectations about what is coming next.

Taking the proposal that consciousness is inherently dynamical seriously has numerous theoretical and methodological consequences. But it also opens up the mesmerizing possibility that conscious awareness is itself a product of plasticity-driven dynamics. In other words, from this perspective, we learn to be conscious. To dispel possible misunderstandings of this proposal right away, I am not suggesting that consciousness is something that one learns like one would learn about the Hundred Years War, that is, as an academic endeavour, but rather that consciousness is the result (vs. the starting point) of continuous and extended interaction with the world, with ourselves, and with others. The brain, from this perspective, continuously (and unconsciously) learns to anticipate the consequences of its own activity on itself, on the environment, and on other brains, and it is from the practical knowledge that accrues in such interactions that conscious experience is rooted. This perspective, in short, endorses the enactive approach introduced by O'Regan and Noë (O'Regan & Noë, 2001), but extends it both inwards (the brain learning about itself) and further outwards (the brain learning about other brains), so connecting with the central ideas put forward by the predictive coding approach to cognition. In this light, the conscious mind is the brain's (implicit, enacted) theory about itself, expressed in a language that other minds can understand.

The theory rests on several assumptions and is articulated over three core ideas. A first assumption is that information processing as carried out by neurons is intrinsically unconscious. There is nothing in the activity of individual neurons that make it so that their activity should produce conscious experience. Important consequences of this

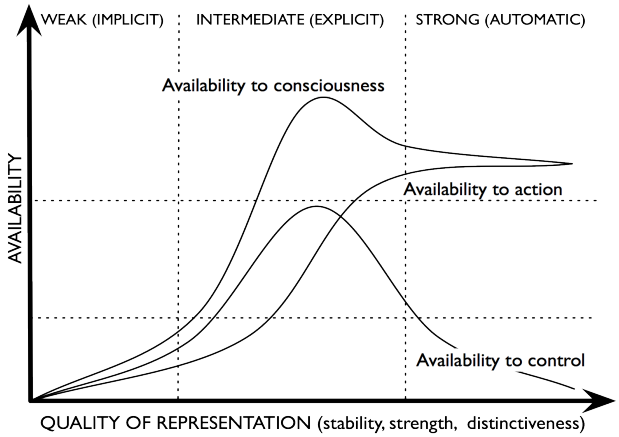

Figure 1: The "QoR" Framework

assumption are (1) that conscious and unconscious processing must be rooted in the same set of representational systems and neural processes, and (2) that tasks in general will always involve both conscious and unconscious influences, for awareness cannot be "turned off" in normal participants. A second assumption is that information processing as carried out by the brain is graded and cascades (McClelland, 1979) in a continuous flow (Eriksen & Schultz, 1979) over the multiple levels of a heterarchy (Fuster, 2008) extending from posterior to anterior cortex as evidence accumulates during an information processing episode. An implication of this assumption is that consciousness takes time. The third assumption is that plasticity is mandatory: The brain learns all the time, whether we intend to or not. Each experience leaves a trace in the brain (Kreiman, Fried, & Koch, 2002). With these assumptions in place, the theory is articulated around three core ideas.

The first is that consciousness depends on quality of representation (see Figure 1). "Quality of representation" (QoR), here, designates graded properties of neural representations, specifically their Strength, their Stability in time, and their Distinctiveness. QoR depends both on bottom-up factors such as stimulus properties and on top-down factors such as attention. QoR determines the extent to which a representation is available to (1) influence behaviour, (2) form the contents of awareness, (3) be the object of cognitive control and other high-level processes. Crucially, QoR *changes* as a function of learning and plasticity over different time scales (processing within a single trial, learning, and development), as depicted in Figure 1. The first region of the figure, labeled "Implicit Cognition", corresponds to the point at which processing starts in the context of a single trial, or to some early stage of development or skill acquisition. This stage is characterized by weak, poor-quality representations. Implicit representations are capable of influencing behaviour, but only weakly so (e.g., through priming). The second region corresponds to the emergence of higher-quality explicit representations, here defined as representations over which one can exert control. Such representations are good candidates for redescription and can thus be recoded in different ways, e.g., as linguistic propositions (supporting verbal report). The third region involves what I call automatic representations, that is, representations that have become so strong that their influence on behavior can no longer be inhibited (e.g., as in the Stroop situation). Such representations exert a mandatory influence on processing. Importantly, however, and unlike the weak representations characteristic of implicit cognition, one is (at least potentially) aware of possessing such strong representations and of their influence on processing. Thus, both the weak representations characteristic of implicit cognition and the very strong representations characteristic of automaticity cannot be controlled, but for very different reasons. This leaves intermediate-quality (explicit) representations, that is, representations that are strong enough that their influence on behaviour needs to be monitored yet not sufficiently adapted that they can be "trusted", as those representations that require the most cognitive control. Crucially, this also predicts that intermediate-quality representations are the most susceptible to be influenced by other sources of knowledge, as they are the most flexible. One would thus expect non-

