Learning and Reasoning

Matteo Colombo
m.colombo @ uvt. nl
TiLPS and Department of Philosophy, Tilburg University

1 Introduction

Learning consists in the acquisition of new psychological traits, including abilities, knowledge and concepts. The improvement of some measure of performance when carrying out some task is the main source of evidence for learning. Reasoning consists in a change in view, where certain beliefs or intentions are taken to provide reasons for acquiring, retaining, dropping, or updating some other belief, or for acting in a certain way (Harman 2008).

Learning and reasoning are prominent targets of research in computational cognitive science. Central questions include: On what kinds of innate knowledge and biases does learning rely? Does the human mind deploy multiple kinds of learning and reasoning systems? How are these learning and reasoning systems neurally realized? How do they interact? When are human learning and reasoning optimal or rational? This chapter will explore some of these questions in the light of recent advances in computational cognitive science such as the following two success stories.

First, the Google-owned artificial intelligence company DeepMind created a system called *deep Q-network* (DQN) that could learn how to play 49 different arcade games—including Pong, Pac-Man, and Space Invaders (Mnih et al 2015). Starting with the same minimal body of built-in knowledge, and after about 924 hours training on each game, a deep Q-network learned how to play all 49 games. For more than half of the games, its performance was comparable to that of expert human gamers. Key to the network's versatile learning and human-level performance was the ability of deep Q-learning to combine a model-free reinforcement learning algorithm, *Q-learning*, with a brain-inspired architecture that supported pattern recognition, *a deep convolutional neural network*.

Second, a computational system was developed within the *Bayesian* framework that could learn concepts associated with handwritten characters based on just a single example (Lake, Salakhutdinov & Tenenbaum 2015). This *Bayesian program learning* (BPL) system had the built-in knowledge that characters in writing systems consist of strokes, demarcated by the lifting of a pen, and that the strokes consist of sub-strokes, demarcated by points at which the pen's velocity is zero. Representing concepts as *probabilistic generative models*, its performance on a series of recognition, parsing, and generation tasks was indistinguishable from human behaviour. Key to this system's performance was its ability to *recombine* its existing pieces of causal knowledge to construct hierarchical probabilistic models that *best explained* new observations.

In what follows, I will refer to these two cases to sketch a taxonomy of computational approaches to learning and reasoning (Section 2), and to discuss some implications for three issues. These three issues concern, first, the character of humans' innate cognitive architecture (Section 3); second, the distinction between two kinds of thinking: one fast and intuitive, the other slow and deliberative (Section 4); and third, the nature of rational behaviour (Section 5).

2 An overview of computational approaches

This section distinguishes reinforcement learning from supervised, unsupervised, and Bayesian learning, and clarifies the functional significance of deep and hierarchical architectures.[1]

[1] For thorough overviews, readers may consult Mitchell (1997), Halpern (2005), Russell & Norvig (2009, Parts III-V), and Pearl (2009). Danks (2014a) and Wing (2006) are excellent, concise surveys of machine learning and computational thinking.

2.1 Reinforcement Learning: Model-based and Model-free algorithms

Mnih et al's (2015) DQN learned how to master a wide range of video games by combining a deep learning architecture, a convolutional neural network, and a model-free reinforcement learning algorithm. *Reinforcement learning* (RL) is one of the most popular computational frameworks in machine learning and artificial intelligence. RL offers a collection of algorithms to solve the problem of learning what to do in the face of rewards and punishments, which are received by taking different actions in an unfamiliar environment (Sutton & Barto 1998).

RL systems observe the current state of their environment and take an action—for example, DQN observes an array of pixels representing the current screen in the game it's playing. Then, it selects actions from a set of legal moves in the game—like 'move up', 'move right', or 'fire'. Actions produce rewards or punishments—for example, DQN receives a reward when the game score increases—and cause a transition from the current state to a new state in the environment.

Rewards and punishments can cause changes in behaviour, as they serve as positive and negative reinforcers. The specific causal impact of rewards and punishments on behaviour depends on their magnitude, which is specified by a scalar quantity that can be positive, negative or zero.

Action selection is based on a value (or loss) function defined with respect to a *policy*, which is a mapping from each state and action to the probability of taking a certain action when in a certain state. Given a policy, a *value function* assigns values to either states or state-action pairs, and is updated on the basis of the rewards and punishments the system receives. Given the magnitude of rewards and punishments in a system's environment, the *value* of a state is the expected sum of future rewards (and punishments) that can be achieved by starting to act from that state in that environment under a certain policy. The system learns by pursuing the goal of maximizing its long-term cumulative reward—for example, maximizing its final score in a video game.

DQN implements a *model-free* reinforcement learning algorithm called "Q-learning" (Watkins & Dayan 1992). Systems implementing model-free algorithms learn a value function that specifies the expected value of each pair of a current state of the environment and an action. Model-free learning algorithms learn the value function without building or searching any model of how an action might change the state of the environment. What drives learning is the *reward prediction-error* signal. This signal quantifies the difference between the predicted and currently experienced reward for a particular action in a particular environmental state, and it is used to update the value function.[2] Speaking to its neurobiological plausibility, a wealth of neurobiological evidence suggests that the phasic activity of dopaminergic neurons in the basal ganglia encode reward prediction-errors signals (Montague, Dayan & Sejnowski 1996; Colombo 2014).

Systems implementing *model-based* algorithms build a model of the transition relations between environmental states under actions and the values associated with environmental states. This model may consist in a forward-looking decision tree representing the relationships between a sequence of states and actions, and their associated values. Model-based algorithms evaluate actions by searching through the model to find the most valuable action. In comparison to model-free computing, model-based computing produces more accurate and more flexible solutions to learning tasks. However, model-based computing is slow, and may not tractably solve complex reasoning and learning tasks.

2.2 Supervised and unsupervised learning

[2] The basic insight is similar to the one informing Bush & Mosteller's (1951) pioneering work on animal learning: learning depends on error in prediction. As Rescorla and Wagner put it: "Organisms only learn when events violate their expectations. Certain expectations are built up about the events following a stimulus complex; expectations initiated by the complex and its component stimuli are then only modified when consequent events disagree with the composite expectation" (Rescorla & Wagner, 1972, 75).

