14

Reinforcement Learning, High-Level Cognition, and the Human Brain

Massimo Silvetti and Tom Verguts
Ghent University
Belgium

1. Introduction

Reinforcement learning (RL) has a rich history tracing throughout the history of psychology. Already in the late 19th century Edward Thorndike proposed that if a stimulus is followed by a successful response, the stimulus-response bond will be strengthened. Consequently, the response will be emitted with greater likelihood upon later presentation of that same stimulus. This proposal already contains the two key principles of RL. The first principle concerns *associative learning*, the learning of associations between stimuli and responses. This theme was developed by John Watson. Building on the work of Ivan Pavlov, John Watson investigated the laws of classical conditioning, in particular, how a stimulus and a response become associated after repeated pairing. In the classical "Little Albert" experiment, Watson and Rayner (1920) repeatedly presented a rabbit together with a loud sound to the kid (little Albert); the rabbit initially evoked a neutral response, the loud sound initially evoked a fear response. After a while, also presentation of the rabbit alone evoked a fear response in the subject. In this same paper, the authors proposed that this principle of learning by association more generally is responsible for shaping (human) behavior. According to psychology handbooks John Watson hereby laid the foundation for behaviorism. The second principle is that *reinforcement* is key for human learning. Actions that are successful for the organism, will be strengthened and therefore repeated by the organism. This aspect was developed into a systematic research program by the second founder of behaviorism, Burrhus Skinner (e.g., Skinner, 1938).

The importance of RL for explaining human behavior started to be debated from the late 1940s. Scientific criticism toward RL arrived from two main fronts. The first was internal, deriving from experimental findings and theoretical considerations within psychology itself. The second derived from external developments, in particular, advancements in information theory and control theory. These criticisms led to a disinterest for RL lasting several decades. However, in recent years, RL has been revived, leading to a remarkable interdiscplinary confluence between computer science, neurophysiology, and cognitive neuroscience. In the current chapter, we describe the relevant mid-20th century criticisms and developments, and how these were considered and integrated in current versions of RL. In particular we focus on how RL can be used as a model for understanding high-level cognition. Finally, we link RL to the broader framework of neural Darwinism.

2. Internal criticisms to RL

During the 40s, more and more data were piling up demonstrating the insufficiency of behaviorism to account for human and animal behavior. For example, Tolman and colleagues showed that animals can and do learn even without obtaining reinforcement (Tolman, 1948). They performed a series of experiments on maze learning in rats. It was shown that animals left free to familiarize themselves with the maze before the reinforcement experimental session, were afterwards able to find the food in the maze much more efficiently than completely naive animals. To explain these findings Tolman introduced the concept of the "cognitive map", i.e. an internal representation of the maze that the rats used to find reinforcers more efficiently. Because of this and other demonstrations that animals hold some kind of internal representation of the environment (memory), Tolman formed part of what became known as "the cognitive revolution".

During the same period, but in the field of psychobiology, Donald Hebb wrote *The Organization of Behaviour* (1949), a seminal work in which for the first time a neurobiological theory of learning was proposed. Hebb suggested that the synaptic connection between two neurons improves its efficacy after repeated simultaneous activity of them. This law, properly called "Hebbian rule" and describing what was called "Hebbian Learning", provided the first neural hypothesis on the basis of memory, thus opening the "black box", which behaviorists considered not scientifically investigable. The depth of Hebb's intuition can be better understood if we consider that the Hebbian rule has been experimentally proven almost twenty years after its formulation, with the discover of synaptic long term potentiation (LTP) in the rabbit hippocampus (Lømo, 1966).

Another strong criticism came from psycholinguistics. In a famous review study, Noam Chomsky (1959) argued that the RL paradigm was not suitable to explain the generative feature of natural language (i.e. the possibility to express a quasi-infinite variety of verbal expressions). In the same work, Chomsky also provided a survey on research in animal behavior (e.g., imprinting) that seemed to be in striking contrast with key behaviorist tenets. Finally, and most importantly from the theoretical point of view, Chomsky showed that Skinner himself was obliged to introduce hypotheses about internal variables (e.g., internal self-reinforcements), in order to explain human verbal behavior.

3. External developments

An important role in the demise of RL derived from advances in information theory and control theory in engineering. This happened during the 1940-50s with the publication of several seminal works like those of Shannon (1948), Turing (1936) and Wiener (1948). Their importance consisted in showing that it was possible to formulate rigorous mathematical theories and models to study information processing. In control theory (Wiener, 1948), for instance, the term "control" referred to the auto-correction of internal parameters of a system based on a feedback signal indicating the error between the wished (or the expected) value of an internal parameter and its real value, typically provided by the environment. This general theory of control (called cybernetics) (literally from ancient Greek: "the art of piloting"), did not refer to a particular system: instead, it provided mathematical models to study control phenomena occurring *inside* any system, being animal or artificial or even social. A similar story holds for information theory (Shannon, 1948), which provided the concept of "information", a measure that did not refer to any directly measurable physical variable, but instead to the *internal* "surprise" of any system receiving an external signal.

These new disciplines showed that it was possible, and indeed a proficient and powerful approach, to investigate the internal functioning of systems (including biological organisms), by mathematical modelling of their hidden machinery that was not directly investigable. In this way, the philosophical-methodological assumption of behaviorism, according to which the scientific approach should be limited to strictly empirical investigation, was shown to be unnecessary for scientific progress.

