**Review**

**Associative learning and animal cognition**

**Anthony Dickinson**

Department of Experimental Psychology, University of Cambridge, Downing Street, Cambridge CB2 3EB, UK

Associative learning plays a variety of roles in the study of animal cognition from a core theoretical component to a null hypothesis against which the contribution of cognitive processes is assessed. Two developments in contemporary associative learning have enhanced its relevance to animal cognition. The first concerns the role of associatively activated representations, whereas the second is the development of hybrid theories in which learning is determined by prediction errors, both directly and indirectly through associability processes. However, it remains unclear whether these developments allow associative theory to capture the psychological rationality of cognition. I argue that embodying associative processes within specific processing architectures provides mechanisms that can mediate psychological rationality and illustrate such embodiment by discussing the relationship between practical reasoning and the associative-cybernetic model of goal-directed action.

**Keywords:** animal cognition; associationism; associative learning; psychological rationality

---

1. **INTRODUCTION**

Although the antecedents of contemporary animal associative theory are to be found in the psychology of the British empiricist philosophers of the eighteenth century, most notably Locke and Hume [1], it was only during the latter half of the past century that the theory really began to flourish. It is true that neo-behaviourism, which dominated North American psychology during the first half of the twentieth century, subscribed to the very limited form of associationism prescribed by Thorndikes’ Law of Effect that allowed for stimuli to become associated with responses through a process of reinforcement. Although we now have good reason to believe that such a stimulus–response/reinforcement process does in fact mediate the acquisition of habitual behaviour, contemporary theorists have liberated both the syntax and semantics of associationism from the shackles of stimulus–response psychology.

Although some students of animal cognition appear to be what Heyes [2] has called ‘association-blind’, my brief is to examine the role of contemporary associative theory within the study of animal cognition in general. To do so, however, requires some discussion of what we mean by associationism and cognition, and, although I shall have much more to say about each of these questions, somewhat cursory answers will suffice for the moment. Within contemporary associationism, knowledge takes the form of what is often called a representation of the relationship between events, be they stimuli or responses, by a connection (or association) between representations of these events. The process by which this knowledge is deployed is the transmission of excitation or activation (and inhibition) from one event representation to the other via the connection with stronger connections producing greater transmission. Finally, this associative knowledge is acquired by the progressive strengthening of the connection with each effective experience of a relationship between the events.

By contrast, it is much more difficult to say what exactly is meant by the term ‘cognition’ within the field of contemporary animal cognition. Shettleworth [3] argued that ‘cognition refers to the mechanisms by which animals acquire, process, store and act on information from the environment’ (p. 5), a conception that is so catholic as to embrace associationism as one of its cognitive mechanisms. Such *catholic cognition* can be contrasted with what I refer to as *imperial cognition* whose proponents argue that associative learning plays no role, not just in higher order cognition, but also in what are often thought to be the paradigmatic cases of associative learning as exemplified by conditioning. This view has a long history from Tolman to contemporary neo-propositional [4,5] and computational [6,7] theories. In one way or another, such theories argue that animal knowledge takes the form of propositional-like or explicitly symbolic representations that are deployed in the control of behaviour by processes that conform to some normative standard, such as conditional and Bayesian reasoning. Imperial cognition maintains that associative conceptions of knowledge and learning are psychological fictions and that even the relatively simple forms of learning found in Pavlovian and instrumental conditioning involve representations and processes that transcend those instantiated in associative theory.

The third view argues that mammals and birds, at least, are creatures with a dual psychology. On the one hand, certain behavioural capacities are

*ad15@cam.ac.uk

One contribution of 14 to a Discussion Meeting Issue ‘Animal minds: from computation to evolution’.
mediated by associative mechanisms, whereas others require cognitive processes that cannot be captured by these mechanisms. I myself have argued for such a position when fractioning instrumental learning [8]. Whereas stimulus–response habit learning along with Pavlovian conditioning are best explained by associative mechanisms, the processes mediating goal-directed action and future planning lie outside the scope of these mechanisms, a view that I shall challenge in this paper.

Whatever stance is taken concerning the relationship between associationism and animal cognition, all three positions assign a significant role to associative processes. Whereas the catholic stance regards these processes as an important component of cognition, the imperial cognitive and dual psychology view associationism as a contrast against which to evaluate the contribution of cognitive processes to a particular behavioural capacity. From these theoretical perspectives, associationism provides the null hypothesis against which to assess cognition and, as a result, many empirical studies of animal cognition attempt to rule out, or at least assess, the contribution of associative processes.