monotic effects as expertise develops, in different paradigms ranging from perception to motor skill learning.

The second core idea is that consciousness depends on metarepresentations. Even strong stimuli can fail to enter conscious awareness — this is what happens in change blindness (Simons & Levin, 1997), in the attentional blink (Shapiro, Arnell, & Raymond, 1997) or in inattentional blindness (Mack & Rock, 1998). States of altered consciousness like hypnosis, and pathological states such as blindsight (Weiskrantz, 1986) or hemineglect likewise suggest that high-quality percepts can fail to be represented in awareness while remaining causally efficacious. This suggests that quality of representation, while necessary for conscious awareness, is not sufficient.

One way of understanding what is missing is to appeal to the central hypothesis of

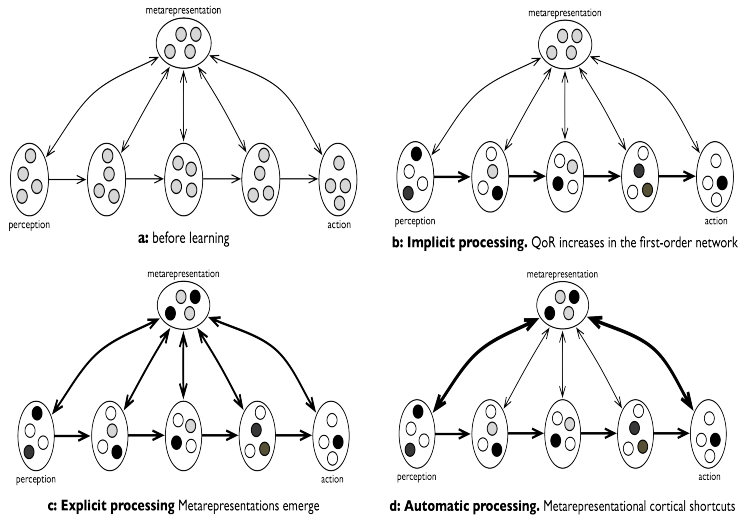

Figure 2: Implicit, explicit & automatic processing

the Higher-Order Thought (HOT) Theory of consciousness (Rosenthal, 1997), namely that a representation is a conscious representation when one knows *that* one is *conscious of* the representation. This roots conscious awareness in a system's capacity to redescribe its own states to itself, a process ("representational redescription") also viewed as central during cognitive development (Karmiloff-Smith, 1992) and metacognition in general (Nelson & Narens, 1990). A system's ability to redescribe its own knowledge to itself depends (1) on the existence of recurrent structures that enable the system to access its own states, and on (2) the existence of predictive models (metarepresentations) that make it possible for the system to characterize and anticipate the occurrence of first-order states (Bar, 2009; Friston, 2006; Wolpert, Doya, &

Kawato, 2004). Such redescription is also uniquely facilitated, in humans, by language, viewed here as the metarepresentational tool per excellence. A natural spot for such metarepresentations to play their functions is the prefrontal cortex (i.e., Crick & Koch's "the front is looking at the back" principle (Crick & Koch, 2003)). Importantly however, here, such metarepresentational models (1) may be local and hence occur anywhere in the brain, (2) can be subpersonal, are (3) are subject, just like first-order representations, to plasticity and hence can themselves become automatic. Metacognition, just like cognition, can thus involve implicit, explicit, or automatic metarepresentations.

