23 Conscious and unconscious cognition: A graded, dynamic perspective

Axel Cleeremans

Consider the following three situations: learning to perform a complex skill such as gymastics (a stunning demonstration of which participants to ICP 2004 experienced during the opening ceremony), learning a complex game such as the ancient Chinese game of Weichi (more widely known as Go), or learning natural language. What these situations have in common, beyond the sheer complexity of the required skills, is the fact that most of what we learn about each appears to proceed in a manner that does not depend so much on the acquisition of explicit, declarative information or on the deployment of intentional strategies, but instead critically depends on repeated practice: Developing the skills needed to execute complex movements in gymnastics, to "see" the correct move in a game of Go, or to fluently produce natural language all involve what Zhu and Simon (1987) dubbed "learning by doing".

Interestingly, in all such cases, it is also often found that people have limited conscious knowledge of what it is they have learned. Thus, people may be, to some degree at least, unaware *that* they have learned anything, unaware of *what* it is that they have learned, or unaware that their behaviour is influenced by something they have previously learned about. Hence we obtain changes in performance that are not accompanied by correlated changes in subjective experience or in people's ability to verbalize knowledge about what was learned. These observations, replicated numerous times over the years (see below for several illustrations) in different domains, have generated and continue to generate controversy. Authors disagree on two apparent implications of such dissociation findings.

First, such findings suggest that *unconscious knowledge can be causally efficacious*. However, many authors would disagree with the notion that the sorts of representations one would want to call "cognitive" can also be unconscious (Perruchet & Vinter, 2003; Shanks & St. John, 1994). Second, such dissociation findings also suggest that there may be exist *separable conscious and unconscious processing systems*. In this paper I will attempt to show that both implications are in fact unwarranted. First, the notion of unconscious representation is only problematic under certain assumptions of what one means by representation. Second, dissociation findings do not necessarily imply the existence of distinct systems whose function it is to

subserve implicit or explicit learning. I will instead suggest that traditional accounts of the differences between cognition with and without consciousness need to be replaced by subsymbolic accounts, which characterize both representation and processing in a profoundly different manner.

This analysis is rooted in a novel conceptual framework (Cleeremans, forthcoming; Cleeremans & Jiménez, 2002) that takes as its starting points (1) that the main function of consciousness is to make flexible, adaptive control over behavior possible; (2) that consciousness is best viewed as involving a graded continuum expressed over representations; and (3) that learning is a mandatory process that always accompanies information processing, and through which our conscious representations of the world are made to reflect those contents which are most in need of control at some point in time. The central claim of the framework is that implicit and explicit learning do not depend on separable, distinct learning and memory systems (although the framework is not wholly incompatible with this perspective either, but not in the specific sense that there exist separable implicit and explicit memory systems), but rather that they are different manifestations of a single set of interacting learning mechanisms and representational systems dedicated to serve specific, and sometimes incompatible, computational, objectives.

In the following, I begin with an overview of some recent findings about the ubiquity of learning. Next, I illustrate through several examples how changes in behavior can fail to be accompanied by changes in subjective experience, and briefly discuss how these findings of dissociation have tended to be interpreted by traditional models in cognitive psychology. I then present the sketch of an alternative framework through which to link learning and consciousness.

Learning and consciousness

The recent years have brought forward a staggering amount of evidence suggesting not only that the brain is far more plastic than previously thought, but also that the effects of learning can be tracked all the way down to the organization of local connectivity. For instance, expert players of stringed instruments exhibit larger-than-normal areas of the somatosensory cortex dedicated to representing input from the fingering digits (Elbert, Pantey, Wienbruch, Rockstroh, & Taub, 1995). Likewise, not only is posterior hippocampus – a region of the brain involved in episodic and spatial memory – enlarged in experienced taxi drivers compared to subjects who do not have extensive experience in memorizing complex maps, but the observed size differences further depend on the amount of driving experience (Maguire et al., 2000). There is also considerable evidence that the brain can recover in various flexible ways after trauma, and even suggestions that the very organization of the somatosensory cortex (the famous Penfield homunculus) depends on pre-natal sensory experience (Farah, 1998). These spectacular findings all reassert that adaptation plays a fundamental role in cognition and, just

as importantly, that learning actually modifies the very structure of the "machine" – the brain – in which information processing takes place.

What, then, can we say about the relationships between learning and consciousness? Is it the case, as some authors contend (Perruchet & Vinter, 2003; Shanks & St. John, 1994) that learning is always accompanied by conscious awareness, or is change possible without concomitant changes in subjective experience? More generally, is information processing in general accompanied by conscious experience, or can one find cases where action seems to be dissociated from introspective accounts?

If learning often fails to produce changes in subjective experience, examples of dissociation between action and consciousness should be abundant. As it turns out, they are. Indeed, everyday experience suggests that we often seem to know more than we can tell. Riding a bicycle, playing tennis, or driving a car, for instance, all involve mastering complex sets of motor skills, yet we are at a loss when it comes to explaining exactly how we perform such physical feats. These dissociations between our ability to report on cognitive processes and the behaviors that involve these processes are not limited to action but extend to higher-level cognition as well. Most native speakers of a language are unable to articulate the grammatical rules they nevertheless follow when speaking natural language. Likewise, expertise in domains such as medical diagnosis or chess, as well as social or aesthetic judgments, all involve intuitive knowledge that one seems to have little introspective access to.