4. Precursors to the return of RL

Because of these developments, behaviorism, and with it RL, was discredited for several decades. Instead an alternative paradigm became dominant, according to which the human mind could be construed as a computer that manipulates abstract symbols (e.g., Neisser, 1967; Atkinson and Shiffrin, 1968). However, in recent years the RL framework became influential again. At least two developments in the second part of the 20th century prepared a renewed interest for RL. The first originated in human learning theory; the second from a new discipline called connectionist psychology, which proposed itself as an alternative to the then canonical symbol-manipulation paradigm for the study of cognition.

4.1 Human learning theory and the Rescorla-Wagner model

Important phenomena observed in the behavioral lab could not be accounted for with the standard behaviorist conceptualization (Rescorla and Wagner, 1972). For example, blocking (Kamin, 1969) refers to the fact that an organism only learns about the contingency between two events to the extent that one of the events is unexpected. To account for blocking, Rescorla and Wagner added a crucial ingredient to an associative learning framework, namely prediction error. Prediction error refers to the difference between an external feedback signal indicating the correct response or stimulus on the one hand, and the response or stimulus predicted by the organism on the other. Here it is worth noting the influence (and indeed similarity) of the cybernetic concept of feedback on the formulation of the concept of prediction error. Rescorla and Wagner proposed a formal model which learned by updating associations between events (e.g., stimulus and response) using prediction error (Rescorla and Wagner, 1972). This model formed the basis for many human learning theories (e.g., Kruschke, 2008; Pearce and Hall, 1980; Van Hamme and Wasserman, 1994), and can be represented by the following equations:

$$\delta_t = \lambda_t - V_t \tag{1}$$

$$V_{t+1} = V_t + \alpha\delta_t \tag{2}$$

where δ is the prediction error, V is the prediction of the organism, and λ is the actual outcome from the environment. Equation 2 shows how the new expectations are updated by the prediction error from time point t to $t + 1$; α is a learning rate parameter modulating the prediction error.

4.2 The connectionist approach

A second development preparing the cultural ground for reviving the field of RL was connectionist psychology. Here, the study of psychological phenomena was grounded on the construction of artificial neural networks, i.e. models simulating both the nervous

system and cognitive processes, providing what was called a sub-symbolic explanation of cognition. This new field was inspired by the fast developing neurosciences; in particular, the scientists developing this new branch not only did not adhere to the dogma that theorizing should remain at the behavioral level, but they also attempted to bridge the explanatory gap between the biological level of neurons and synapses on the one hand, and the psychological level of language and other forms of high-level cognition on the other.
An important step was taken by McClelland, Rumelhart and colleagues (Rumelhart and McClelland, 1986). Models similar to theirs had been developed by other researchers before (Grossberg, 1973) but Rumelhart and McClelland developed a series of applications that made these connectionist models almost instantly influential. At the core of these models is again the Rescorla-Wagner idea that learning consists of updating associations based on prediction errors. However, the authors proposed a generalized learning rule (backpropagation), which allowed learning also for so-called "hidden units", that is, neurons that do not receive external feedback. In backpropagation, such neurons use as a prediction error a linear combination of prediction errors of other neurons that do receive external feedback. This development made the learning rule many orders more powerful than that of Rescorla and Wagner. With the more powerful learning rule, the connectionists were able to investigate linguistic phenomena such as past tense formation (Rumelhart and McClelland, 1987), naming aloud (Seidenberg and McClelland, 1989), and sentence comprehension (St. John and McClelland, 1990).

4.3 The new RL approach

With these important historical precedents, RL learning became influential again during the early 1990s partly because of its important contributions to Machine Learning, a branch of Artificial Intelligence. One of the main protagonists of this revival was Richard Sutton, who developed another generalization of the Rescorla-Wagner rule, called temporal-difference (TD) learning (Sutton, 1988). The original Rescorla-Wagner rule had a *spatial* limitation in the sense that not all neurons received feedback, and this problem was solved by backpropagation. Similarly, the Rescorla-Wagner rule also has a *temporal* limitation in the sense that feedback is not always available to the model – only when there is explicit supervisory feedback. The TD learning algorithm solved this latter problem, because it allowed learning by not only comparing a prediction with external feedback (which may or may not be available, depending on an appropriate teacher's availability), but additionally by comparing a prediction with an earlier prediction (which is always available). In this case the learning signal is the TD error (here denoted as δ^{TD}), in which both the comparisons between previous prediction and external feedback and previous prediction and current prediction play a role. The TD error signal can be written as follows:

$$\delta_{t+1}^{TD} = \lambda_{t+1} + \gamma V_{t+1} - V_t \tag{3}$$

where λ is the external feedback already defined in Equation (1) and γ is a discount factor. The symbol V was used before to denote the organism's prediction; in RL applications, it refers specifically to reward prediction. This rule is more powerful than the Rescorla-Wagner rule: For example, Tesauro (1989) demonstrated that a neural network equipped with TD learning can learn to play backgammon at a worldmaster level.