A recent example comes from a study of tool use by Eurasian jays. Inspired by Aesop’s fable ‘The crow and the pitcher’, Cheke et al. [9] found that two of their jays rapidly learned to deposit a number of stones in a vertical tube to raise the water level in order to reach a mealworm floating on the surface. At issue is whether this behaviour was mediated by physical cognition, namely reasoning by some form of comprehension of Archimedes’ principle, or ‘simple’ instrumental learning about the association between depositing the stones and access to the worm reward. To investigate this issue, Cheke and colleagues devised a U-tube device in which the connection between the two adjacent arms of the tube was obscured from the bird’s view beneath the base of the apparatus. Moreover, the width of the arm containing the floating reward was too narrow for the stones so that the jay had to learn to deposit the stones in the other tube in order to raise the water level in the reward tube.

Assuming that the jays were ignorant of the connection between the two arms of the tube, the elegance of this design is that it removes the information necessary for making causal inferences about displacement while at the same time presenting the bird with at least similar immediate instrumental consequence, the approach of the mealworm contingent upon depositing the stone. Across a series of trials, neither of the jays learned to deposit stones preferentially in the effective tube relative to a control tube that was unconnected to the reward tube. When contrasted with their highly efficient performance in the standard task that manifests the underlying causal process, this failure suggests that instrumental learning about the association between action and outcome was insufficient to establish and maintain this form of tool use.

Whatever stance one takes to the relationship between associationism and cognition, it is clear that students of animal cognition should be sensitive to the complexity of cognitive processes that occur in the wild. By themselves they do not enable associative representations and processes to implement simple cognitive theorems. Instead, they appear to finesse cognition through the means of associative learning.

2. FINESSING COGNITION

It has been known for nearly half a century that stimulus–response reinforcement processes do not encompass even the simple forms of conditioning, and associative theorists now generally assume that such learning can involve the acquisition of associative structures that in some sense represent the predictive and causal relationships embodied in the conditioning procedures. The development of quasi-representational theories has greatly enhanced the sophistication of associative explanations. The second major development arises from the realization that the learning processes controlling the acquisition of such associative representation are neither simple nor unitary, which has led to complex, hybrid accounts of these processes. However, while acknowledging that these two developments have greatly enhanced the explanation power of associationism, I shall argue that by themselves they do not enable associative representations and processes to implement simple cognition but rather simply allow them to finesse a cognitive explanation.

(a) Mediated learning

Imagine a scenario in which a foraging rat comes across adjacent orange and lime trees, both of whose fruits are novel to the animal. The rat forages in both trees and, as a consequence, discovers that it feels well fed. On the basis of this experience alone, the rat has no way of knowing whether the repletion is owing to the orange or lime (or both). However, if subsequently it remains unsatisfied by feeding on the lime fruit alone, the rat has grounds for believing that it is the orange fruit that is nutritious. And, indeed, rats appear to be capable of this inference. Balleine et al. [10] initially fed rats on an orange–lime drink that was sweetened with sugar before giving one group the opportunity to sample an unsweetened lime-alone drink. When subsequently tested with an unsweetened orange solution, they drank more than control rats for which the lime-alone drink had been sweet. The rats retrospectively revalued the orange drink as a consequence of discovering that the lime drink was not the source of sugar.

Retrospective revaluation is problematic for classic associative theory, which assumes that learning about a stimulus can only occur when that stimulus is present. At variance with this assumption, retrospective revaluation shows that experience with one stimulus (the lime-alone flavour) changes the evaluation of another stimulus (the orange flavour) without any

Phil. Trans. R. Soc. B (2012)
further exposure to this revalued stimulus. This form of retrospective revaluation invites an account in terms of reasoning by a disjunctive syllogism: either the lime or the orange drink is nutritious; the lime is not nutritious; therefore, the orange is nutritious.

However, before accepting that rats are capable of such syllogistic reasoning, we should consider whether retrospective revaluation is really problematic for contemporary associationism. More than 20 years ago, Holland [11] argued that associative theory should be liberalized to allow animals to learn not only about directly perceived stimuli, but also about associatively retrieved representations. Therefore, while initially consuming the sweet orange–lime drink, the lime flavour may have been associated not only with the sweet taste but also with the orange flavour so that when the rats subsequently drank the lime-alone solution, representations of both the orange and sugar would have been conjointly activated, thereby allowing them to be further associated [12]. This association between absent but retrieved representation of the orange and sweet tastes could then explain why the rats drank more of the unsweetened orange solution on test.