RL differs from both supervised and unsupervised learning. *Supervised learning* algorithms rely on an "external supervisor" that can supply the learner with the correct answer when learning. In supervised learning, for every input, the corresponding output is provided, and the learning algorithm seeks a function from inputs to the respective target outputs. For example, an algorithm for solving a classification task may rely on a training set of labelled examples, that is, a set of pairs consisting of some input (a certain example) and the correct output value (the right label to classify that example). After being trained on labelled examples, the behaviour of the supervised learning system may generalize to solve the classification task for new, unlabelled examples. Thus, the main difference between RL and supervised learning is that RL algorithms, unlike supervised learning algorithms, pursue the goal of maximizing their expected reward without relying on a set of supervised instructions about input-output pairings.

Semi-supervised learning is a class of supervised learning, where a system learns by relying on both labelled and unlabelled training data. *Active learning* is a species of semi-supervised learning, where the system can actively query the "external supervisor" for the labelled outputs of the input data (or for the correct response to some problem) from which it chooses to learn. The advantage of active learning over supervised learning is that, for many tasks, providing the system with labelled input-output pairings is difficult and time-consuming; in these tasks, if the system could actively choose its input data and query the external supervisor for labelled outputs, its learning may be quicker and more efficient.

Unsupervised learning does not rely on an "external supervisor." In unsupervised learning, the system receives only a set of unlabelled input data. As no datasets about the corresponding outputs are provided, the learning algorithm seeks to find structure or hidden patterns in the input data. Popular unsupervised learning algorithms, like k-Means and hierarchical clustering algorithms, seek to cluster input data, where a cluster is a grouping of data that are similar between them and dissimilar to data in other clusters. Because unsupervised learning algorithms, unlike RL algorithms, do not rely on reward signals and do not pursue the goal of maximizing expected reward, RL differs from unsupervised learning too.

2.3 Convolutional neural networks and deep learning

DQN is a complex learning system consisting of two main sub-systems: a convolutional neural network and a model-free RL algorithm. *Convolutional neural networks* (LeCun et al. 1989) are a species of connectionist neural networks that make the assumption that their inputs are images. While every image can be represented as a matrix of pixel values, convolutional neural networks leverage two statistical properties of images: that local groups of pixel values are highly correlated (so as to form distinct local visual motifs), and that the local statistics of images are invariant to location (so that a certain motif can appear anywhere in an image) (LeCun, Bengio, & Hinton 2015).

Leveraging these statistical properties of images, convolutional neural networks take arrays of data (e.g., pixel matrices) as inputs, pass them through a series of feed-forward processing stages (or "layers") to get an output. The output is the probability that, for example, the input image is a specific state in a video game. The convolutional network in the DQN learns these probabilities by extracting low level features from input images, and then building up more abstract representations through a series of transformations carried out by multiple hidden layers of neurons. Low-level features extracted by the first layer might consist of dark and light pixels. The next layer might recognize that some of these pixels form edges. The next up the hierarchy might distinguish horizontal and vertical lines. Eventually, the network recognizes objects like trees and faces, and complex scenes like a video game frame.

The convolutional neural network in DQN has many layers, and can then be considered as having a "deep architecture." Inspired by the hierarchical organization of the visual cortex in the brain, a *deep learning architecture* is a multilayer stack of nodes, whose processes can handle large set of input data and extract progressively more abstract representations (Hinton 2007). Recall that

DQN takes as input screen pixels. Screen pixels implicitly contain the relevant information about the current state of the game, but the number of possible game states picked out by different combinations of screen pixels is *very* large. This is where deep learning helps. With greater "depth," that is, with many layers of neurons, the convolutional network in DQN can reliably and tractably recognize the current state of the environment from screen pixels.

2.4 Bayesianism: Probabilistic models, conditionalization, and hierarchical hypotheses

Lake et al's (2015) Bayesian program learning system (BPL) can use one single example of a handwritten alphabetic character to successfully classify new characters, to generate new characters of the same type, and to generate new characters for a given alphabet. While Lake et al (2015) took these successes to be evidence that BPL can learn *concepts*, BPL's performance depends on its capacity of constructing hierarchical probabilistic models that can explain central causal properties of handwriting under a Bayesian criterion.

Despite several important differences, BPL shares a common core structure with other Bayesian systems[3]: BPL aims at constructing an accurate model of a target data-generating process, where a *data-generating process* is the process, system, or phenomenon we want to learn about on the basis of the data it generates. For BPL, the handwriting of alphabetic characters is the generative process, and the data this processes generates include sequences of pen strokes and raw images of characters.

To learn a model of the generative-process, the system assumes a space of hypotheses (a *generative model*) about what raw image and sequences of pen strokes are likely to be associated with what alphabetic character. BPL's hypotheses correspond to probability distributions defined by parameters—for example, by a mean μ and standard deviation σ. Because BPL's hypothesis space is *hierarchically* structured, it includes *over-hypotheses* about the values of the parameters defining the hypotheses at the level below—for example, over-hypotheses about the probability distribution of the values of parameters μ and σ. In turn, BPL's over-hypotheses correspond to probability distribution defined by hyper-parameters—for example, by hyper-parameters α and β.

BPL employs *Bayesian conditionalization* to compute an approximation of the posterior probability of each hypothesis, at each level in its hierarchical hypothesis space, in the light of observed raw images and sequences of pen strokes. Based on the hypotheses with the highest posterior probability, the system infers a specific alphabetic character—for example, an 'A'. BPL uses this inference to achieve human-level performance in a number of classification and generation tasks involving handwritten characters.

3 Learning and innateness[4]

This section makes two claims. First, the empirical successes of Bayesianism and deep learning vindicate *enlightened empiricism* as the correct way to understand the character of our innate cognitive architecture. Second, in contrast to deep learning approaches, Bayesianism displays learning as a rational inductive process of hypothesis-construction and testing, but this does not entail that "there is no such thing as learning a new concept" (Fodor 1975, 95).