The theory thus proposes a novel conception of skill acquisition that links automaticity with the observation that conscious awareness seems to proceed from the top down (i.e., Crick & Koch's "the high levels first" principle (Crick & Koch, 2003)): We become aware of the higher-level aspects (the gist) of a scene before becoming aware of its lower-level features. I suggest that this stems from the fact that, from a computational point of view, metarepresentations implement what one could call cortical reflexes or shortcuts: A system that has *learned* to redescribe the activity of an entire feedforward pathway can now also anticipate the consequences of early activity in such a chain on its output faster than the pathway itself can compute the output. As a result, *adapted* metarepresentations (and only adapted metarepresentations) make it possible to bypass the first-order pathway altogether. I surmise that this accounts not only for the fact that the time course of (expert) perception seems to follow a reverse hierarchy (Ahissar & Hochstein, 2004), but also for the fact that automaticity entails loss of access to the contents computed along the first-order pathway. By the same token, this also opens up the possibility for *postdictive effects* in conscious experience, as metarepresentations are shaped by first-order processing. This top-down view of automaticity contrasts with extant theories (Chein & Schneider, 2012).

The theory distinguishes four reasons why knowledge may remain unconscious. First (Figure 2a), knowledge embedded in synapses is assumed not be accessible at all, for such knowledge fails to be instantiated in the form of active patterns of neural activity (Koch, 2004), a necessary condition for their contents to be available to awareness. The provocative idea here is that the brain *does not know*, e.g., that SMA activity consistently precedes M1 activity. To represent this causal link to itself, it therefore has to learn to redescribe its own activity so that the causal link is now represented explicitly as a metarepresentation. Second, weak representations (Figure 2b), while they can influence behaviour, remain unconscious for they fail to be sufficiently strong to be the target of metarepresentations. Third, when sufficiently strong, first-order representations can begin to be redescribed into metarepresentations (Figure 2c), yet, other conditions (e.g., lack of attention induced by distraction, failure to properly redescribe first-order contents) may make such redescription impossible or difficult. Fourth, the very strong representations characteristic of automaticity (Figure 2d) are not necessary anymore to drive behaviour since the learned metarepresentations now implement a faster "shortcut" pathway from input to output. This also accounts for the

fact that metacognitive accuracy often lags first-order performance initially, but *precedes* first-order performance with expertise (i.e., I know that I know the answer to a query before I can actually answer the query).

The distinctions introduced here overlap partially with the distinctions introduced by existing theories of consciousness: Dehaene's conscious - preconscious - unconscious taxonomy (Dehaene, Changeux, Naccache, Sackur, & Sergent, 2006), Lamme's Stages 1/2/3/4 framework (Lamme, 2006), and Kouider's partial awareness hypothesis (Kouider, de Gardelle, Sackur, & Dupoux, 2010), but uniquely frames the transitions dynamically as resulting from leaning.

The third core idea is that consciousness depends on theory of mind (Timmermans, Schilbach, Pasquali, & Cleeremans, 2012) The emergence of an agent's ability to redescribe its own representations to itself in the way sketched above, I argue, critically depends on the agent being embedded in interaction with other agents. From this perspective, as Frege pointed out, *conscious experience cannot be understood*

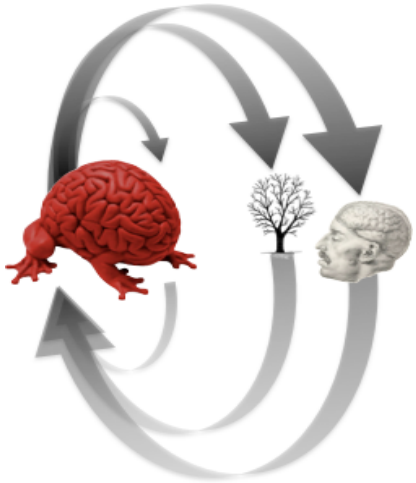

Figure 3: Tangled Loops

independently from the agent who experiences these experiences. Yet, as obvious as this may seem, neuroscientists have approached the question as though the differences between conscious and unconscious representations could be understood independently of the subject, from a purely "objective", third-person point of view. The entire "search for the Neural Correlates of Consciousness" is, in this sense at least, misguided. As Donald (2001) put it, "the human mind is unlike any other on this planet, not because of its biology, which is not qualitatively unique, but because of its ability to generate and assimilate culture" (p. xiii). Thus, I build a model of myself not only by developing a

non-conceptual understanding of how my goals are eventually expressed in action, but also by understanding how agents similar to me react to actions directed towards them. It is thus essential that we strive to understand how interactions with other agents shape our own conscious experiences.