We also often seem to tell more than we can know. In a classic article, social psychologists Nisbett and Wilson (1977) reported on many experimental demonstrations of the fact that accounts of our own behavior frequently reflect reconstructive and interpretative processes rather than genuine introspection.

Demonstrations of dissociations between conscious awareness and stimulation and/or behavior have now been reported in many domains of psychology, and in fact date back to the very beginnings of scientific psychology in the nineteenth century (e.g., Pierce & Jastrow, 1880).

One of the most convincing demonstrations of dissociations between conscious knowledge and behaviour was obtained in a eye-blink conditioning situation (Perruchet, 1985). In this experiment, people were exposed to a series of identical tones, 50% of which could be followed after a short interval by an air puff directed to the left cornea. Immediately after each tone was presented (and before the puff occurred in reinforced trials), people were asked to indicate the extent to which they expected the tone to be followed by an air puff on a 0–7-point scale. A trial-by-trial analysis of the results indicated that eye-blink responses were increasingly likely to occur after presentation of a tone if the corresponding trial had been preceded by a series of reinforced trials (i.e., trials during which the tone had indeed been followed by an air puff). In stark contrast, however, people's subjective expectancy of the occurrence of an air puff tended to *decrease* with the number of

reinforced trials that preceded the trial under consideration. In other words, people's eye-blink responses were completely dissociated from their conscious expectations about when each tone would be followed by an air puff.

Dissociations of a somewhat different kind can also be observed in situations that do not involve the motor system (see Cleeremans, Destrebecqz, & Boyer, 1998, for a review). Arthur Reber, in a classic series of studies conducted in 1967, first suggested that learning might be "implicit", to the extent that people appear to be able to learn new information without intending to do so and in such a way that the resulting knowledge is difficult to express. In Reber's seminal study of *artificial grammar learning* (Reber, 1967), subjects were asked to memorize a set of meaningless letter strings generated by a simple set of rules embodied in a finite-state grammar. After this memorization phase, they were told that the strings followed the rules of a grammar, and were asked to classify novel strings as grammatical or not. In this experiment and in many subsequent replications, subjects were able to perform this classification task better than chance would predict, despite remaining unable to describe the rules of the grammar in verbal reports. This dissociation between classification performance and verbal report is the finding that prompted Reber to describe learning as implicit, for subjects appeared sensitive to and could apply knowledge (the rules of the grammar) that they remained unable to describe and had had no intention to learn.

Today, another paradigm – sequence learning – has become dominant in the investigation of implicit learning. Nissen and Bullemer (1987) first demonstrated that subjects asked to respond as fast and as accurately as possible to a series of visual events progressively learned about the sequential structure of the stimulus sequence, in spite of showing little evidence of being aware that the material contained structure. Numerous subsequent studies of this effect have indicated that subjects can learn about complex sequential relationships despite remaining unable to fully express this knowledge in corresponding direct, explicit tasks (e.g., Cleeremans & McClelland, 1991), thus once again suggesting that learning can occur without corresponding changes in our ability to express the acquired knowledge explicitly.

Examples of dissociations between action, memory, perception on the one hand, and subjective experience on the other, are thus commonplace in many different domains. How can we account for these dissociations? How should we conceptualize the relationships between conscious and unconscious cognition? In the next section, I briefly overview traditional perspectives on these issues.

The function of consciousness: Commander Data meets the Zombies

In a recent overview article, Dehaene and Naccache (2001) conclude that "The present view associates consciousness with a unified neural workspace through which many processes can communicate. The evolutionary advantages that

this system confers to the organism may be related to the increased independence that it affords" (p. 31). Dehaene and Naccache thus suggest that consciousness allows organisms to free themselves from acting out their intentions in the real world, relying instead on less hazardous simulation made possible by the neural workspace. While I certainly agree with this conclusion, it begs the question of how consciousness came to play these functions in the first place. How can conscious states of the system come to reflect the most adaptive representation of the current situation, given that "what is most adaptive" continuously changes?

This complex, dynamical relationship between consciousness and learning, however, often tends to be overlooked in classical models of cognition. As argued in Cleeremans (1997) and also in Cleeremans and Jiménez (2002), this is most likely due to the fact that classical models of cognition (the "Computational Theory of Mind", see Fodor, 1975) take it as a starting point that *cognition is symbol manipulation*. If, however, one takes cognition to be exclusively and exhaustively about symbol manipulation, the only way to separate conscious from unconscious cognition is either (1) to assume that unconscious cognition is just like conscious cognition, only minus consciousness (Searle, 1992), or (2) to deny unconscious cognition altogether. Jiménez and I have dubbed these two (admittedly caricatural) perspectives "Zombie" and "Commander Data" models respectively.

Star Trek: The Next Generation's character Data is an android whose bodily and cognitive innards are fully transparent to himself. Except in rare circumstances, Data thus exhibits perfect episodic memory and perfect introspection. Commander Data theorists likewise assume that cognition is fully transparent – that is, (1) that whatever knowledge is expressed through behavior is also transparently available to introspection, and (2) that consciousness allows access, with sufficient effort or attention, to all aspects of our inner lives.