[3] Not all Bayesian models are meant to make substantial claims about the mechanisms and representations underlying cognition and behaviour. Some Bayesian models are meant to offer only an encompassing mathematical template that can be applied to a wide range of phenomena in order to provide computational-level analyses (Anderson 1990) and/or in order to unify these phenomena without making commitments to underlying mechanisms and representations (Colombo & Hartmann 2017; Danks 2014b, Ch. 8). Here I set aside questions about the psychological reality of Bayesian models (Colombo & Seriès 2012). Rather, I assume that Bayesianism offers not only a mathematical template or computational-level analyses, but can also make substantial empirical claims about the nature of the mechanisms and representations underlying reasoning and learning (Pouget et al 2013).

[4] This section is based on Colombo (2017).

3.1 Nativism and empiricism

Contemporary nativists and empiricists agree that the acquisition of psychological traits depends on a certain amount of innate structure. The disagreement concerns the exact amount and character of this innate structure (Cowie 1999, 26; Margolis & Laurence 2013, 695). According to empiricists, the innate architecture of the mind includes few *general-purpose* (aka domain-general) *algorithms* for acquiring psychological traits. According to nativists, instead, the innate architecture of the mind includes many domain-specific algorithms or bodies of knowledge for acquiring new psychological traits, where *domain-specific algorithms* operate in a restricted class of problems in a specific psychological domain, and *domain-specific bodies of knowledge* are systems of mental representations that encode information about a specific subject matter such as physics and psychology, and that apply to a distinct class of entities and events (Spelke 1994; Carey 2001).

As nativism and empiricism admit of degrees, Cowie (1999, 153-9) offers a more fine-grained taxonomy. She distinguishes between *Chomskyan Nativism*, *Weak Nativism*, and *Enlightened Empiricism*. In relation to language, *Chomskyan Nativism* is committed to three ideas: that learning a language requires bodies of knowledge specific to the linguistic domain (DS); that the bodies of knowledge specified in (DS) are innate (I)[5]; and that the bodies of knowledge specified in (DS) as being required for language learning are the principles of a Universal Grammar (UG). *Weak Nativism* accepts (DS) and (I), but rejects (UG); *Enlightened Empiricism* accepts (DS), but rejects (I) and (UG). This taxonomy will be helpful for understanding how computational learning systems might be relevant for the nativism debate.

3.2 Bayesianism and deep learning vindicate enlightened empiricism

Deep learning neural networks and Bayesianism are not intrinsically anti-nativist. However, empirically successful Bayesian and deep learning neural models have two interesting consequences for the nativism debate. On the one hand, they show that the acquisition and developmental trajectory of psychological traits need *not* depend on a system of innate, domain-specific representations with combinatorial syntactic and semantic structure (a *language of thought*), which were once assumed to be essential, psychologically primitive (i.e., innate) ingredients of the human cognitive architecture (Fodor 1975, 1981; Chomsky 1980). On the other hand, empirically successful Bayesian and deep learning neural models vindicate enlightened empiricism.

Mnih et al's (2015) DQN shows that psychological traits such as the ability to play a video game can be acquired courtesy of three ingredients: a deep convolutional network for recognizing the current state of the game; a model-free RL algorithm searching for an optimal action-selection policy to maximize its expected score in the game; and inbuilt knowledge that input data are visual images and that only a specific set of actions can be taken in a certain game. Lake, Salakhutdinov & Tenenbaum's (2015) BPL shows that acquiring new concepts of handwritten characters can be carried out by probabilistic algorithms that operate on hierarchical generative models and that exploit primitive representations of edges, and primitive non-propositional knowledge of general spatial and causal relations.

Learning how to play a video game, and learning new concepts of handwritten characters require learners' thoughts about video gaming, and about handwritten characters be constrained by knowledge that is specific to the target domains—respectively, by knowledge of which actions are most likely to yield a high score in a video game, and by knowledge of which pen strokes are most likely to constitute a handwritten character.

This domain-specific knowledge need not built in. As the DQN was trained anew for each game, it acquired visual representations and decision policies specialized for each new game. BPL

[5] According to a prominent explication, innate psychological traits are *psychologically primitive*, which means that their acquisition cannot be explained by any adequate theory in cognitive science (Cowie 1999; Samuels 2002).

acquired concepts of handwritten characters by building them compositionally from primitive representations of edges and of causal and spatial relations. Edge representations were combined to make lines. Lines were combined according to certain spatial and causal relations. In turn, newly acquired representations of handwritten characters could be re-used to build new representations.

In summary, both Mnih et al's (2015) DQN and Lake et al.'s (2015) BPL can learn domain-specific constraints on future learning, on the basis of minimal bodies of innate knowledge and powerful general-purpose algorithms. To the extent that these two computational systems are representative of Bayesian and deep learning connectionist approaches, and are empirically successful, *enlightened empiricism* is the correct way of understanding the character of our innate cognitive architecture.

3.3 Is concept learning impossible for Bayesian systems?

Unlike connectionism, Bayesianism transparently displays learning as a rational inductive process of hypothesis-construction and testing. However, as Fodor famously argued, "[i]f the mechanism of concept learning is the projection and confirmation of hypotheses (and what else *could* it be), then there is a sense in which there is no such thing as learning a new concept" (1975, 95).

According to Fodor, one cannot acquire new concepts via hypothesis-testing because hypotheses are thoughts, and thoughts are constituted by concepts. So, hypothesis-testing presupposes concepts, which means that learning presupposes concepts. From this argument, it would follow that Bayesianism makes the very idea of learning all one's concepts incoherent.

In a sense, Fodor is right that Bayesian systems do not learn. After all, any Bayesian system assumes knowledge of a hypothesis space, which specifies the space of possible structures in the environment that could have generated input data. Each hypothesis in that space is defined as a probability distribution over the possible input data. The system acquires new psychological traits by searching and evaluating hypotheses in its hypothesis space. In this sense, Bayesian systems cannot acquire new psychological structures that were not built into their hypothesis space.

In a different sense, Fodor is wrong. Bayesian systems can acquire novel psychological structures. To understand how, we should distinguish between the *latent* and the *explicit* hypothesis space of a system (Perfors 2012), and between *parametric* and *nonparametric* Bayesian models (Austerweil et al 2015).