This analysis immediately raises the question of what is the relationship between accounts of retrospective revaluation in terms of associatively mediated learning and syllogistic reasoning. One interpretation is that mediated learning provides a mechanism for implementing the syllogistic reasoning. However, there are grounds for resisting such an interpretation because the mediated learning can equally well support irrational behaviour. This point is well illustrated by a Danziger cartoon entitled ‘Russian government unable to pay its army’, which illustrates the musing of two unpaid Red Army soldiers in some Siberian posting just after the collapse of the Soviet Union. One soldier says to the other ‘I’m beginning to associate communism with paycheques’. Mediated learning offers a mechanism for this reflection through the conjoint activation of the representations of two absent events, communism that is retrieved by the context of the gulag posting and a pay-cheque that is retrieved by the context of the pay-time. However, this is not necessarily a rational inference as acknowledged by the soldier himself when he concludes his musing with the thought ‘that’s not right, is it?’

This associative analysis raises the issue of whether the retrospective revaluation observed by Balleine and colleagues is an example of rodent disjunctive reasoning or whether their rats simply finessed the inference problem by mediated learning. A study of mediated learning by Dwyer et al. [13], in which they effectively implemented an analogue of the military scenario, favours the associative account. The design is illustrated in figure 1. In the first stage, the rats were allowed to drink a peppermint solution in a distinctive context, context 1, represented by a square chamber. As a result of this experience, the peppermint flavour, which plays the role of communism, should have been associated with context 1, the gulag posting in the military scenario. Then, in a different context, context 2, illustrated as a cylindrical chamber, rats drank an almond-flavoured sugar solution, so that the almond flavour, which functions as pay-time, was associated with the sugar, the pay-cheque. Finally, the rats were replaced in context 1 and allowed to consume an unsweetened almond solution.

This third stage implements an episode analogous to that represented in the cartoon. Being in context 1 (the gulag) should have retrieved a representation of the peppermint flavour (communism) conjointly with the activation of a sugar representation (pay-cheque) by the almond flavour (pay-time). If conjoint activation of retrieved representations is sufficient for associative learning, the peppermint flavour should have become associated with the sugar taste, thereby enhancing its palatability. In accord with this prediction, Dwyer et al. [13] observed that rats that had undergone these experiences drank more of the peppermint solution when presented alone outside any of the training contexts. In other words, rodent associative thinking appears to correspond to that of Red Army soldiers, at least in this respect.

When Holland first discovered that animals learn about associatively retrieved representation, he also demonstrated that such mediated learning shows many of the phenomena of the standard forms of conditioning. However, at the time, mediated learning was thought to be an associative curiosity, but subsequently theorists have come to realize that this form of learning greatly enhances the apparent inferential power of the associative system. However, as Dwyer et al.’s study of mediated learning makes clear, the

Figure 1. An illustration of the design of a mediated learning experiment by Dwyer et al. [14]. See text for details. P, peppermint flavour; A, almond flavour; S, sugar solution.
processes deploying the ‘imaginary’ associations are not constrained to implement a rational thought but rather just finesse a problem that would seem to require a rational solution.

(b) Representational poverty
A second strategy for finessing cognition is to solve what would appear to be a representational problem with a learning process. A major limitation of associationism arises from the way in which it represents the relationship between events in that the only semantics available to the system is carried by the strength of excitatory or inhibitory associations between event representations with the consequence that only a generic relationship, such as a causal or predictive one, can be represented. Traditionally, students of human associative memory [14,15] have attempted to overcome this limitation by positing units of a semantic network that represent specific relationships. So, for example, the proposition that ‘Bill is the brother of Susie’ is instantiated in the associative pathway between representations of Bill and Susie by the interpolation of a ‘brother of’ unit. However, to the best of my knowledge, such relational elements have not been developed within the field of comparative cognition.

It is true that contemporary animal learning has appealed to intermediate or so-called ‘hidden units’ in associative pathways to explain the acquisition of various forms of nonlinear discrimination. Although the propositional interpretation of retrospective revaluation assumes an ‘exclusive-or’ representation of the compound training (‘orange’ or ‘lime’), there are other examples of learning about this class of relationship that are not handled by mediated learning. An example is a negative patterning discrimination. It is impossible for a single-layered associative system to learn that both stimuli X and Y alone predict the same outcome, whereas an XY compound does not (X+ Y+ XY−). If X and Y have excitatory associations with the outcome, then their compound must, if anything, activate the outcome representation to an even greater extent. However, if we assume that, in addition to direct X → outcome and Y → outcome associations, there is an intermediate unit representing the XY compound that is only excited when X and Y are presented conjointly, then standard associative learning processes predict the acquisition of an inhibitory connection from the XY unit to the outcome unit, thereby enabling the compound to counteract the excitatory influences of the X and Y elements [16].