In contrast, philosophical zombies (Chalmers, 1996) are perfectly opaque, and in this sense instantiate absolutely implicit beings: Whatever internal knowledge currently influences their behavior can neither be explicit nor conscious because, by definition, they lack conscious experience. Zombie theorists thus take it as a starting point that consciousness has an epiphenomenal character: There is a zombie within you and, while you may not be aware of its existence, it could in fact be responsible for most of your actions. It is capable of processing all the information you can process in the same way that you do, with one crucial difference: "All is dark inside" (Chalmers, 1996, p. 96); your zombie is unconscious.

Needless to say, both of these perspectives are profoundly unsatisfactory. On the one hand, zombie perspectives (ZP) ascribe no role whatsoever to consciousness in information processing and – because it is absurd to deny consciousness altogether – are ultimately forced to assume the existence of equally powerful conscious and unconscious systems. On the other hand, Commander Data perspectives (CDP), by assuming that all of cognition is

conscious, paradoxically likewise depict consciousness as epiphenomenal. Crucially, both perspectives assume that consciousness does not change cognition in any principled way, and hence that *consciousness plays no functional role* beyond that of a epiphenomenon that accompanies either a functionally redundant subset of (ZP) or all (CDP) cognitive events.

In the face of the deeply counterintuitive flavor of both perspectives, it seems surprising to see that the past few years have witnessed the appearance of several broad theoretical proposals that intentionally or inadvertently endorse either of these perspectives. Thus for instance, Holender (1986), based on an extensive review of the subliminal perception literature, found no evidence for the existence of unconscious priming. Likewise, Shanks and St. John (1994) conclude their target article dedicated to implicit learning with the statement that "Human learning is almost invariably accompanied by conscious awareness" (p. 394). O'Brien and Opie (1999) propose that the contents of phenomenal consciousness include all stable neural states, and that it is only those stable states that are causally efficacious, that is, susceptible to influence further processing and, ultimately, behavior. Perruchet and Vinter (2002), consider that unconscious influences on behavior should be ascribed exclusively to noncognitive, neural processes and state that "Mental life [. . .] is co-extensive with consciousness" (p. 299, see also Dulany, 1997).

The debate is rooted not so much in equivocal empirical findings but rather in the deep conceptual problems associated with the notion of unconscious representation. Hence, defenders of the claim that cognition can be unconscious often succumb to some version of the ZP, while defenders of the opposite view can often be taken to endorse some variant of the CDP.

In the next section, I sketch out an alternative, subsymbolic, framework through which to think about the relationship between learning and consciousness – one that I believe offers a clear function to consciousness by linking it with adaptability in cognitive systems.

The framework

If the central assumption that the function of consciousness is to offer adaptive control over behavior is correct, then consciousness is necessarily closely related to processes of learning, because one of the central consequences of successful adaptation is that conscious control is no longer required over the corresponding behavior. Indeed, it might seem particularly adaptive for complex organisms to be capable of behavior that does not require conscious control, for instance because behavior that does not require monitoring of any kind can be executed faster or more efficiently than behavior that does require such control. Like Perruchet and Vinter (2003), I assume that there is a dynamic relationship between consciousness and learning such that (1) awareness of a particular state of affairs triggers learning and (2) this learning in turn changes the contents of subjective experience so as to make these contents more adapted.

I would now like to introduce the set of assumptions that together form the core of the framework (see Cleeremans, forthcoming; Cleeremans & Jiménez, 2002, for more detailed accounts). It is important to keep in mind that the framework is based on the connectionist framework (e.g., Rumelhart & McClelland, 1986). It is therefore based on many central ideas that characterize the connectionist approach, such as the fact that information processing is graded and continuous, and that it takes place over many interconnected modules consisting of processing units. In such systems, long-term knowledge is embodied in the pattern of connectivity between the processing units of each module and between the modules themselves, while the transient patterns of activation over the units of each module capture the temporary results of information processing.

This being said, a first important assumption is that *representations are graded, dynamic, active, and constantly causally efficacious*. Patterns of activation in neural networks and in the brain are typically distributed and can therefore vary on a number of dimensions, such as their stability in time, their strength. or their distinctiveness. *Stability* in time refers to how long a representation can be maintained active during processing. There are many indications that different neural systems involve representations that differ along this dimension. For instance, prefrontal cortex, which plays a central role in working memory, is widely assumed to involve circuits specialized in the formation of the enduring representations needed for the active maintenance of task-relevant information. *Strength* of representation simply refers to how many processing units are involved in the representation, and to how strongly activated these units are. As a rule, strong activation patterns will exert more influence on ongoing processing than weak patterns. Finally, *distinctiveness* of representation is inversely related to the extent of overlap that exists between representations of similar instances. Distinctiveness has been hypothesized as the main dimension through which cortical and hippocampal representations differ (McClelland, McNaughton, & O'Reilly, 1995; O'Reilly & Munakata, 2000), with the latter becoming active only when the specific conjunctions of features that they code for are active themselves.

In the following, I will collectively refer to these different dimensions as "quality of representation" (see also Farah, 1994). The most important notion that underpins these different dimensions is that representations, in contrast to the all-or-none propositional representations typically used in classical theories, instead have a *graded* character that enables any particular representation to convey the extent to which what it refers to is indeed present.