A system's *latent* hypothesis space consists of the system's representational resources. It defines the possible thoughts the system can have. The space of thinkable thoughts of biological systems might be determined by their brain's structural and functional patterns of connectivity (Park & Friston 2013). The *explicit* hypothesis space consists of the representations available to the system for evaluation, manipulation, and inference. It defines the system's actual thoughts.

Two types of hypothesis spaces correspond to parametric and nonparametric models. *Parametric* models define the set of possible hypotheses with a fixed number of parameters. For example, the hypotheses might all be Gaussian distributions with a known variance, but unknown mean. Based on input data, the system acquires new psychological traits by estimating the mean of these Gaussians. *Nonparametric* models make weaker assumptions about the family of possible structures in the environment that could have generated input data. The number of parameters of a nonparametric model increases with the number of observed data. This allows for the acquisition of new psychological traits that need not have a fixed, predetermined shape.

Hierarchical Bayesian systems like Lake et al.'s (2015) BPL maintain multiple hypothesis spaces. At the highest level of the hierarchy, we find over-hypotheses concerning the shape of the hypotheses at the levels below (Gaussian, Dirichlet, Beta, etc.). Over-hypotheses are defined by psychologically primitive hyper-parameters that need not pick out any lexical concept—for example, if the system is using a beta distribution to model the distribution of the parameter μ of a Gaussian distribution at the level below, then α and β are parameters of an over-hypothesis, hence hyper-parameters. Hierarchical Bayesian systems can include both parametric and nonparametric hypothesis spaces. In hierarchical, nonparametric systems, the number of parameters is not fixed

with respect to the amount of data, but it grows with the number of sensory data observed by the system.

During hypothesis testing, over-hypotheses in hierarchical Bayesian systems are compared and evaluated in the light of incoming data. Hyper-parameters that best fit the data lead to the generation of a class of hypotheses at the levels below. Specific hypotheses are constructed, and in turn evaluated in terms of their explanatory power over observed sensory data. As data flow into the system and tune hyper-parameters, new sets of representations become available at lower levels in the hierarchy.

The entire process of evaluation of different hypotheses at different levels in the hierarchy, and of construction of specific lower-level hypotheses allows the system to acquire novel psychological traits. These traits are not built into the systems, in the sense that they cannot be simply read off from the system's inbuilt hypothesis space (Kemp & Tenenbaum 2008; Kemp, Perfors, & Tenenbaum 2007).

Bayesian nonparametric hierarchical hypothesis testing need *not* involve psychologically primitive lexical concepts, and can lead to genuine conceptual learning. In this sense, Bayesian systems learn genuinely novel concepts: a richer system of representations can be acquired on the basis of a more impoverished one. This learning depends on a complex interplay of soft innate biases, environmental structure, and general-purpose mechanisms for hypothesis-testing and hypothesis-construction.

4 Reasoning beyond two systems

A popular distinction in the psychology and philosophy of reasoning is between intuition and deliberation. In social and cognitive psychology, this dichotomy is at the core of two-system theories of reasoning. These theories posit two kinds of reasoning systems in the human cognitive architecture, which are distinguished on the basis of two clusters of co-varying properties. System 1 is said to be fast, affective, autonomous, unconscious, effortless, it does not significantly engage working memory resources, and produces intuitive judgements. System 2 is said to be slow, cognitive, controlled, conscious and effortful, it loads on working memory and produces deliberate judgements (Sloman 1996; Kahneman 2011; Evans & Stanovich 2013). Samuels (2009, 134-5) points out that if the human cognitive architecture is comprised of two reasoning systems, "there needs to be some way of distinguishing […] a reasoning system from the rest of cognition so that there are plausibly just two reasoning systems." This section argues that advances in RL and Bayesianism in computational neuroscience are progressively eroding the distinction between two kinds of reasoning systems. RL and Bayesianism appeal to computational features on a continuum, allowing for several different kinds of reasoning systems and hybrid processes. There is no convincing case for an empirically supported distinction between just two reasoning systems.

4.1 RL and two reasoning systems

The distinction between model-free and model-based systems maps onto the distinction between System 1 and System 2 reasoning. To the extent that this mapping is plausibly grounded in computational, neural and behavioural evidence, it would offer a promising way of carving out two reasoning systems in the human cognitive architecture. Here are three preliminary considerations in support of the mapping.

First, like System 1 reasoning, a model-free algorithm is typically fast, knowledge-sparse, and computationally frugal. A model-based algorithm is instead slower, knowledge-involving, and computationally expensive, similarly to System 2 reasoning. Second, the neural circuits that have been associated with System 1 and System 2 reasoning are roughly the same as those that have been associated with model-free and model-based RL. Like System 1 processing, model-free control has been associated with activity in a subcortical neural circuit comprised of the striatum and its dopaminergic afferents. Like System 2 processing, model-based control has been associated with activity in the prefrontal cortex (Niv 2009). Third, model-based systems underlie goal-directed

behaviour, which is typically said to be supported by System 2 processing. Model-free systems underlie habitual behaviour, which is instead said to be supported by System 1 processing (Daw et al 2005).[6]

In particular, Daw et al (2005) proposed that activity in the prefrontal cortex is responsible for implementing model-based strategies, thereby supporting goal-directed behaviour, whereas the dorsolateral striatum and its dopaminergic afferents would implement model-free strategies such as Q-learning, thereby supporting habitual behaviour. These two systems would be "opposite extremes in a trade-off between the statistically efficient use of experience and computational tractability" (Daw et al 2005, 1704). For Daw and collaborators, when the model-based and model-free strategies are in disagreement, the nervous system would rely on the relative accuracy of the evaluations of the two strategies to arbitrate between them. The relative accuracy of the two strategies depends on such factors as the amount of training (which increases accuracy in the model-free system) and the depth of search in the model (which requires heuristic approximations for making value estimates in deeper models, and consequently increases inaccuracy in the model-based system) (see also Keramati, Dezfouli, & Piray 2011).

Despite the extraordinary fruitfulness of the model-based/model-free dichotomy, recent advances call into question the mapping between System 1 and model-free control, and between System 2 and model-based control. First, there are several intermediate modes of learning and reasoning between model-free and model-based control, which vary considerably both in their algorithmic specification, representational formats, and computational properties. These intermediate modes of reasoning exhibit a combination of computational properties that crisscross the distinguishing clusters associated with System 1 and System 2 reasoning (Dolan & Dayan 2013, 320; on this 'crossover' problem see also Samuels 2009).