A second limitation arises from the fact that the only variable for representing even a general causal or predictive relationship is the strength of the association. So an associative system cannot distinguish between a belief about the strength of a relationship and confidence in that belief. Both these features are collapsed into the concept of associative strength. However, perhaps the clearest illustration of the representational poverty of associations is learned irrelevance. In a learned irrelevance procedure, an animal is first exposed to a random relationship between two events. So, for example, Baker & Mackintosh [17] exposed thirsty rats to a random relationship between a light cue and the delivery of small aliquots of water in a distinctive experimental context for half an hour each day for a number of days. At issue was what, if anything, did the rats learn about the relationship between the light and water during this experience.

Cognitive folk psychology predicts that the rats should have learned that the light was irrelevant with respect to the occurrence of water, at least within that particular context, learning that should have been manifest if the rats were then required in a second stage to learn that the light now has a predictive relationship to the water. Learning that the light and water are unrelated in the first stage should have retarded subsequent learning of a predictive relationship between these events in the second stage. This is just what Baker and Mackintosh found in that their rats were subsequently slower to learn that the light predicted either the occurrence of the water reward (excitatory relationship) or its omission (inhibitory relationship).

But how can an association represent this irrelevance? Prediction of occurrence is represented by an excitatory association, and prediction of omission by an inhibitory association, whereas the absence of any association just reflects ignorance in the form of a zero associative strength. So associative representations cannot distinguish between an untrained signal and an irrelevant one. Associative theory has finessed this representational poverty by arguing that animals do not in fact represent the absence of a relationship but rather change the way in which they process the stimulus. To take this analysis further, however, requires discussion of associative learning processes.

(c) Prediction error
When faced with the problem of explaining how his rats learned that the light was irrelevant to the occurrence of the water, Mackintosh [18] argued that, rather than representing the random relationship between these events, the rats learned to change the way in which they processed the light in relationship to the water. He noted that during the random training, the light was a relatively worse predictor of the water than was just being in the experimental context. Under the random relationship, the water was just as likely when the light was off as when it was on in the experimental context, whereas this reward only ever occurred when the rat was in the experimental context for half an hour each day. To capture the idea of how good a stimulus (or context) is as a predictor of a particular outcome, in this case the water, Mackintosh deployed the concept of a prediction error, which is a cardinal concept in contemporary associative learning [19]. The prediction error generated by an outcome, such as the water reward, is a measure of how unexpected or surprising it is with respect to a potential signal and is represented by the difference between two terms. The first is the extent to which the signal currently predicts the outcome, which is determined by the strength of the association between the signal and outcome (V'). The greater this associative strength, the better the signal predicts the outcome, whereas the second

Phil. Trans. R. Soc. B (2012)
term is the associative strength that a signal requires to fully predict the outcome ($V^p$). Therefore, the prediction error ($V^* - V$) is large when the signal is a poor predictor of the outcome and small when it is a good predictor. When the signal perfectly predicts the outcome, the error is zero.

The second component of Mackintosh’s theory is a comparison process. When a target signal occurs, the animal compares the prediction error generated by the signal itself with those generated by all the other potential signals that are present. When the target signal generates a relatively smaller error, and therefore is a relatively better predictor than other signals, the learning system changes the way in which the signal is processed in future so that it will be more readily associated with this class of outcome. Conversely, if the signal is a relative worse predictor, its future associability will decrease because it generates a relatively larger prediction error. Through the reiteration of this comparison process across a series of learning episodes, the associative system acquires a set of associabilities for different classes of stimuli that reflect their causal and predictive status with respect to the respective outcomes.

Because the prediction error for the light was larger than that for the experimental context, experience with the random relationship between light and the water should have driven down the associability of the light for the water, which, in turn, should have decreased the rate of learning when the two events were subsequently presented in a predictive relationship. I have examined this example at some length because it illustrates another way in which associative theory can finesse its representational poverty to explain a behavioural effect, learned irrelevance, which on first analysis seems to require inference processes operating on the propositional-like representations.