Another important aspect of this characterization of representational systems in the brain is that, far from being static propositions waiting to be accessed by some process, representations instead continuously influence processing regardless of their quality. This assumption takes its roots in McClelland's (1979) analysis of cascaded processing which, by showing how modules interacting with each other need not "wait" for other modules to

have completed their processing before starting their own, demonstrated how stage-like performance could emerge out of such continuous, non-linear systems. Thus, even weak, poor-quality traces are capable of influencing processing, for instance through associative priming mechanisms – that is, in *conjunction* with other sources of stimulation. Strong, high-quality traces, in contrast have *generative capacity*, in the sense that they can influence performance independently of the influence of other constraints – that is, whenever their preferred stimulus is present.

A second important assumption is that *learning is a mandatory consequence of information processing*. Indeed, every form of neural information processing produces adaptive changes in the connectivity of the system, through mechanisms such as Long Term Potentiation (LTP) or Long Term Depression (LTD) in neural systems, or hebbian learning in connectionist systems. An important aspect of these mechanisms is that they are mandatory in the sense that they take place whenever the sending and receiving units or processing modules are co-active. O'Reilly and Munakata (2000) have described hebbian learning as instantiating what they call *model learning*. The fundamental computational objective of such unsupervised learning mechanisms is to enable the cognitive system to develop useful, informative models of the world by capturing its correlational structure. As such, they stand in contrast with *task learning* mechanisms, which instantiate the different computational objective of mastering specific input–output mappings (i.e., achieving specific goals) in the context of specific tasks through error-correcting learning procedures.

Having put in place assumptions about representations and learning, the central ideas that I would now like to explore are (1) that the extent to which a particular representation is available to consciousness depends on its quality, (2) that learning produces, over time, higher-quality (and therefore adapted) representations, and (3) that the function of consciousness is to offer necessary control over those representations that are strong enough to influence behavior, yet not sufficiently adapted that their influence does not require control any more.

Figure 23.1 aims to capture these ideas by representing the relationships between quality of representation (X-axis) on the one hand and (1) potency, or the extent to which a representation can influence behavior, (2) availability to control, (3) availability to subjective experience. I discuss the figure at length in the following section. Let us simply note here that the X-axis represents a continuum between weak, poor-quality representations on the left and very strong, high-quality representations on the right.

Two further points are important to keep in mind with respect to Figure 23.1. First, the relationships depicted in the figure are intended to represent *availability* to some dimension of behavior or consciousness independently of other considerations. Many potentially important modulatory influences on the state of any particular module are thus simply not meant to be captured, neither by Figure 23.1, nor by the framework presented here. Second, the

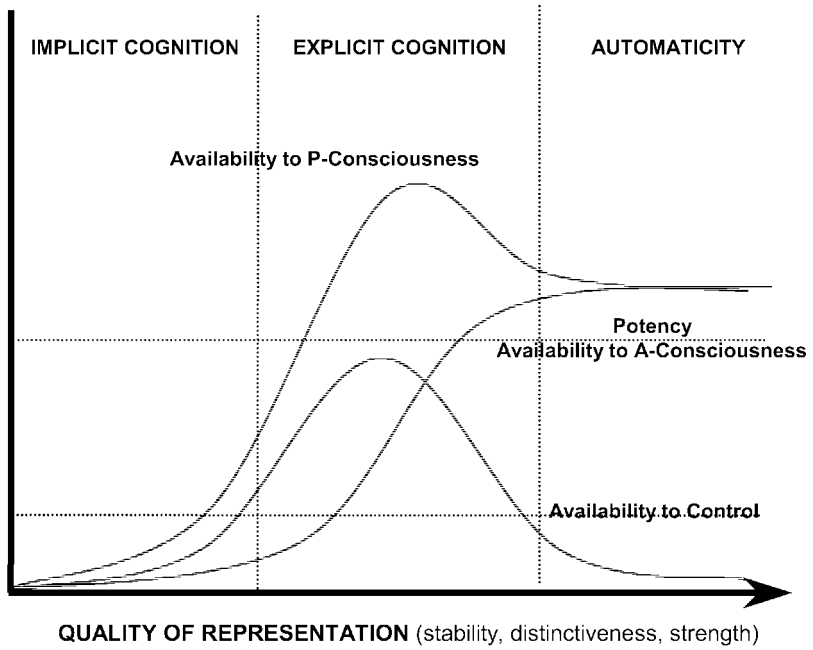

Figure 23.1 Graphical representation of the relationships between quality of representation (X-axis) and (1) potency, (2) availability to control, (3) availability to subjective experience. See text for further details.

figure is intended to represent what happens in *each* of many processing modules involved in any particular cognitive task. Thus, at any point in time, there will be many such modules active, each contributing to some extent to behavior and to conscious experience; each modulating the activity of other modules. With these caveats in mind, let me now turn to four assumptions about consciousness and its relationship with learning.