Second, the two neural pathways that have been associated with System 1/model-free reasoning and System 2/model-based reasoning are densely connected by cortico-basal ganglion loops, and span several parallel and integrative circuits that play several functional roles in reasoning, learning, decision-making, and memory (Haber 2003; Haruno & Kawato 2006; Hazy et al 2007).

In particular, one cannot maintain that dopaminergic activity is the signature feature of System 1/model-free processing. While initial findings suggested that the phasic firing of dopaminergic neurons reports the temporal difference reward prediction error featuring in model-free processing and underwriting several properties of System 1 processing, more recent studies indicate that sub-second dopamine fluctuations might actually encode a superposition of different prediction errors: reward prediction errors and counterfactual prediction errors (Kishida et al 2016). Counterfactual prediction errors signal "how much better or worse than expected the experienced outcome could have been" had the agent performed a different action (Ibid, 200; see also Doll, Simon, & Daw 2012). Counterfactual error signals would speed up model-free learning, but would also be involved in mental simulation of alternative possible outcomes, which has been said to be a defining property of System 2 (Evans & Stanovich 2013). This suggests that dopamine fluctuations in the striatum might implement hybrid forms of RL, involving aspects of both model-free and model-based computation.

Third and finally, a growing number of studies highlight that, in several reasoning tasks, people do not rely purely on either model-free or model-based control. Instead, they capitalize on aspects of both controllers at the same time: adaptive reasoners integrate the computational

[6] In the literature in computational neuroscience and animal learning goal-directed behaviour "is defined as one that is performed because: (a) the subject has appropriate reason to believe it will achieve a particular goal, such as an outcome; and (b) the subject has a reason to seek that outcome" (Dayan 2009, 213; Dickinson 1985). Since the propensity of an agent to select a goal-directed action is sensitive to changes in the variables associated with conditions (a) and (b), goal-directed behaviour is flexible. If behaviour is not affected by these two types of changes, then it is habitual, and is triggered by learned cues and associations.

efficiency of model-free, habitual control with the flexibility of model-based, goal-directed control (Dezfouli & Balleine 2013; Cushman & Morris 2015). DQN, for example, does not purely rely on model-free control, but its reasoning integrates a Q-learning algorithm with stored *experience replay representations*; that is, during gameplay DQN can store in a replay memory representations of state-action-reward-state transitions, which it can use for reasoning and learning in an off-line mode. As Keramati et al (2016, 12871) explain, "humans are equipped with a much richer repertoire of strategies, than just two dichotomous systems, for coping with the complexity of real-life problems as well as with limitations in their cognitive resources."

4.2 Approximate Bayes

Many believe that Bayesianism in cognitive science is committed to positing one general purpose reasoning algorithm, which consists of the Bayesian rule of conditionalization for computing posterior distributions. But the type of neurocomputational mechanism that might account for concept learning, categorization, causal reasoning and so on, cannot implement Bayesian conditionalization, because Bayesian conditionalization makes these learning and reasoning tasks intractable (Gigerenzer et al 2008).

Almost all Bayesian computational systems, including Lake and colleagues' (2015) BPL, deploy a variety of algorithms that compute approximations of a target posterior distribution. For example, two Monte Carlo algorithms that have been used to solve category learning problems are Gibbs sampling and particle filtering (Sanborn, Griffiths, & Navarro 2010). Both Gibbs sampler and particle filter algorithms are flexible, can successfully solve different types of problems, and are computationally less expensive than Bayesian conditionalization. They present important differences too. Perhaps, the most notable difference is that the particle filter algorithms are path dependent, while Gibbs samplers are not.

The Gibbs sampler assumes that all data are available at the time of learning and reasoning: if new data arrive over the course of processing of the Gibbs sampler, then the Gibbs sampler must start its processing anew, which makes it unsuitable for online, rapid, sequential learning and reasoning, whose trajectories may be constrained by evidence observed in the past. Instead, the particle filter algorithm assumes that data are collected progressively over time: posterior distributions are approximated by propagating samples, whose weights are updated as a function of incoming observations. Thus, the order in which different pieces of evidence are encountered has substantial effect on learning and reasoning underlain by particle filter algorithms.

Bayesian systems that compute approximations of target posterior distributions display computational features that straddle the System 1 vs System 2 distinction. Like Carruthers (2014, 199) and others (e.g., Kruglanski & Gigerenzer 2011) have argued, "cognitive scientists would be well-advised to abandon the System 1/System 2 conceptual framework. The human mind is messier and more fine-grained than that." Reasoning is produced and supported by a hodgepodge of computationally diverse systems.

5 Computational (ir)rationality and optimality

Since the 1970s, psychologists Amos Tversky and Daniel Kahneman introduced the term 'cognitive bias' in psychology to describe the systematic and purportedly mistaken patterns of responses that characterize human judgement and decision-making in many situations (Tversky & Kahneman 1974). RL and Bayesian models, however, show a good fit (at least on the aggregate) with people's performance in a variety of psychophysical and cognitive tasks (Colombo & Hartmann 2017; Niv 2009), which has been taken to indicate "a far closer correspondence between optimal statistical inference and everyday cognition than suggested by previous research" (Griffiths & Tenenbaum, 2006, 771). So, there is an apparent tension between the finding that learning and thinking often instantiate "optimal statistical inference" and the observation that people are systematically biased, and often make errors in reasoning.

This final section puts this tension into sharper focus by asking two questions: How can RL and Bayesianism in computational cognitive neuroscience help us diagnose and assess instances of alleged irrationality? What challenges do computational approaches to rationality face?

5.1 Diagnosing (ir)rationality

To begin address these two questions, two distinctions are helpful. The first distinction is between the *personal and the sub-personal* level of explanation (Dennett 1969; Drayson 2014). Explanation of people's behaviour couched in terms of intentional mental states like beliefs, desires, preferences, expectations and emotions is the paradigm case of explanation at the *personal* level. The contents of these mental states stand in logical and semantic relationships, and are easy to express in familiar propositional language. Explanation of behaviour and cognitive functions couched in terms of parts or systems of cognitive agents—for example in terms of networks in the brain or of features of one's perceptual system—is at the *sub-personal* level.