(d) Hybrid theories

Mackintosh’s theory is just one of a number of learning models that have deployed prediction errors to control associative learning. Most famously, Rescorla & Wagner [20] introduced the contemporary concept of prediction error to control the acquisition of associative strength, but in their case, the prediction error, rather than operating through associability, has a direct impact on changes in associative strength. Although the Rescorla–Wagner and Mackintosh theories are often presented as competing accounts of associative learning, there is a growing consensus that no one theory can account for the gamut of associative learning and, as consequence, recent theoretical developments have focused upon developing hybrid theories that integrate the various ways in which prediction errors have been deployed to control learning. For example, Le Pelley [21] has offered a theory that integrates not only the Rescorla–Wagner and Mackintosh processes but also a theory that uses prediction errors in the control of attention [22]. Such a hybrid account makes evolutionary sense. The ancestral form of associative learning may have been based on simple temporal contiguity between events. However, this simple system was prone to developing superstitious ‘beliefs’ based on fortuitous event pairings. Consequently, a Rescorla–Wagner process evolved in which the fundamental contiguity process was modulated by a prediction error signal [23]. This ensured that learning only occurred when the outcome was unexpected or surprising. However, even this more complex system failed to capture knowledge about the general causal structure of the environment so that natural selection led to the superimposition of a modulating system that deployed prediction errors to control associability in the way envisaged by Mackintosh. Once a system had evolved for computing prediction errors, these signals could then be deployed in a variety of learning processes.

The development of such hybrid learning systems makes the job of determining whether a particular ‘cognitive’ feat lies outside the scope of associative processes even less tractable. This point can be illustrated by recent studies of apparent reasoning about common causes by rats. The relationship between the weather and a barometer reading provides an illustration of this form of reasoning. If the barometer reading is low, there are grounds for believing that it is likely to rain. However, this is not because the low reading causes the rain but rather because these two effects have a common cause in a low atmospheric pressure. However, this inference is not warranted if one intervened in the normal causal processes by interfering with the mechanism of the barometer to produce the low reading. Blaisdell et al. [24,25] investigated whether the humble rat could perform this form of common-cause reasoning.

The typical design of their studies is illustrated in figure 2. In the first stage, rats are presented with a series of trials in which on some occasions a light signal (analogous to the low pressure) is followed by a tone (low barometer reading) and on others equally frequently by food (rain). An issue was whether this experience led the rats to form a common-cause model in which light indicates a common cause of the tone and food. To investigate whether this was so, Blaisdell and colleagues then gave a group of observing rats simple presentations of the tone, having made the indicator of the common cause unobservable by covering the light. They predicted that, on hearing the tone, these rats should infer that the tone indicated that the common cause of food had occurred and, as a consequence, they should run over to the food source in the expectation of receiving food. And this is what they did. This behaviour contrasted with that of a second, intervening group that produced the tone presentation by pressing a lever. Blaisdell et al. argued that because these rats had produced the tone by an independent causal intervention, pressing the lever (analogous to interfering with the barometer), they should have discounted the tone as an effect of the common cause and therefore not have bothered to search food. Again, this was the observed behaviour (but see [26] for further analysis of the effect and an alternative interpretation).

Without going into details, it is difficult to account for the effect of intervention on food seeking in terms of simple associative learning. However, Kutlu & Schmajuk [27] have recently simulated this pattern of results using a complex hybrid attentional-associative...
model. Inevitably, the explanation offered is complex and not amenable to a simple description, but suffice it to say that, according to the model, common-cause training endows the tone with the capacity to activate a representation of the food through a complex chain of associations, an activation that rapidly becomes inhibited in the intervention group by the acquisition of an inhibitory association between lever pressing and the food representation during testing.

The point of addressing this study and model is not to advance the model as an account of common-cause reasoning but rather to illustrate that the enterprise of contrasting contemporary hybrid associative theories with a cognitive account is not straightforward, and I should take the hybrid explanation of apparent common-cause reasoning as yet another example of associative processes finessing rather than implementing cognition.

3. ASSOCIATIVE ARCHITECTURES AND COGNITION

Of necessity, I have presented a sketchy portrait of the contemporary associative theories that have been developed from the study of animal learning. However, this sketch has been sufficient, I hope, to make the point that distinguishing between the behavioural predictions of cognitive (in the imperial sense) and associative accounts is not straightforward, and I should take the hybrid explanation of apparent common-cause reasoning as yet another example of associative processes finessing rather than implementing cognition.