C1. Consciousness involves two dimensions: Subjective experience and control

As argued by Ned Block (1995, 2005) and even though there is continuing debate about this issue, consciousness involves at least two separable aspects, namely access consciousness (A-consciousness) and phenomenal consciousness (P-consciousness). According to Block (1995), "A perceptual state is access-conscious roughly speaking if its content – what is represented by the perceptual state – is processed via that information processing function, that is, if its content gets to the Executive system, whereby it can be used to control reasoning and behavior" (p. 234). In other words, whether a state is A-conscious is defined essentially by the causal efficacy of that state; the extent to which it is available for global control of action. Control refers to

the ability of an agent to control, to modulate, and to inhibit the influence of particular representations on processing. In this framework, control is simply a function of potency, as described in assumption C3. In contrast, P-consciousness refers to the phenomenal aspects of subjective experience: A state is P-conscious to the extent that there is something it is like to be in that state: I am currently experiencing a pain, hearing a beautiful piece of music, entertaining the memory of a joyful event. While the extent to which potency (i.e., availability to access consciousness) and control on the one hand, and subjective experience (i.e., availability to phenomenal consciousness) on the other, are dissociable is debatable, the framework suggests that potency, control and phenomenal experience are closely related to each other.

C2. Availability to consciousness correlates with quality of representation

This assumption is also a central one in this framework. It states that explicit, conscious knowledge involves higher-quality memory traces than implicit knowledge. "Quality of representation" designates several properties of memory traces, such as their relative strength in the relevant information-processing pathways, their distinctiveness, or their stability in time. The assumption is consistent with the theoretical positions expressed by several different authors over the last few years. O'Brien and Opie (1999) have perhaps been the most direct in endorsing a characterization of phenomenal consciousness in terms of the properties of mental representations in defending the idea that "consciousness equals stability of representation", that is, that the particular mental contents that one is aware of at some point in time correspond to those representations that are sufficiently stable in time. Mathis and Mozer (1996) have also suggested that consciousness involves stable representations, specifically by offering a computational model of priming phenomena in which stability literally corresponds to the state that a dynamic "attractor" network reaches when the activations of a subset of its units stop changing and settle into a stable, unchanging state.

A slightly different perspective on the notion of "quality of representation" is offered by authors who emphasize not stability, but strength of representation as the important feature through which to characterize availability to consciousness. One finds echoes of this position in the writings of Kinsbourne (1997), for whom availability to consciousness depends on properties of representations such as duration, activation, or congruence.

In Figure 23.1, I have represented the extent to which a given representation is available to the different components of consciousness (phenomenal consciousness, access-consciousness/potency, and control) as functions of a single underlying dimension expressed in terms of the quality of this representation. Availability to access-consciousness is represented by the curve labeled "potency", which expresses the extent to which representations can influence behavior as a function of their quality: High-quality, strong,

distinctive representations, by definition, are more potent than weaker representations and hence more likely to influence behavior. "Availability to control processes" is represented by a second curve, so labeled. We simply assume that both weak and very strong representations are difficult to control, and that maximal control can be achieved on representations that are strong enough that they can begin to influence behavior in significant ways, yet not so strong that they have become utterly dominant in processing. Finally, availability to phenomenal experience is represented by the third curve, obtained simply by adding the other two. The underlying intuition, discussed in the context of assumption C4, is that which contents enter subjective experience is a function of both availability to control and of potency.

C3. Developing high-quality representations takes time

This assumption states that the emergence of high-quality representations in a given processing module takes time, both over training or development, as well as during processing of a single event. Figure 23.1 can thus be viewed as representing not only the relationships between quality of representation and their availability to the different components of consciousness, but also as a depiction of the dynamics of how a particular representation will change over the different time scales corresponding to development, learning, or within-trial processing.

Both skill acquisition and development, for instance, involve the long-term progressive emergence of high-quality, strong memory traces based on early availability of weaker traces. Likewise, the extent to which memory traces can influence performance at any moment (e.g., during a single trial) depends both on available processing time, as well as on overall trace strength. These processes of change operate on the connection weights between units, and can involve either task-dependent, error-correcting procedures, or unsupervised procedures such as hebbian learning. In either case, continued exposure to exemplars of the domain will result in the development of increasingly congruent and strong internal representations that capture more and more of the relevant variance. Although I think of this process as essentially continuous, three stages in the formation of such internal representations (each depicted as separate regions in Figure 23.1 can be distinguished: Implicit representations, explicit representations, and automatic representations.

The first region, labeled "Implicit Cognition" in Figure 23.1, is meant to correspond to the point at which processing starts in the context of a single trial, or to some early stage of development or skill acquisition. In either case, this stage is characterized by weak, poor-quality representations. A first important point is that representations at this stage are already capable of influencing performance, as long as they can be brought to bear on processing together with other sources of constraints, that is, essentially through mechanisms of associative priming and constraint satisfaction. A second

important point is that this influence is best described as "implicit", because the relevant representations are too weak (i.e., not distinctive enough) for the system as a whole to be capable of exerting control over them: You cannot control what you cannot identify as distinct from something else.

The second region of Figure 23.1 corresponds to the emergence of explicit representations, defined as representations over which one can exert control. In the terminology of attractor networks, this would correspond to a stage during learning at which attractors become better defined – deeper, wider, and more distinctive. It is also at this point that the relevant representations acquire generative capacity, in the sense that they now have accrued sufficient strength to have the potential to determine appropriate responses when their preferred stimulus is presented to the system alone. Because awareness is partially tied to control in this framework, one would thus also be aware both of these internal representations and of their influence on our behavior. Because one is aware of these representations, one can then also possess metaknowledge about them, and recode them in various different ways, for instance, as linguistic propositions.