The second distinction is between *constructivist and ecological* models of rationality (Gigerenzer et al 1999; Smith 2008). *Constructivist models* are built under the assumption that their target systems are designed to carry out well-defined functions in order to solve well-defined problems. Given this assumption, constructivist models embody some set of general norms that are used to justify the acceptance of the model, and to specify guidelines for building other general-purpose models of rational behaviour. *Ecological models* of rationality are built on the assumption that their target systems display adaptive responses in certain environments. Given this assumption, ecological models embody some domain-specific strategy that agents can employ quickly and frugally to pursue some goal they care about when coping with a certain problem in a certain environment. Where constructivist models allow us to evaluate behaviour and cognition against *norms* of rationality, ecological models allow us to evaluate behaviour and cognition against *goals* (Polonioli 2015).

Now, behaviour that conforms to predictions based on Bayesian or RL models is often claimed to be 'optimal' or 'rational.' When these models are used in epistemology and decision theory, claims of optimality and rationality should be understood as *constructivist*, *personal*-level claims. Both Bayesianism and RL are in fact *normative* frameworks. Both embed norms of coherence for evaluating the rationality of people's beliefs and decisions.

RL agents aim to maximize the amount of reward they receive from interacting with the environment. While this aim coheres with the idea that rational behaviour consists in maximizing expected utility, RL agents try to achieve this aim by learning and using an optimal policy that yields for each state (or state-action pair) visited by the agent the largest expected reward. Optimal policies in RL should respect a consistency condition defined by *Bellman's principle of optimality*. This principle expresses a relationship between the value of some initial state and the values of the states following from the initial one. It says that, in multi-stage problems, "an optimal policy has the property that whatever the initial state and initial decisions are, the remaining decisions must constitute an optimal policy with regard to the state resulting from the first decisions" (Bellman 1954, 504). The basic idea is that a value function under an optimal policy can be decomposed into two parts: the immediate reward obtained from an action taken from an initial state, and the (discounted) value of the successor state. This condition makes it possible to construct RL algorithms that solve difficult multi-stage problems by breaking them down into simpler sub-problems.

Bayesianism in epistemology is committed to the ideas that people's beliefs come in degrees, that degrees of belief are probabilities, that people's degrees of belief at any given time ought to cohere with the axioms of the probability calculus, and that belief update ought to consist in transforming prior degrees of belief to posterior degrees of belief by conditionalizing on the evidence. Acceptance of such Bayesian models in epistemology depends on norms of epistemic rationality justified on the basis of a variety of considerations like Dutch book arguments and accuracy-based arguments (Hájek & Hartmann 2010).

Taking a constructivist, personal-level approach, Tversky and Kahneman's (1974) work diagnoses that people are irrational when their beliefs do not conform to the epistemic norms embodied in Bayesianism. More recent studies in computational cognitive science, like Griffiths and Tenenbaum's (2006), take the same constructivist, personal-level approach; yet, unlike Tversky and Kahneman's, these studies are concerned with judgements about everyday phenomena, of which people may have extensive experience, and which might explain why people's judgements and behaviour are sometimes found in line with Bayesian and RL norms of rationality (but see Marcus & Davis 2013 for a criticism). In either case, RL and Bayesian models are used as benchmark for rational behaviour in a given task. If people's actual behaviour falls short of the benchmark, then cognitive scientists may try to explain this discrepancy by tweaking the models including parameters that would reflect people's limitations in memory, attention or some other cognitive resource (Zednik & Jäkel 2016).

Many current Bayesian and RL models, however, do *not* target whole people. Mnih et al's (2015) DQN targeted perceptual, learning, and decision-making sub-systems. Lake et al's (2015) BPL targeted sensorimotor and learning sub-systems. As these types of models target parts of cognitive agents, they lie at the sub-personal level of explanation.

RL and Bayesian sub-personal models define the kind of computational problem faced by a target system and the function the system computes to solve the problem. This way, RL and Bayesian models can afford explanations of the *functional significance* of the behaviour exhibited by some target sub-system in some information-processing task. How well the system is functioning is defined only with respect to a loss (or cost) function, which measures the cost associated with each combination of state of the environment and action taken by the system. A system is functioning optimally if its actions (or outputs) minimize expected loss.

Because optimality is defined only with respect to a system's environment and its loss function, the claim that a system functions (sub)optimally depends on correct modelling assumptions about the structure of its environment, on a correct identification of the system's available information and goal in that environment, and on a correct specification of the computational limitations of the system. This means that systems functioning in line with the predictions of a Bayesian or RL model may actually behave sub-optimally, when, for example, they have an incorrect representation of their environment.

5.2 Challenges towards an ecological computational rationality

The computational approaches discussed here are not properly grounded in ecological considerations. Properly grounding computational approaches in ecological considerations has several dimensions. *Time* is one of these dimensions. Learning and reasoning should be sensitive to the expected costs and values of computation under variable time constraints in a changing environment. DQN's training required five hundred times more than the amount of time required by a human gamer to reach a similar performance.[7] By relying on a richer body of symbolic and sub-symbolic representations with compositional structure, BPL developed a capacity for one-shot learning with relatively less training, on around one hundred and fifty different types of handwritten characters. The benefits of one-shot learning and, more generally, of making quick judgements by relying on computationally frugal heuristics that operate on richly structured representations often outweigh the costs of making sub-optimal judgements.

Flexibility is another dimension of ecological rationality. DQN requires re-training to play a novel game, or to pursue a new goal (e.g., 'Die as quickly as possible' instead of 'Make as many points as possible') in the same game. DQN cannot re-deploy knowledge gained in one game to play a new game or to pursue a new goal. BPL shows more flexibility, as it learns models of

[7] As Lake et al (2016) clarify, "DQN was trained on 200 million frames from each of the games, which equates to approximately 924 hours of game time (about 38 days), or almost 500 times as much experience as the human received."

different classes of characters that can be re-deployed and recombined to adapt to new situations. While an ability for building models that show compositional structure seems necessary for flexible and timely reasoning, context-sensitive meta-reasoning and meta-learning systems supplement this ability that consider "the current uncertainties, the time-critical losses with continuing computation, and the expected gains in precision of reasoning with additional computation" (Gershman, Horvitz, & Tenenbaum 2015, 275). Rather than aiming for optimal performance, the challenge becomes how to manage finite time, limited memory and attention, ambiguous data, and unknown unknowns to flexibly navigate a changing environment populated by other agents with different goals and abilities.