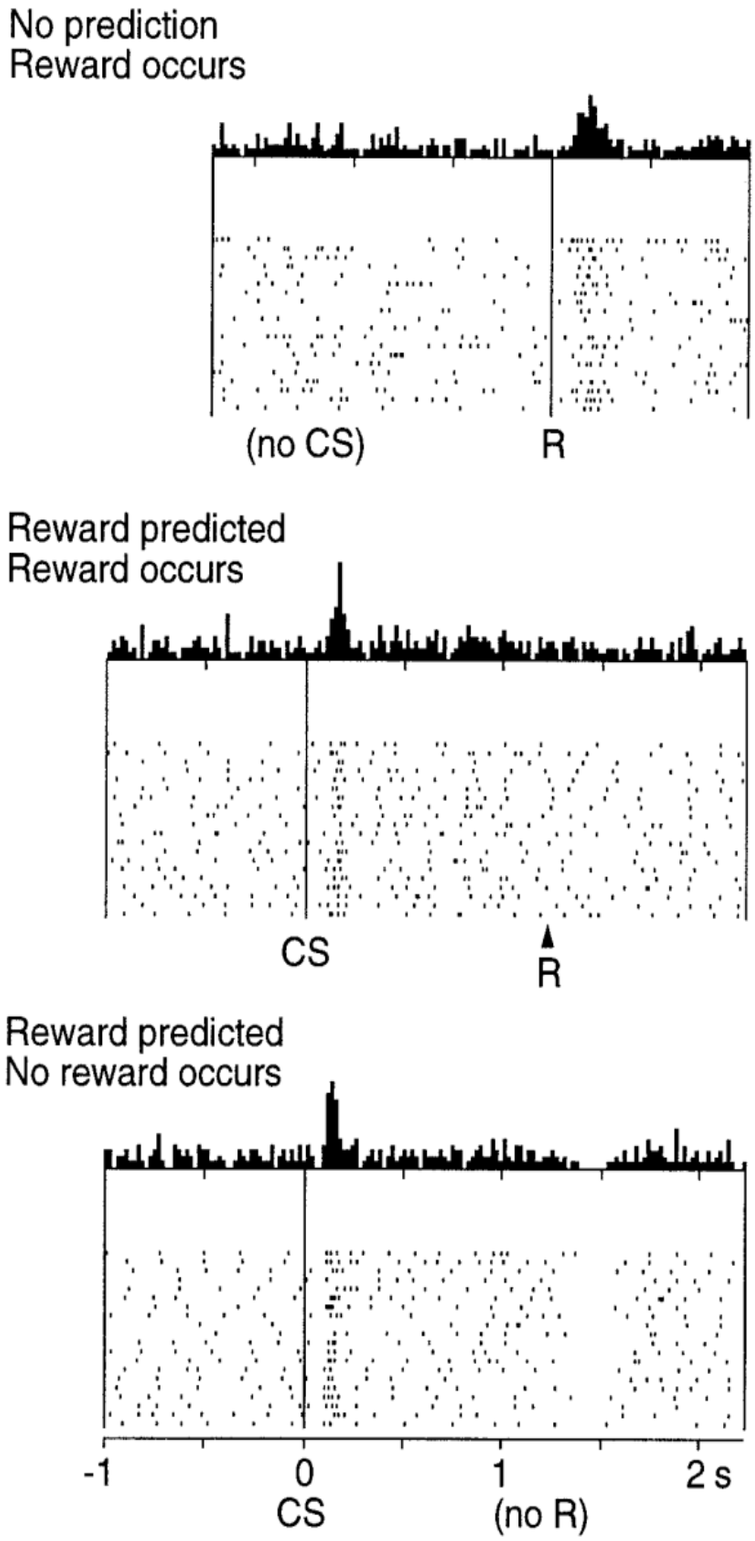


Fig. 1. Shifting of dopaminergic activity from reward to predictor of reward (CS). Reprinted with permission from Schultz et al. (1997).

A few years later, the RL paradigm received the decisive boost to come back to the attention of the broad scientific community. This derived from its official entrance into the domain of neurophysiology. In particular, with single-unit recording Wolfram Schulz and colleagues discovered dopaminergic neurons in the brainstem ventral tegmental area (VTA) and substantia nigra (SN) of macaque monkeys that exhibited a prediction error signature. In a classical conditioning experiment, Schultz et al. (1993) presented a conditioned stimulus (CS, e.g. a light), followed by an unconditioned stimulus (US, e.g. a drop of juice) some seconds later. Initially, dopaminergic neurons respond to the US only. After some trials, the dopaminergic neurons respond to the CS, but no longer to the US (Figure 1). This backward shift in time is exactly what was predicted by TD learning (Montague et al., 1996). Hence, this strongly suggested that the mammalian nervous system implements a RL (in particular, TD) algorithm to learn associations between stimuli.

5. RL in high-level cognition: Conceptual and empirical advances

Ever since the seminal findings of Schultz et al. (1993), the marriage between neuroscience and RL never stopped providing benefits for the study of learning and the nervous system. We here discuss a few highlights from the recent literature.

One conceptual development of RL consisted in the discovery that besides reward value, other value dimensions can be estimated and used to discount reward value (e.g., effort, Kennerley et al., 2006, or delay, Rudebeck et al., 2006). More generally, not only value but also upcoming states of the world can be estimated (Sutton and Barto, 1998). This allows the organism to make more far-sighted actions than with immediate values estimates only. Further, RL models have been proposed with the same computational power as the benchmark backpropagation algorithm (O'Reilly & Frank, 2004; Roelfsema and Van Ooyen, 2005), providing a biologically plausible alternative to backpropagation.

At the empirical level, clever experimental paradigms in combination with modern imaging technology allowed demonstrating the validity of RL models for human cognition. Using fMRI, Seymour et al. (2004) identified a TD signal in the human brain, similar to what was found by Schultz and colleagues in the monkey brain. Seymour et al. used a cued pain learning paradigm, in which a first CS (CS1) predicted (statistically) a second CS (CS2), which then (deterministically) predicted the upcoming pain level. In the striatum (ventral putamen), they observed a pain prediction error signal which responded to CS1 onset, and to CS2 if it differed from CS1 (i.e., was unpredicted based on CS1). Similar paradigms were used using appetitive learning (O'Doherty et al., 2003). The TD learning framework has also been applied extensively to EEG data, in particular the error-related negativity, for example in the work of Holroyd and Coles (2002). These authors successfully compared the performance of a TD learning-based computational model with the dynamics of the error-related negativity (ERN) from human volunteers. The model was aimed to clarify the roles of anterior cingulate cortex (ACC) and the ventral striatal structures in an instrumental conditioning paradigm. The authors proposed that the ventral striatum implements TD learning in order to estimate the value of external stimuli in terms of expected reward, while the ACC functions as a filter of several possible motor responses. In their proposal, the ACC would select the motor plans that are expected to be the most effective to achieve future rewards, based on the reward predictions computed by the ventral striatum. In this model, the ERN would be the result of ACC activity following the suppression of dopaminergic input from the ventral striatum. In a series of EEG experiments, Holroyd and Coles showed

that their model was indeed able to predict several effects linked to the ERN, for example the fact that this EEG component appears only when there is a violation of the reward prediction.