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 controlled (i.e., inhibited). Such representations exert a mandatory influence on processing. Importantly, however, one is aware both of possessing them (that is, one has relevant metaknowledge) and of their influence on processing (see also Tzelgov, 1997), because availability to conscious awareness depends on the quality of internal representations, and that strong representations are of high quality. In this perspective then, one can always be conscious of automatic behavior, but not necessarily with the possibility of control over these behaviors.

In this framework, skill acquisition and development therefore involve a continuum at both ends of which control over representations is impossible or difficult, but for very different reasons: Implicit representations influence performance but cannot be controlled because they are not yet sufficiently distinctive and strong for the system to even know it possesses them. This might in turn be related to the fact that, precisely because of their weakness, implicit representations cannot influence behavior on their own, but only in conjunction with other sources of constraints. Automatic representations, on the other hand, cannot be controlled because they are too strong, but the system is aware both of their presence and of their influence on performance.

C4. The function of consciousness is to offer flexible, adaptive control over behavior

The framework gives consciousness a central place in information processing, in the sense that its function is to enable flexible control over behavior. Crucially, however, consciousness is not necessary for information processing, or for adaptation in general, thus giving a place for implicit learning

in cognition. I believe this perspective to be congruent with theories of adaptation and optimality in general.

Indeed, another way to think about the role of learning in consciousness is to ask: "When does one *need* control over behavior?". Control is perhaps not necessary for implicit representations, for their influence on behavior is necessarily weak (in virtue of the fact that precisely because they are weak, such representations are unlikely to be detrimental to the organism even if they are not particularly well adapted). Likewise, control is not necessary for automatic representations, because presumably, those representations that have become automatic after extensive training should be *adapted* (optimal) as long as the processes of learning that have produced them can themselves be assumed to be adaptive. Automatic behavior is thus necessarily optimal behavior in this framework, except, precisely, in cases such as addiction, obsessive-compulsive behavior, or laboratory situations where the automatic response is manipulated to be non-optimal, such as in the Stroop situation. Referring again to Figure 23.1, my analysis therefore suggests that the representations that most require control are the explicit representations that correspond to the central region of Figure 23.1: Representations that are strong enough that they have the potential to influence behavior in and of themselves (and hence that one should really care about, in contrast to implicit representations), but not sufficiently strong that they can be assumed to be already adapted, as is the case for automatic representations. It is for those representations that control is needed, and, for this reason, it is these representations that one is most aware of.

Likewise, this analysis also predicts that the dominant contents of subjective experience at any point in time consist precisely of those representations that are both strong enough that they can influence behavior, yet weak enough that they still require control. Figure 23.1 reflects these ideas by suggesting that the contents of phenomenal experience depend on the potency of currently active representations as well as on their availability to control. Since availability to control is inversely related to potency for representations associated with automatic behavior, this indeed predicts weaker availability to phenomenal experience of "very strong" representations as compared to "merely strong" representations. In other words, such representations can become conscious if appropriate attention is directed towards their contents – as in cases where normally automatic behavior (such as walking) suddenly becomes conscious because the normal unfolding of the behavior has been interrupted (e.g., because I've stumbled upon something) – but they are not normally part of the central focus of awareness nor do they require conscious control.

The framework thus leaves open four distinct possibilities for knowledge to be implicit. First, knowledge that is embedded in the connection weights within and between processing modules can never be directly available to conscious awareness and control. This is simply a consequence of the fact

that I assume that consciousness necessarily involves representations (patterns of activation over processing units). The knowledge embedded in connection weights will, however, shape the representations that depend on it, and its effects will therefore be detectable – but only indirectly, and only to the extent that these effects are sufficiently marked in the corresponding representations.

Second, to enter conscious awareness, a representation needs to be of sufficiently high quality in terms of strength, stability in time, or distinctiveness. Weak representations are therefore poor candidates to enter conscious awareness. This, however, does not necessarily imply that they remain causally inert, for they can influence further processing in other modules, even if only weakly so.

Third, a representation can be strong enough to enter conscious awareness, but fail to be recognized as relevant to the particular situation that is currently unfolding. Conscious contents, indeed, have to be linked together in a coherent manner before they can be made available globally for conscious report and for the control of action. There are thus many opportunities for a particular conscious content to remain implicit, not because its representational vehicle does not have the appropriate properties, but because it fails to be integrated with other conscious contents. Dienes and Perner (2002) offer an insightful analysis of the different ways in which what I have called high-quality representations can remain implicit.

Finally, a representation can be so strong that its influence can be no longer be controlled. In these cases, it is debatable whether the knowledge should be taken as genuinely unconscious, because it can certainly become fully conscious as long as appropriate attention is directed to it, but the point is that such very strong representations can trigger and support behavior without conscious intention and without the need for conscious monitoring of the unfolding behavior.

Conclusions

In this article I have attempted to outline a framework that offers a clear functional role to consciousness by linking conscious awareness with adaptation in general, and with learning in particular. I have argued that if we take consciousness as the only mechanism through which flexible control can be achieved over action, then it follows that learning should be the most important factor that determines the contents of conscious experience. Learning thus shapes consciousness, and consciousness, in turn, reflects the adapted appreciation of the dynamics of the current situation that is necessary to make flexible control over action possible (Perruchet & Vinter, 2003).