Putting the three core ideas together, we end up with the radical plasticity thesis (Cleeremans, 2008, 2011), that is, with the idea that consciousness emerges in cognitive systems that are capable of *learning* to redescribe their own activity to themselves. In other words, one "learns to be conscious". From this perspective, the brain is continuously and unconsciously learning to anticipate the consequences of action or activity on itself, on the world, and on other people.

3 Conclusion

Thus, we have three closely interwoven loops (Figure 3) all driven by the very same prediction-based mechanisms. A first, internal or "inner loop"**,** involves the brain redescribing its own representations to itself as a result of its continuous unconscious attempts at predicting how activity in one region influences activity in other regions. In this light, consciousness amounts to the brain's performing signal detection on its own representations (Lau, 2008), so continuously striving to achieve a coherent (prediction-based) understanding of itself. It is important to keep in mind that this inner loop in fact involves multiple layers of recurrent connectivity, at different scales throughout the brain. A second "perception-action loop", results from the agent as a whole predicting the consequences of its actions on the world. The third loop is the "self-other loop", and links the agent with other agents, again using the exact same set of mechanisms as involved in the other two loops. The existence of this third loop is constitutive of conscious experience, I argue, for it is in virtue of the fact that as an agent I am constantly attempting to model other minds that I am able to develop an understanding of myself. In the absence of such a "mind loop", the system can never bootstrap itself into developing the implicit, embodied, transparent (Metzinger, 2003) model of itself that forms the basis, through Higher-Order Thought Theory, of conscious experience. The processing carried out by the inner loop is thus causally dependent on the existence of both the perception-action loop and the self-other loop, with the entire system thus forming a "tangled hierarchy" (e.g., Hofstadter's concept of "a strange loop" (Hofstadter, 2007)) of predictive internal models (Pacherie, 2008; Wolpert et al., 2004). Consciousness, in this light, is thus the brain's implicit, embodied theory about itself, achieved through continuously operating prediction-driven learning mechanisms.

References

Ahissar, M., & Hochstein, S. (2004). The reverse hierarchy theory of visual perceptual learning. *Trends in Cognitive Sciences, 8*(10), 457-464.

Bar, M. (2009). Predictions: a universal principle in the operation of the human brain. *Philosophical Transactions of the Royal Society B, 364*, 1181-1182.

Blakemore, S. J., Frith, C. D., & Wolpert, D. M. (1999). Spatiotemporal prediction modulates the perception of self-produced stimuli. *Journal of Cognitive Neuroscience, 11*(5), 551-559.

Chein, J. M., & Schneider, W. (2012). The brain's learning and control architecture. *Current Directions in Psychological Science, 21*(2), 78-84.

Clark, A. (2013). Whatever next? Predictive brains, situated agents, and the future of cognitive science. *Behavioral and Brain Sciences, 36*(3), 181-204.

Cleeremans, A. (2008). Consciousness: The radical plasticity thesis. *Progress in Brain Research, 168*, 19-33.

Cleeremans, A. (2011). The radical plasticity thesis: How the brain learns to be conscious. *Frontiers in Psychology, 2*, 1-12.

Crick, F. H. C., & Koch, C. (2003). A framework for consciousness. *Nature Neuroscience, 6*(2), 119-126.

Dayan, P., Hinton, G. E., & Neal, R. M. (1995). The Helmholz machine. *Neural Computation, 7*(889-904).

Dehaene, S., Changeux, J.-P., Naccache, L., Sackur, J., & Sergent, C. (2006). Conscious, preconscious, and subliminal processing: A testable taxonomy. *Trends in Cognitive Sciences, 10*(5), 204-211.

Donald, M. (2001). *A mind so rare*. New York: W.W. Horton.

Eriksen, C. W., & Schultz, D. W. (1979). Information processing in visual search: A continuous flow conception and experimental results. *Attention, Perception & Psychophysics, 25*(4), 249-263.

Friston, K. (2006). A free energy principle for the brain. *Journal of Physiology (Paris), 100*, 70-87.

Friston, K. (2010). The free energy principle: a unified brain theory? *Nature Reviews Neuroscience, 11*(2), 127-138.