A third dimension of ecological rationality concerns the *tuning* between the informational and morphological structure of computational systems and the structure of their environment. While the idea is widespread that prior knowledge should be tuned to environmental statistics, "tuning an organism to its environment involves somewhat more than collecting statistics from the environment, interpreting them as the true priors, and endowing the organism with them" (Feldman 2013, 23). Evolution need *not* put adaptive pressure on organisms to have true priors. It is an important challenge to sort between environments that may favour the evolution of true priors and environments where true priors are ecologically irrational and brittle.

Conclusion

This chapter has surveyed prominent computational frameworks that are used as a basis for analysing and understanding biological learning and reasoning. After laying out a taxonomy of learning systems and clarifying the functional significance of deep connectionist architectures and hierarchically structured hypothesis spaces, the chapter has focus on three implications of computational approaches to learning and reasoning. It has argued that advances in computational cognitive science vindicate enlightened empiricism as the correct way of understanding the character of our innate cognitive architecture, and are progressively eroding the distinction between two kinds of reasoning systems, System 1 and System 2. The final section has put into sharper focus the complex relationship between rationality, optimality, and computational approaches.

References

Anderson, J.R. (1990). *The adaptive character of thought*. Hillsdale, NJ: Erlbaum.

Austerweil, J. L., Gershman, S. J., Tenenbaum, J. B., & Griffiths, T. L. (2015). Structure and flexibility in bayesian models of cognition. *Oxford handbook of computational and mathematical psychology*, 187-208.

Bellman, R. (1954). The theory of dynamic programming. *Bulletin of the American Mathematical Society*, 60(6):503–515.

Bush, R. R., & Mosteller, F. (1951). A mathematical model for simple learning. *Psychological Review*, 58, 313–323.

Carey, S. (2011). The origin of concepts: A précis. *The Behavioral and Brain Sciences*. 34, 113-123.

Carruthers, P. (2014). The fragmentation of reasoning. In P. Quintanilla, C. Mantilla, & P. Cépeda (Eds.), *Cognición social y lenguaje. La intersubjetividad en la evolución de la especie y en el desarrollo del niño*. Lima: Fondo Editorial de la Pontificia Universidad Católica del Perú, 181-204.

Chomsky, N. (1980). *Rules and Representations*. New York: Columbia University Press.

Colombo, M. (2017). Bayesian cognitive science, predictive brains, and the nativism debate. *Synthese*, 1-22. Doi: 10.1007/s11229-017-1427-7.

Colombo, M. (2014). Deep and beautiful. The reward prediction error hypothesis of dopamine. *Studies in history and philosophy of science part C: Studies in history and philosophy of biological and biomedical sciences*, *45*, 57-67.

Colombo, M., & Hartmann, S. (2017). Bayesian cognitive science, unification, and explanation. *The British Journal of Philosophy of Science*, 68, 451-484.

Colombo, M. & Seriès, P. (2012). Bayes in the brain. On Bayesian modeling in neuroscience. *The British Journal for Philosophy of Science*, 63, 697–723.

Cushman, F., & Morris, A. (2015). Habitual control of goal selection in humans. *Proceedings of the National Academy of Science USA*, 112(45), 13817–13822.

Danks, D. (2014a). Learning. In K. Frankish & W. Ramsey (Eds.), *Cambridge handbook to artificial intelligence*. Cambridge: Cambridge University Press, 151-167.

Danks, D. (2014b). *Unifying the mind: Cognitive representations as graphical models*. Cambridge (MA): MIT Press.

Daw, N.D., Niv, Y., & Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, 8, 1704–1711.

Dayan, P. (2009). Goal-directed control and its antipodes. *Neural Networks*, 22, 213–219.

Dezfouli, A., & Balleine, B. W. (2013). Actions, action sequences and habits: Evidence that goal-directed and habitual action control are hierarchically organized. *PLoS Computional Biology*, 9, e1003364. doi: 10.1371/journal.pcbi.100336

Dickinson, A. (1985). Actions and habits: The development of behavioural autonomy. *Philosophical Transactions of the Royal Society of London, Series B, Biological Sciences*, 308, 67-78.

Dolan, R. J., & Dayan, P. (2013). Goals and habits in the brain. *Neuron*, *80*(2), 312-325.

Doll, B. B., Simon, D. A., & Daw, N. D. (2012) The ubiquity of model-based reinforcement learning. *Current Opinions in Neurobiology* 22(6), 1075–1081.

Drayson, Z. (2014). The personal/subpersonal distinction. *Philosophy Compass*, *9*(5), 338-346.

Evans, J. S. B., & Stanovich, K. E. (2013). Dual-process theories of higher cognition advancing the debate. *Perspectives on psychological science*, *8*(3), 223-241.

Feldman, J. (2013). Tuning your priors to the world. *Topics in cognitive science*, *5*(1), 13-34.

Fodor, J. (1981) The present status of the innate controversy. In J. Fodor , *RePresentations*, Cambridge MA: MIT Press, pp. 257-316.

Fodor, J. A. (1975). *The language of thought*. Cambridge, MA: Harvard University Press.

Gershman, S. J., Horvitz, E. J., & Tenenbaum, J. B. (2015). Computational rationality: A converging paradigm for intelligence in brains, minds, and machines. *Science*, *349*(6245), 273-278.

Gigerenzer, G., Todd, P.M., & the ABC Research Group. (1999). *Simple heuristics that make us smart.* New York, NY: Oxford University Press.

Griffiths, T.L., & Tenenbaum, J.B. (2006). Optimal predictions in everyday cognition. *Psychological Science*, 17, 767–773.

Haber, S. N. (2003). The primate basal ganglia: parallel and integrative networks. *Journal of chemical neuroanatomy*, *26*(4), 317-330.