The framework does not assume that there exists a strong distinction between conscious and non-conscious aspects of cognition. Rather, it assumes that conscious and unconscious aspects of cognition are simply that – *aspects* of a single ensemble of learning mechanisms and of representational systems.

The analysis presented above resonates well with recent computational

models of overall cerebral function. O'Reilly and colleagues (Atallah, Frank, & O'Reilly, 2004; McClelland et al., 1995; O'Reilly & Munakata, 2000), for instance, have recently proposed that different regions of the brain have evolved to solve different – and incompatible – computational problems by using different representational formats and different learning regimes (McClelland et al., 1995). In their "tripartite" proposal, the brain is organized in three broad interacting systems: the hippocampus (HC), prefrontal cortex/basal ganglia (FC), and posterior cortex (PC). In this framework, each system uses similar, but not identical, learning mechanisms and representational formats. The main function of the hippocampus is to rapidly learn about specific novel facts (episodic memory). Posterior cortex's function, in contrast, is to learn about the statistical regularities shared by many exemplars of a given domain (semantic memory). Finally, the main function of frontal cortex is to maintain information in an active state (active maintenance, subtending working memory) and to rapidly switch between active representations. Achieving each of these functions requires different (but germane) learning mechanisms and different representational formats. Thus, the hippocampus uses the sparse, conjunctive representations necessary to avoid catastrophic interference, and a high learning rate that makes it possible to rapidly bind together the various elements of the current percept. Posterior cortex, in contrast, slowly accumulates information over largely overlapping, distributed representations, so that broad semantic knowledge can progressively emerge over learning and development. Finally, frontal cortex is characterized by self-sustaining representational systems involving the recurrent connectivity necessary for active maintenance as well as the gating mechanisms necessary for rapid switching.

The three systems also differ from each other in terms of processing and learning mechanisms. Thus, O'Reilly and Munakata (2000) argue that the functions typically attributed to frontal cortex (i.e., working memory, inhibition, executive control, and monitoring) require "activation-based processing", characterized by mechanisms of active maintenance through which representations can remain strongly activated for long periods of time as well as rapidly updated so as to make it possible for these representations to modulate processing elsewhere in the brain. Note how this is consistent with Crick and Koch's (2003) notion that the front of the brain is looking at the back. Because of these properties, frontal representations are thus more accessible to verbalization and other reporting systems. To this, they oppose "weight-based processing", characteristic of posterior cortex, in which knowledge is encoded directly by the pattern of connectivity between processing units and hence tends to remain tacit to the extent that this knowledge only manifests itself through the effects it exerts on ongoing processing rather than through the form of representations themselves.

Finally, one might wonder about an essential aspect of consciousness that I have not discussed at all in the context of this article: Who, or what, is doing the controlling? How are we to conceptualize the self in this framework? I

believe that the framework makes it possible to cast that story without resorting to a homunculus, but explaining how is clearly beyond the goals of this article. Suffice it to say here that while the framework indeed involves something of a homunculus, this homunculus need not be a real one but can instead itself be viewed as emergent.

Acknowledgements

Axel Cleeremans is a Senior Research Associate with the National Fund for Scientific Research (FNRS, Belgium). This work was supported by an institutional grant from the Université Libre de Bruxelles to Axel Cleeremans and by grant HPRN-CT-1999-00065 from the European Commission. The FNRS and the Ministère de la Communauté Française de Belgique both contributed to make my participation to the congress possible. Significant portions of this article were adapted from the following publications: (1) Cleeremans, A. & Jiménez, L. (2002). Implicit learning and consciousness: A graded, dynamic perspective. In R.M. French & A. Cleeremans (Eds.), *Implicit learning and consciousness: An empirical, computational and philosophical consensus in the making*? Hove, UK: Psychology Press. (2) Cleeremans, A. (2002). Handlung und Bewusstsein: Ein Rahmenkonzept für den Fertigkeitserwerb. *Psychologie und Sport*, *9*, 2–19. (3) Cleeremans, A., (2005). Computational correlates of consciousness. In S. Laureys (Ed.), *Progress in Brain Research* (Vol. 150, pp. 81–98). Amsterdam: Elsevier.

References

Atallah, H., Frank, M.J., & O'Reilly, R.C. (2004). Hippocampus, cortex, and basal ganglia: Insights from computational models of complementary learning systems. *Neurobiology of Learning and Memory*, *82*, 253–267.

Block, N. (1995). On a confusion about a function of consciousness. *Behavioral and Brain Sciences*, *18*, 227–287.

Block, N. (2005). Two neural correlates of consciousness. *Trends in Cognitive Science*, *9*(2), 46–52.

Chalmers, D.J. (1996). *The conscious mind: In search of a fundamental theory*. Oxford: Oxford University Press.

Cleeremans, A. (1997). Principles for implicit learning. In D.C. Berry (Ed.), *How implicit is implicit learning?* (pp. 195–234). Oxford: Oxford University Press.

Cleeremans, A. (forthcoming). *Being virtual*. Oxford: Oxford University Press.

Cleeremans, A., Destrebecqz, A., & Boyer, M. (1998). Implicit learning: News from the front. *Trends in Cognitive Sciences*, *2*, 406–416.