Fuster, J. M. (2008). *The prefrontal cortex* (4th ed.). London: Academic Press.

Helmholz, H. (1860/1962). *Handbuch der physiologgischen optik* (J. P. C. Southall Ed. English translation ed. Vol. 3). New York: Dover.

Hinton, G. E. (2007). Learning multiple layers of representation. *Trends in Cognitive Sciences, 11*, 428-434.

Hofstadter, D. R. (2007). *I am a strange loop*. New York: Basic Books.

Karmiloff-Smith, A. (1992). *Beyond modularity : A developmental perspective on cognitive science*. Cambridge: MIT Press.

Koch, C. (2004). *The quest for consciousness. A neurobiological approach*. Englewood, CO: Roberts & Company Publishers.

Kouider, S., de Gardelle, V., Sackur, J., & Dupoux, E. (2010). How rich is consciousness: The partial awareness hypothesis. *Trends in Cognitive Sciences, 14*(7), 301-307.

Kreiman, G., Fried, I., & Koch, C. (2002). Single-neuron correlates of subjective vision in the human medial temporal lobe. *Proceedings of the National Academy of Sciences of the U.S.A., 99*(8378-8383).

Lamme, V. A. F. (2006). Toward a true neural stance on consciousness. *Trends in Cognitive Sciences, 10*(11), 494-501.

Lau, H. (2008). A higher-order Bayesian Decision Theory of consciousness. In R. Banerjee & B. K. Chakrabarti (Eds.), *Models of brain and mind. Physical, computational and psychological approaches. Progress in Brain Research. Progress in Brain Research* (Vol. 168, pp. 35-48). Amsterdam: Elsevier.

Mack, A., & Rock, I. (1998). *Inattentional Blindness*. Cambridge, MA: MIT Press.

McClelland, J. L. (1979). On the time-relations of mental processes: An examination of systems in cascade. *Psychological Review, 86*, 287-330.

McClelland, J. L., & Rumelhart, D. E. (1986). *Parallel Distributed Processing. Explorations in the Microstructure of Cognition. Volume 2: Psychological and Biological Models*. Cambridge, MA: MIT Press.

Metzinger, T. (2003). *Being No One: The self-model theory of subjectivity*. Cambridge, MA: Bradford Books, MIT Press.

Nelson, T. O., & Narens, L. (1990). Metamemory: A theoretical framework and new findings. *The Psychology of Learning and Motivation, 26*, 125-173.

O'Regan, J. K., & Noë, A. (2001). A sensorimotor account of vision and visual consciousness. *Behavioral and Brain Sciences, 24*(5), 883-917.

Pacherie, E. (2008). The phenomenology of action: A conceptual framework. *Cognition, 107*, 179-217.

Rao, R., & Ballard, D. (1999). Predictive coding in the visual cortex: A functional interpretation of some extra-classical receptive field effects. *Nature Neuroscience, 2*(1), 79.

Rosenthal, D. (1997). A theory of consciousness. In N. Block, O. Flanagan, & G. Güzeldere (Eds.), *The Nature of Consciousness: Philosophical Debates*. Cambridge, MA: MIT Press.

Rumelhart, D. E., Hinton, G. E., & Williams, R. (1986). Learning representations by back-propagating errors. *Nature, 323*, 533-536.

Shapiro, K. L., Arnell, K. M., & Raymond, J. E. (1997). The Attentional Blink. *Trends in Cognitive Sciences, 1*, 291-295.

Simons, D. J., & Levin, D. T. (1997). Change Blindness. *Trends in Cognitive Sciences, 1*, 261-267.

Smith, B. C. (2006). *Questions of taste: The philosophy of wine*. New York: Oxford University Press.

Timmermans, B., Schilbach, L., Pasquali, A., & Cleeremans, A. (2012). Higher order thoughts in action: consciousness as a unconscious re-description process. *Philosophical Transcations of the Royal Society B, 367*, 1412-1423.

Weiskrantz, L. (1986). *Blindsight: A case study and implications*. Oxford, England: Oxford University Press.

Wolpert, D. M., Doya, K., & Kawato, M. (2004). A unifying computational framework for motor control and social interaction. In C. D. Frith & D. M. Wolpert (Eds.), *The neuroscience of social interaction* (pp. 305-322). Oxford, UK: Oxford University Press.