Hájek, A. & Hartmann, S. (2010). Bayesian Epistemology. In: J. Dancy et al. (Eds), *A Companion to Epistemology*. Oxford: Blackwell 2010, 93-106.

Halpern, J.Y. (2005). *Reasoning about uncertainty*. Cambridge (MA): MIT press.

Harman, G. (2008). *Change in view: Principles of reasoning*. Cambridge: Cambridge University Press.

Haruno, M., & Kawato, M. (2006). Heterarchical reinforcement-learning model for integration of multiple cortico-striatal loops: fMRI examination in stimulus-action-reward association learning. *Neural Networks*, *19*(8), 1242-1254.

Hazy, T. E., Frank, M. J., & O'Reilly, R. C. (2007). Towards an executive without a homunculus: computational models of the prefrontal cortex/basal ganglia system. *Philosophical Transactions of the Royal Society of London B: Biological Sciences*, *362*(1485), 1601-1613.

Hinton, G. E. (2007). Learning multiple layers of representation. *Trends in cognitive sciences*, *11*(10), 428-434.

Kahneman, D. (2011). *Thinking, fast and slow*. Macmillan.

Kemp, C. & Tenenbaum, J. B. (2008). The discovery of structural form. *Proceedings of the National Academy of Sciences*, 105(31):10687–10692.

Kemp, C., Perfors, A., & Tenenbaum, J. B. (2007). Learning overhypotheses with hierarchical bayesian models. *Developmental Science*, 10(3):307–321.

Keramati, M., Smittenaar, P., Dolan, R. J., & Dayan, P. (2016). Adaptive integration of habits into depth-limited planning defines a habitual-goal–directed spectrum. *Proceedings of the National Academy of Sciences USA*, *113*(45), 12868-12873.

Keramati, M., Dezfouli, A., & Piray, P. (2011). Speed/accuracy trade-off between the habitual and the goal- directed processes. *PLoS Computational Biology*, 7,e1002055. doi:10.1371/journal.pcbi.1002055

Kishida, K. T., Saez, I., Lohrenz, T., Witcher, M. R., Laxton, A. W., Tatter, S. B., ... & Montague, P. R. (2016). Subsecond dopamine fluctuations in human striatum encode superposed error signals about actual and counterfactual reward. *Proceedings of the National Academy of Sciences*, *113*(1), 200-205.

Kruglanski, A. W., & Gigerenzer, G. (2011). Intuitive and deliberative judgements are based on common principles. *Psychological Review*, 118, 97–109.

Lake, B. M., Ullman, T. D., Tenenbaum, J. B., & Gershman, S. J. (2016). Building machines that learn and think like people. *Behavioral and Brain Sciences*, 1-101. Doi:10.1017/S0140525X16001837

Lake, B. M., Salakhutdinov, R., & Tenenbaum, J. B. (2015). Human-level concept learning through probabilistic program induction. *Science*, 350 (6266), 1332-1338.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., & Jackel, L.D. (1989). Backpropagation applied to handwritten zip code recognition. *Neural Computation*, 1 , 541-551.

LeCun, Y., Bengio, Y., & Hinton, G. (2015). Deep learning. *Nature*, *521*(7553), 436-444.

Marcus, G. F., & Davis, E. (2013). How robust are probabilistic models of higher-level cognition?. *Psychological science*, *24*(12), 2351-2360.

Mitchell, T.M. (1997). *Machine Learning*. New York: McGraw-Hill.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., … & Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, 518 (7540), 529-533.

Montague, P. R., Dayan, P., & Sejnowski, T. J. (1996). A framework for mesencephalic dopamine systems based on predictive Hebbian learning. *The Journal of neuroscience*, *16*(5), 1936-1947.

Niv, Y. (2009). Reinforcement Learning in the Brain. *Journal of Mathematical Psychology*, 53, 139-54.

Park, H.-J., & Friston, K. (2013). Structural and Functional Brain Networks: From Connections to Cognition. *Science*, 342, 579-587.

Pearl, J. (2009). *Causality: Models, Reasoning, and Inference* (2nd edition). Cambridge: Cambridge University Press.

Polonioli A. (2015). The uses and abuses of the coherence–correspondence distinction. Frontiers in Psychology 6:507. doi:10.3389/fpsyg.2015.00507

Pouget, A., Beck, J. M., Ma, W. J., & Latham, P. E. (2013). Probabilistic brains: knowns and unknowns. *Nature neuroscience*, *16*(9), 1170-1178.

Rescorla, R.A., & Wagner, A. R. (1972). A theory of Pavlovian conditioning: Variations in the effectiveness of reinforcement and nonreinforcement. In A. H. Black & W. F. Prokasy (Eds.), *Classical conditioning II: Current research and theory* (pp. 64–99). New York: Appleton Century Crofts.

Russell, S.J. & Norvig, P. (2009). *Artificial Intelligence: A Modern Approach* (3rd edition). Upper Saddle River: Prentice Hall.

Samuels, R. (2009). The magic number two plus or minus: Some comments on dual-processing theories of cognition. In J. St. B. T. Evans & K. Frankish (Eds) *In two minds: Dual processes and beyond* (pp. 129–146). Oxford, England: Oxford University Press.

Sanborn, A. N., Griffiths, T. L., & Navarro, D. J. (2010). Rational approximations to rational models: Alternative algorithms for category learning. *Psychological Review, 117*, 1144–1167.

Sloman, S.A. (1996). The empirical case for two systems of reasoning. *Psychological Bulletin*, 119, 3-22.

Smith, V. L. (2008). *Rationality in Economics: Constructivist and Ecological Forms*. Cambridge: Cambridge University Press.

Spelke, E. (1994). Initial knowledge: six suggestions. *Cognition*, 50, 431-445.

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

Tversky, A., & Kahneman, D. (1974). Judgment under uncertainty: Heuristics and biases. *Science*, 185, 1124-31.

Watkins, C. J., & Dayan, P. (1992). Q-learning. *Machine learning*, *8*(3-4), 279-292.

Wing, J. M. (2006). Computational thinking. *Communications of the ACM*, *49*(3), 33-35.

Zednik, C., & Jäkel, F. (2016). Bayesian reverse-engineering considered as a research strategy for cognitive science. *Synthese*, *193*(12), 3951-3985.