Cleeremans, A., & Jiménez, L. (2002). Implicit learning and consciousness: A graded, dynamic perspective. In R.M. French & A. Cleeremans (Eds.), *Implicit learning and consciousness: An empirical, computational and philosophical consensus in the making?* (pp. 1–40). Hove, UK: Psychology Press.

Cleeremans, A., & McClelland, J.L. (1991). Learning the structure of event sequences, *Journal of Experimental Psychology*: General, *120*, 235–253.

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

Dehaene, S., & Naccache, L. (2001). Towards a cognitive neuroscience of consciousness: Basic evidence and a workspace framework. *Cognition*, *79*, 1–37.

Dienes, Z., & Perner, J. (2002). A theory of the implicit nature of implicit learning. In R.M. French & A. Cleeremans (Eds.), *Implicit learning and consciousness: An empirical, computational and philosophical consensus in the making?* (pp. 68–92). Hove, UK: Psychology Press.

Dulany, D.E. (1997). Consciousness in the explicit (deliberative) and implicit (evocative). In J.D. Cohen & J.W. Schooler (Eds.), *Scientific approaches to consciousness* (pp. 179–212). Mahwah, NJ: Lawrence Erlbaum Associates Inc.

Elbert, T., Pantey, C., Wienbruch, C., Rockstroh, B., & Taub, E. (1995). Increased cortical representation of the fingers of the left hand in string players. *Science*, *270*, 305–307.

Farah, M.J. (1994). Visual perception and visual awareness after brain damage: A tutorial overview. In C. Umiltà & M. Moscovitch (Eds.), *Attention and Performance XV: Conscious and nonconscious information processing* (pp. 37–76). Cambridge, MA: MIT Press.

Farah, M.J. (1998). Why does the somatosensory homunculus have hands next to face and feet next to genitals: A hypothesis. *Neural Computation*, *10*(8), 1983–1985.

Fodor, J.A. (1975). *The language of thought*. New York: Harper & Row.

Holender, D. (1986). Semantic activation without conscious activation in dichotic listening, parafoveal vision, and visual masking : A survey and appraisal. *Behavioral and Brain Sciences*, *9*, 1–23.

Kinsbourne, M. (1997). What qualifies a representation for a role in consciousness? In J.D. Cohen & J.W. Schooler (Eds.), *Scientific approaches to consciousness* (pp. 335–355). Mahwah, NJ: Lawrence Erlbaum Associates Inc.

Maguire, E.A., Gadian, D.G., Johnsrude, I.S., Good, C.D., Ashburner, J., Frackowiak, R.S. et al. (2000). Navigation-related structural change in the hippocampi of taxi drivers. *Proceedings of the National Academy of Sciences of the USA*, *10*, 1073.

Mathis, W.D., & Mozer, M.C. (1996). Conscious and unconscious perception: A computational theory. In *Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society* (pp. 324–328). Hillsdale, NJ: Lawrence Erlbaum Associates Inc.

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., McNaughton, B.L., & O'Reilly, R.C. (1995). Why there are complementary learning systems in the hippocampus and neocortex: Insights from the successes and failures of connectionist models of learning and memory. *Psychological Review*, *102*, 419–457.

Nisbett, R.E., & Wilson, T.D. (1977). Telling more than we can do: Verbal reports on mental processes. *Psychological Review*, *84*, 231–259.

Nissen, M.J., & Bullemer, P. (1987). Attentional requirement of learning: Evidence from performance measures. *Cognitive Psychology*, *19*, 1–32.

O'Brien, G., & Opie, J. (1999). A connectionist theory of phenomenal experience. *Behavioral and Brain Sciences*, *22*, 175–196.

O'Reilly, R.C., & Munakata, Y. (2000). *Computational explorations in cognitive neuroscience: Understanding the mind by simulating the brain*. Cambridge, MA: MIT Press.

Perruchet, P. (1985). A pitfall for the expectancy theory of human eyelid conditioning. *Pavlovian Journal of Biological Science*, *20*, 163–170.

Perruchet, P., & Vinter, A. (2002). The self-organizing consciousness. *Behavioral and Brain Sciences*, *25*(3), 297–330.

Pierce, C.S., & Jastrow, J. (1880). On small differences in sensation. *Memoirs of the National Academy of Sciences*, *3*, 75–83.

Reber, A.S. (1967). Implicit learning of artificial grammars. *Journal of Verbal Learning and Verbal Behavior*, *5*, 855–863.

Rumelhart, D.E., & McClelland, J.L. (1986). *Parallel distributed processing: Explorations in the microstructure of cognition. Volume 1: Foundations*. Cambridge, MA: MIT Press.

Searle, J.R. (1992). *The rediscovery of the mind*. Cambridge, MA: MIT Press.

Shanks, D.R., & St. John, M.F. (1994). Characteristics of dissociable human learning systems. *Behavioral and Brain Sciences*, *17*, 367–447.

Tzelgov, J. (1997). Automatic but conscious: That is how we act most of the time. In R.S. Wyer (Ed.), *The automaticity of everyday life* (Vol. X, pp. 217–230). Mahwah, NJ: Lawrence Erlbaum Associates Inc.

Zhu, X., & Simon, H.A. (1987). Learning mathematics by examples and by doing. *Cognition and Instruction*, *4*, 137–166.