Another example comes from the study of Parkinson's disease, a neurological disorder whose pathological basis consists of the degeneration of the dopaminergic neurons in the substantia nigra pars compacta, source of the main brainstem input to basal ganglia. Parkinson patients are impaired in learning from positive outcomes (reward), while performance is preserved for learning based on negative outcomes (punishment) (Frank et al., 2004). A neural model was proposed representing the interactions between basal ganglia, cortex and substantia nigra (Frank, 2005). In this model, the basal ganglia consists of two neural populations; "Go" neurons fire when an action planned in cortex is allowed to be implemented, whereas "No Go" neurons suppress the action planned in the cortex. Both Go and No Go populations learn by dopaminergic (i.e., reinforcement-related) bursts and dips coming from the substantia nigra. One of the advantages of the model consists in explaining several symptoms of Parkinson's disease. For instance, with reduced dopaminergic input (simulation of the substantia nigra degeneration), the basal ganglia are impaired at learning in Go neural populations, and hence impaired specifically in learning by rewards, just like human Parkinson patients. In addition, the model successfully predicts that this distinction between Go versus No Go learning holds true in high-level cognition as well (Frank et al., 2004).

Finally, it is worth describing briefly the work of Gläscher et al. (2010), which showed, by a combined computational and fMRI study, that the human brain also implements RL-like algorithms for creating abstract models of the environment. This study resembled the historical experiment of Tolman (1948). Volunteers were at first exposed to a simplified artificial environment, in which each single state was represented by an abstract figure (a fractal) (Figure 2). The subjects were asked to "navigate" inside this environment by performing binary choices (left or right). Each choice was followed by a transition to one of two possible states, each with some probability. In the first part of the experiment, subjects freely navigated in this environment, resembling Tolman's latent learning phase. In the second part, subjects received a monetary reward in some of the final states. In this way they had to exploit the latent learning acquired during the first part of the experiment to maximize reward, again as in Tolman's paradigm. Through a model-based analysis of the fMRI signal from both experimental phases, the authors localized the brain regions involved in both the latent learning (leading to a cognitive map or model of the environment) and the subsequent model-driven RL. While the RL-related areas were those typically found in the literature (ventral striatum and dopaminergic system), the areas involved in the formation of cognitive maps were the dorsolateral prefrontal cortex and intraparietal sulci.

The merit of this work consisted not only in the localization of two separate circuits for RL and environment-model (cognitive map) learning, but also in the demonstration that the two processes can be based on very similar computational mechanisms. One is the already described "prediction error" (Equation (1)), the other the "state prediction error". The latter is formally similar to the prediction error (comparison between predictions and real outcomes), but it deals with environmental state transitions. The mathematical form of the state prediction error (δ^{SPE}) is the following (note the similarity with Equation (1)):

$$\delta_t^{SPE} = 1 - T_t(s, a, s') \tag{4}$$

where the value 1 corresponds to the probability of being in the current state (s'; with probability 1), and $T(s,a,s')$ is the expected probability of transition from previous state s (the previous state) to the current state given (chosen) action a. The expectation T is updated by means of the state prediction error:

$$T_{t+1}(s,a,s') = T_t(s,a,s') + \alpha\delta_t^{\mathrm{SPE}} \tag{5}$$

In conclusion, this work showed that prediction error and state prediction error are similar computations but calculated in different brain circuits. This suggested a neurophysiological and computational basis of Tolman's discoveries sixty years before.

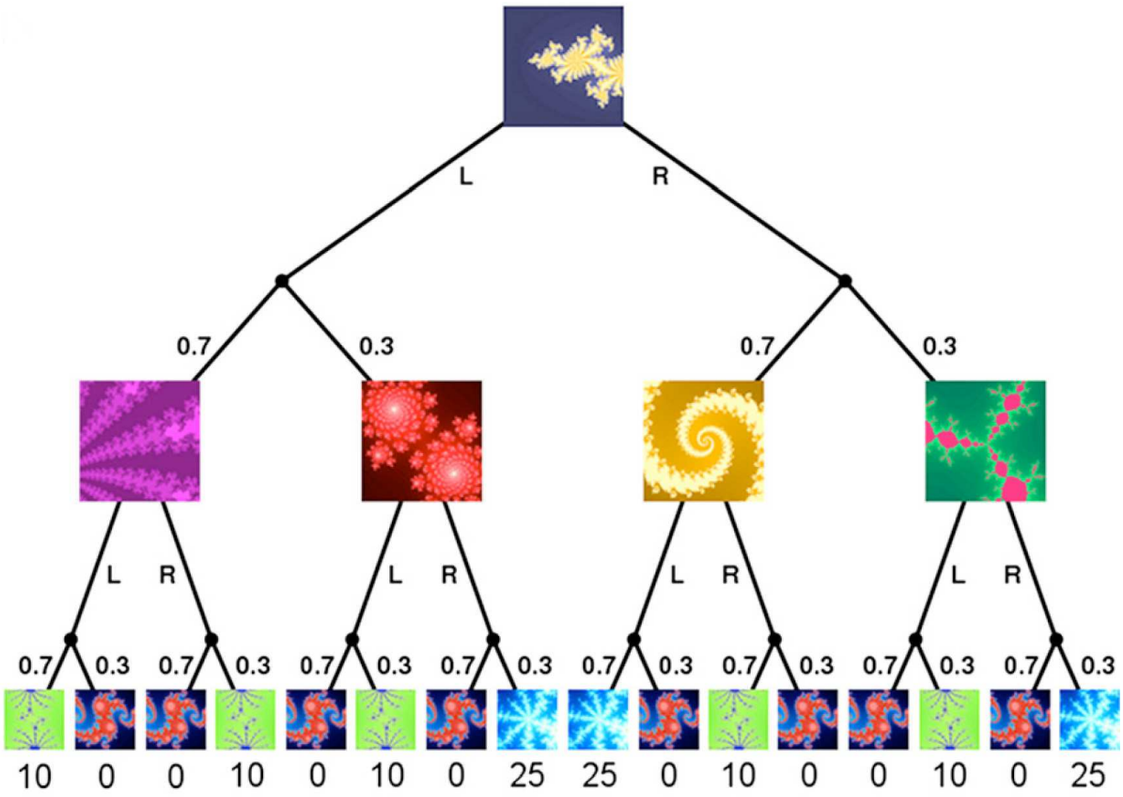


Fig. 2. Formal structure of state space in Gläscher et al.'s experiment. Reprinted with permission from Gläscher et al. (2010).

6. A case study: RL, cognitive control, and anterior cingulate cortex

One of the remaining mysteries of the human mind is executive functioning or cognitive control – the rapid modulation of behavior when called for by unexpected circumstances. In the symbol-manipulation paradigm, executive functioning was proposed to originate from a central executive endowed with two or more "slave systems" (typically the visuospatial sketch pad and the phonologic loop; Baddeley and Hitch, 1974). Detailed models have been developed of the slave systems (Burgess and Hitch, 1999), and in general great progress has been made in understanding them. However, the role of the central executive has remained

poorly understood. To tackle this issue, researchers have tried recasting executive functioning in neural models. We will describe these models and demonstrate how the union of RL and connectionist models provides steps toward understanding the neural basis of cognitive control.

6.1 Associative models of cognitive control

Working within the connectionist framework, Cohen et al. (1990) proposed a model of the Stroop task, a widely used index of cognitive control. In this task, subjects are shown a color word in a given ink color, with the color and word either congruent (e.g., the word RED written in red), or incongruent (e.g., the word RED written in green). The subject's task is to name the ink color. Because word reading is automatic in literate adults, cognitive control is required to override the automatic tendency to read the word. Although subjects can do this, a congruency cost is typically observed, with incongruent trials slower than congruent ones. The Stroop task is widely used in clinical contexts to assess executive functioning, and differentiates between healthy subjects and various patient groups suffering from impairments in cognitive control (e.g., ADHD, Willcutt et al., 2005; Parkinson's disease, Bonnin et al., 2010). In the Cohen et al. model, a distinction is made between an input layer for the relevant dimension and an input layer for the irrelevant dimension, each projecting to a response layer. Crucially, Cohen et al. added task demand units which bias responding toward the relevant dimension (input layer). This brings top-down modulation in an associative model framework.

Botvinick et al. (2001) further developed the model of Cohen et al. They argued that the earlier model did not specify when cognitive control is required. In particular, cognitive control is required only on incongruent trials (e.g., RED written in green), not on congruent ones. For this purpose, they introduced the notion of response conflict, measuring the extent to which responses are simultaneously active. They proposed the conflict monitoring model, according to which response conflict is calculated in anterior cingulate cortex (ACC). Conceptually, this was an advance over the previous model, because not only top-down modulation but also the trial-by-trial cognitive control could be captured in an associative learning framework. In addition, it has been highly influential and allowed accounting for many data. For example, using fMRI Botvinick et al. (1999) demonstrated that human ACC was more active on incongruent trials following a congruent trial than on incongruent trials following an incongruent trial. This finding contradicted the popular notion that ACC activity reflects executive control itself (because the subject should be more "controlled" after an incongruent trial), but was in line with the conflict monitoring model because there should be more conflict after a congruent trial. Note that this model is also a control model in the cybernetic sense mentioned before: It detects when something goes wrong, and when so, it leads to adaptation in the system.

Verguts and Notebaert (2008, 2009) further developed this line of work. They started from the fact that the conflict monitoring model specifies when control should be exerted, but not where (see also Blais et al., 2007). To confront this issue, the authors proposed a neural model in which the implementation of cognitive control was based on an error signal modulating the Hebbian learning between active model neurons. This error signal was, like in Gläscher et al.'s work, borrowed from the RL domain. The new measure, which could be called "conflict prediction error", was computed by comparing the actual amount of conflict, evoked by a stimulus, with the expected mean amount of conflict. This model successfully predicted that cognitive control should not extend across different task input dimensions

(Notebaert and Verguts, 2008) or even across task effectors (Braem et al., in press). Consistent with the model, it was recently demonstrated that ACC responds to item-specific congruencies, not block-level congruencies (Blais and Bunge, 2011).

6.2 New evidence on ACC function: Insights from RL-based neural modelling

Besides conflict monitoring, several other functions have been attributed to the ACC. In humans, evidence using EEG and fMRI pointed toward a role in error processing (Gehring et al., 1993), error likelihood (Brown and Braver, 2005), or volatility (Behrens et al., 2007). Moreover, in the single-cell literature, no direct evidence has been found for conflict monitoring (Cole et al., 2009), while, on the other hand, there is strong evidence for reinforcement processing (Rushworth and Behrens, 2008). More specifically, single-cell recording studies revealed the presence in ACC of three different types of neural units. One population codes for reward expectation, discharging as a function of the expected reward following the presentation of an external cue or the planning of an action. A second population codes for positive prediction error (i.e. when the outcome was better than predicted). Finally, another population codes for negative prediction errors (i.e. when the outcome was worse than predicted). We recently attempted to integrate these different levels of data and theories from the point of view of the RL framework. The model we proposed (Silvetti et al., 2011), the Reward Value Prediction Model (RVPM) demonstrated that all these findings can be understood from the same computational machinery which calculates values and deviations between observed reinforcement and expected values in an RL framework. The global function of the ACC however, remained similar to that in the conflict monitoring model and later versions of it: it is to detect if something is unexpected, and if so, to take action and adapt the cognitive system.

The evolution sketched here, from abstract cybernetic control models to the RVPM, represents a general trend in RL, in which computational, cognitive, and neuroscience concepts are increasingly integrated. Despite this success, not all features of RL have received appropriate attention in the literature. In the final section, we look at an aspect of RL that has been underrepresented.

7. RL and neural Darwinism

Despite the variety in levels of abstraction and purpose of the different models that we described, most of them implement what is sometimes called a triple-factor learning rule (Ashby et al., 2007; Arbuthnott et al., 2000). This means that three factors are multiplied for the purpose of changes in model weights: the first two factors are activation of input and output neurons, constituting the Hebbian component. The third factor is a RL-like signal, which provides some evaluation of the current situation (is it rewarding, unexpected, etc; henceforth, value signal). The value signal indicates the valence of an environmental state or of an internal state of the individual. It can be both encoded by dopaminergic signals (Holroyd & Coles, 2001) or by noradrenergic signals (e.g., Gläscher et al., 2010; Verguts & Notebaert, 2009).

This general scheme of Hebbian learning modulated by value provides an instantiation of the theory of Neural Darwinism (ND; Edelman, 1978). ND is a large scale theory on brain processes with roots in evolutionary theory and immunology. The basic idea of ND consists in the analogy between the Darwinian process of natural selection of individual organisms, and the selection of the most appropriate neural connections between a large population of

them. The general learning rule described above implements such a scheme. Because of the Hebbian component (input and output cells active together), individual synapses (which connect input and output neurons) are selected; and because of the value signal, the most appropriate synapses are chosen.

Just like in Darwinism applied to natural evolution, one key ingredient of ND is variation (called degeneracy by Edelman, 1978), or exploration when the unit of variation is not the individual synapse but rather responses (Aston-Jones & Cohen, 2001). From this variation, a selection can be made, based on an appropriate value signal. Computationally, Dehaene et al. (1987) demonstrated that temporal sequence learning can be achieved by such a variation-and-selection process. In neuroimaging, Daw et al. (2006) demonstrated that frontopolar cortex was used when subjects were in an exploration (rather than exploitation) phase of learning. Besides a few exceptions, however, variation and selection remain poorly studied. Given that it is a key component of RL, we suggest that its further exploration will learn us much more about high-level cognition and its implementation in the human brain.

8. Acknowledgements

MS and TV were supported by BOF/GOA Grant BOF08/GOA/011.

9. References

Arbuthnott, G. W., Ingham, C. A., & Wickens, J. R. (2000). Dopamine and synaptic plasticity in the neostriatum. *Journal of Anatomy, 196,* 587-596.

Ashby, F. G., Ennis, J. M., & Spiering, B. J. (2007). A neurobiological theory of automaticity in perceptual categorization. *Psychological Review, 114,* 632-656.

Aston-Jones, G., & Cohen, J. D. (2005). An integrative theory of locus coeruleus-norepinephrine function: Adaptive gain and optimal performance. *Annual Review of Neuroscience,* 28, 403-450.

Atkinson, R.C., and Shiffrin, R.M. (1968). "Human memory: A proposed system and its control processes," in *The psychology of learning and motivation.* (New York: Academic Press), 89–195.

Baddeley, A.D., and Hitch, G. (1974). "Working memory," in *The psychology of learning and motivation: Advances in research and theory,* ed. G.H. Bower. (New York: Academic Press).

Behrens, T.E., Woolrich, M.W., Walton, M.E., and Rushworth, M.F. (2007). Learning the value of information in an uncertain world. *Nat Neurosci* 10, 1214-1221.

Blais, C., and Bunge, S. (2011). Behavioral and neural evidence for item-specific performance monitoring. *J Cogn Neurosci* 22, 2758-2767.

Blais, C., Robidoux, S., Risko, E.F., and Besner, D. (2007). Item-specific adaptation and the conflict-monitoring hypothesis: a computational model. *Psychol Rev* 114, 1076-1086.

Bonnin, C.A., Houeto, J.L., Gil, R., and Bouquet, C.A. (2010). Adjustments of conflict monitoring in Parkinson's disease. *Neuropsychology* 24, 542-546.

Botvinick, M., Braver, T.S., Barch, D.M., Carter, C.S., and Cohen, J.D. (2001). Conflict monitoring and cognitive control. *Psychol Rev* 108, 624-652.

Botvinick, M., Nystrom, L.E., Fissell, K., Carter, C.S., and Cohen, J.D. (1999). Conflict monitoring versus selection-for-action in anterior cingulate cortex. *Nature* 402, 179-181.

Braem, S., Verguts, T., & Notebaert, W. (in press). Conflict adaptation by means of associative learning. *Journal of Experimental Psychology: Human Perception & Performance.*

Brown, J.W., and Braver, T.S. (2005). Learned predictions of error likelihood in the anterior cingulate cortex. *Science* 307, 1118-1121.

Burgess, N., and Hitch, G.J. (1999). Memory for Serial Order: A Network Model of the Phonological Loop and its Timing. *Psychological Review* 106, 551-581.

Chomsky, N. (1959). Review of Verbal Behavior by B.F. Skinner. *Language* 35, 26-58.

Cohen, J.D., Dunbar, K., and Mcclelland, J.L. (1990). On the control of automatic processes: a parallel distributed processing account of the Stroop effect. *Psychol Rev* 97, 332-361.

Cole, M.W., Yeung, N., Freiwald, W.A., and Botvinick, M. (2009). Cingulate cortex: diverging data from humans and monkeys. *Trends Neurosci* 32, 566-574.

Daw, N. D., O'Doherty, J. P., Dayan, P., Seymour, B., & Dolan, R. J. (2006). Cortical substrates for exploratory decisions in humans. *Nature, 441,* 876-879.

Dehaene, S., Changeux, J.-P., & Nadal, J.-P. (1987). Neural networks that learn temporal sequences by selection. *Proceedings of the National Academy of Sciences: USA, 84,* 2727-2731.

Edelman, G. (1978). *The Mindful Brain.* Cambridge, Ma: MIT press.

Frank, M.J., Seeberger, L.C., and O'reilly R, C. (2004). By carrot or by stick: cognitive reinforcement learning in parkinsonism. *Science* 306, 1940-1943.

Frank, M.J. (2005). Dynamic dopamine modulations in the basal ganglia: A neurocomputational account of cognitive deficits in medicated and nonmedicated Parkinsonism. Journal of Cognitive Neuroscience, 17, 51-72.

Gehring, W.J., Goss, B., Coles, M.G.H., Meyer, D.E., and Donchin, E. (1993). A Neural System for Error Detection and Compensation. *Psychological Science* 4, 385-390

Gläscher, J., Daw, N., Dayan, P., and O'doherty, J.P. (2010). States versus rewards: dissociable neural prediction error signals underlying model-based and model-free reinforcement learning. *Neuron* 66, 585-595.

Grossberg, S. (1973). Contour enhancement, short term memory, and constanciesin reverberating neural networks. *Studies in Applied Mathematics* 11, 213-257.

Hebb, D. (1949). *The organization of behavior; a neuropsychological theory.* New York Wiley-Interscience.

Holroyd, C.B., and Coles, M.G. (2002). The neural basis of human error processing: reinforcement learning, dopamine, and the error-related negativity. *Psychol Rev* 109, 679-709.

Kamin, L.J. (1969). "Predictability, surprise, attention, and conditioning.," in *Punishment and Aversive Behavior.,* eds. B.A. Campbell & R.M. Church. (New York: Appleton-Century-Crofts), 279-296.

Kennerley, S.W., Walton, M.E., Behrens, T.E., Buckley, M.J., and Rushworth, M.F. (2006). Optimal decision making and the anterior cingulate cortex. *Nat Neurosci* 9, 940-947.

Kruschke, J.K. (2008). Bayesian approaches to associative learning: from passive to active learning. *Learn Behav* 36, 210-226.

Lømo, T. (1966). Frequency potentiation of excitatory synaptic activity in the dentate area of the hippocampal formation. *Acta Physiologica Scandinavia* 68 Suppl. 277, 128.

Montague, P.R., Dayan, P., and Sejnowski, T.J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *J Neurosci* 16, 1936-1947.

Neisser, U. (1967). *Cognitive psychology.* New York: Appleton-Century-Crofts.

Notebaert, W., and Verguts, T. (2008). Cognitive control acts locally. *Cognition* 106, 1071-1080.

O'doherty, J.P., Dayan, P., Friston, K., Critchley, H., and Dolan, R.J. (2003). Temporal difference models and reward-related learning in the human brain. *Neuron* 38, 329-337.

O'Reilly, R. C., & Frank, M. J. (2004). Making working memory work: A computational model of learning in the prefrontal cortex and basal ganglia. *Neural Computation, 18,* 283-328.

Pearce, J.M., and Hall, G. (1980). A model for Pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychol Rev* 87, 532-552.

Rescorla, R.A., and Wagner, A.R. (1972). "A theory of Pavlovian conditioning: variation in the effectiveness of reinforcement and nonreinforcement," in *Classical conditioning II: current research and theory,* eds. A.H. Black & W.F. Prokasy. (New York: Appleton-Century-Crofts), 64-99.

Roelfsema, P.R., and Van Ooyen, A. (2005). Attention-gated reinforcement learning of internal representations for classification. *Neural Computation* 17, 2176-2214.

Rudebeck, P.H., Walton, M.E., Smyth, A.N., Bannerman, D.M., and Rushworth, M.F. (2006). Separate neural pathways process different decision costs. *Nat Neurosci* 9, 1161-1168.

Rumelhart, D.E., and Mc Clelland, J.L. (1986). *Parallel Distributed Processing: Explorations in the Microstructure of Cognition.* Cambridge, MA: MIT Press.

Rumelhart, D.E., and Mcclelland, J.L. (1987). Learning the past tenses of english verbs: Implicit rules or parallel distributed processing, in *Mechanisms of Language Acquisition,* ed. B. Macwhinney. (Mahwah, NJ: Erlbaum), 194-248.

Rushworth, M.F., and Behrens, T.E. (2008). Choice, uncertainty and value in prefrontal and cingulate cortex. *Nat Neurosci* 11, 389-397.

Schultz, W., Apicella, P., and Ljungberg, T. (1993). Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. *J Neurosci* 13, 900-913.

Schultz, W., Dayan, P., & Montague, P. R. (1997). A neural substrate of prediction and reward. *Science, 275,* 1593-1599.

Seidenberg, M.S., and Mc Clelland, J.L. (1989). A Distributed, Developmental Model of Word Recognition and Naming. *Psychological Review* 96, 523-568.

Seymour, B., O'doherty, J.P., Dayan, P., Koltzenburg, M., Jones, A.K., Dolan, R.J., Friston, K.J., and Frackowiak, R.S. (2004). Temporal difference models describe higher-order learning in humans. *Nature* 429, 664-667.

Shannon, C.E. (1948). A Mathematical Theory of Communication. *The Bell System Technical Journal* 27, 379–423, 623–656.

Silvetti, M., Seurinck, R., and Verguts, T. (2011). Value and prediction error in the medial frontal cortex: integrating the single-unit and systems levels of analysis. *Frontiers in Human Neuroscience,* 5:75.

Skinner, B. F. (1938). *The behavior of organisms: An experimental analysis.* New York: Appleton-Century-Crofts.

St. John, M.F., and Mcclelland, J.L. (1990). Learning and applying contextual constraints in sentence comprehension. *Artificial Intelligence* 46, 217-257.

Sutton, R.S. (1988). Learning to Predict by the Method of Temporal Differences. *Machine Learning* 3, 9-44.

Sutton, R.S., and Barto, A.G. (1998). *Reinforcement learning : an introduction.* Cambridge (MA): MIT Press.

Tesauro, G. (1989). Neurogammon wins Computer Olympiad. *Neural Computation* 1, 321-323.

Tolman, E.C. (1948). Cognitive maps in rats and men. *Psychological Review* 55, 189-208.

Turing, A.M. (1936). On Computable Numbers, with an Application to the Entscheidungsproblem. *Proceedings of the London Mathematical Society* 2, 230–265.

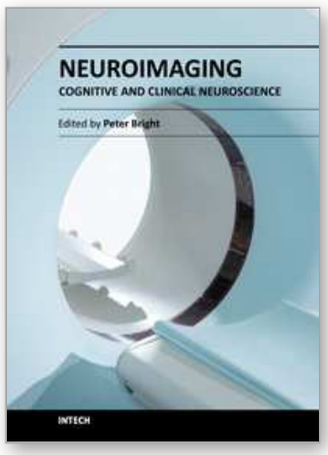
Van Hamme, L.J., and Wasserman, E.A. (1994). Cue competition in causality judgements: The role of nonpresentation of compound stimulus elements. *Learning and Motivation* 25, 127–151.

Verguts, T., and Notebaert, W. (2008). Hebbian learning of cognitive control: dealing with specific and nonspecific adaptation. *Psychol Rev* 115, 518-525.

Verguts, T., and Notebaert, W. (2009). Adaptation by binding: a learning account of cognitive control. *Trends Cogn Sci* 13, 252-257.

Wiener, N. (1948). *Cybernetics or Control and Communication in the Animal and the Machine.* Paris: Hermann & Cie Editeurs.

Willcutt, E.G., Doyle, A.E., Nigg, J.T., Faraone, S.V., and Pennington, B.F. (2005). Validity of the executive function theory of attention-deficit/hyperactivity disorder: a meta-analytic review. *Biol Psychiatry* 57, 1336-1346.



Neuroimaging - Cognitive and Clinical Neuroscience
Edited by Prof. Peter Bright

ISBN 978-953-51-0606-7
Hard cover, 462 pages
Publisher InTech
Published online 16, May, 2012
Published in print edition May, 2012

The rate of technological progress is encouraging increasingly sophisticated lines of enquiry in cognitive neuroscience and shows no sign of slowing down in the foreseeable future. Nevertheless, it is unlikely that even the strongest advocates of the cognitive neuroscience approach would maintain that advances in cognitive theory have kept in step with methods-based developments. There are several candidate reasons for the failure of neuroimaging studies to convincingly resolve many of the most important theoretical debates in the literature. For example, a significant proportion of published functional magnetic resonance imaging (fMRI) studies are not well grounded in cognitive theory, and this represents a step away from the traditional approach in experimental psychology of methodically and systematically building on (or chipping away at) existing theoretical models using tried and tested methods. Unless the experimental study design is set up within a clearly defined theoretical framework, any inferences that are drawn are unlikely to be accepted as anything other than speculative. A second, more fundamental issue is whether neuroimaging data alone can address how cognitive functions operate (far more interesting to the cognitive scientist than establishing the neuroanatomical coordinates of a given function - the where question).

How to reference
In order to correctly reference this scholarly work, feel free to copy and paste the following:

Massimo Silvetti and Tom Verguts (2012). Reinforcement Learning, High-Level Cognition, and the Human Brain, Neuroimaging - Cognitive and Clinical Neuroscience, Prof. Peter Bright (Ed.), ISBN: 978-953-51-0606-7, InTech, Available from: http://www.intechopen.com/books/neuroimaging-cognitive-and-clinical-neuroscience/reinforcement-learning-high-level-cognition-and-the-human-brain

INTECH
open science | open minds

InTech Europe
University Campus STeP Ri
Slavka Krautzeka 83/A
51000 Rijeka, Croatia
Phone: +385 (51) 770 447
Fax: +385 (51) 686 166
www.intechopen.com

InTech China
Unit 405, Office Block, Hotel Equatorial Shanghai
No.65, Yan An Road (West), Shanghai, 200040, China
中国上海市延安西路65号上海国际贵都大饭店办公楼405单元
Phone: +86-21-62489820
Fax: +86-21-62489821

IntechOpen