(a) Goal-directed behaviour
What I mean by ‘goal-directed’ is illustrated by an outcome devaluation experiment that we conducted many years ago [26]. Initially, we trained hungry rats to press a lever for one of two types of food pellets, either grain or sugar pellets, with the other type being presented independently of responding. Not surprisingly, the rats rapidly learned to press the lever, but at issue was whether the rats had learned a direct lever→press association, as claimed by stimulus–response/reinforcement theory, or some form of representation of the causal relationship between lever pressing and the contingent food pellet.

To distinguish between these two possibilities, we then sought to change the rats’ evaluation of one of the foods by conditioning aversion to it in the absence of the lever. If the rats had simply learned to press in the presence of lever without encoding information about the contingent food, devaluing either of the pellets should not have directly affected subsequent responding. By contrast, if the rats had learned that lever pressing caused the delivery of the contingent food, devaluation of this food should have depressed responding relative to devaluation of the non-contingent food, which importantly it did. It is crucial to note that this assessment was conducted in the absence of the foods, i.e. in extinction, so that their differential values could not have had a direct impact on performance through a stimulus–response/reinforcement mechanism.

This outcome devaluation effect allows lever pressing to be characterized as goal-directed by two criteria [28]. The goal criterion requires that performance reflects the current value of the outcome or, in other words, whether or not the outcome is currently a goal for the animal. The finding that the rats pressed less following devaluation of the outcome meets this criterion. The second, instrumental criterion is intended to capture the fact that the action is ‘directed’ at the goal by requiring that the effect of the outcome devaluation is mediated by the causal relationship between, in this case, lever pressing and the food outcome. This criterion was met by contrasting the effect
of devaluing the outcome contingent on lever pressing with that of devaluing the non-contingent outcome.

(b) Practical reasoning
The most straightforward cognitive explanation of sensitivity to outcome devaluation is in terms of practical reasoning or what is often called ‘belief–desire’ psychology. This account assumes that during training the rat acquires a belief that pressing the lever causes delivery of a food pellet and that because the contingent food pellet is attractive, it also has a current desire for this type of food pellet. A process of practical inference then uses the belief and desire as arguments to yield an intention to press the lever.

This explanation may seem no more than a trivial form of ‘folk’ psychology. However, as Searle [29] has pointed out, each component of this explanation (the belief, desire and intention) is a representation because they have a representational relationship to the states of the world. In the case of beliefs, the relationship is one of being true or false, whereas a desire can be fulfilled or unfulfilled. Correspondingly, an intention can be executed or not executed. Although we, as enculturated humans, can understand the representational status of a psychological entity such as a belief because we have an explicit concept of truth, the issue for animal cognition is how to determine whether a component of a psychological theory functions as a representation for an agent lacking any concept of representational relationships, such as truth.

In discussing insect cognition, Webb [30] endorses the idea that cognition involves the deployment of some form of the internal model or simulation in the flexible control of behaviour. While agreeing with this general characterization, I should argue that the nature of the flexibility needs elaboration before we can determine whether the model or simulation functions as a representation for the animal, at least by the standards of imperial cognition. As Anderson [31], among others, pointed out many years ago, one cannot characterize a representation without discussing the nature of the processes that determine and deploy its content, and therefore cognition should be analysed in terms of representational systems rather than representations in isolation. From this perspective, we can elaborate the notion of flexibility by arguing that for a psychological entity to be a representation, the processes that operate on it must respect its representation relationship, which, for example, in the case of a putative belief is the fact that it is an entity that can be true or false.

There are two types of processes that determine whether a psychological system is representational. The first are the processes that determine the content of the putative representation, which again in the case of a belief should function to ensure that the content is true. The obvious process for ensuring the truth of representational content is learning. If the content of a ‘belief’ remains entirely resistant to evidence that it is false, it would surely be perverse to characterize this psychological entity as bearing a representational relationship to the world.

A second way in which a psychological entity manifests the fact that it has a representational relationship is through the processes that deploy its content in the control of behaviour. If the process that governs the interaction between beliefs consistently produces a false product when the two initial beliefs are true, then this process does not respect the representational properties of the putative beliefs. So for a psychological entity to be a representation in the imperial sense, the processes that deploy its content must meet some criterion of rationality or necessity. This is the case for practical reasoning. If the belief about an action–outcome contingency is true and the intention to perform the action is executed, then of necessity the desire for the outcome must be fulfilled. In other words, if lever pressing does in fact yield food pellets, and the rat presses the lever, then its desire for a food pellet must be satisfied. In this sense, practical inference is a rational deployment of representations in the control of behaviour that respects the representational status of its arguments [32].

(c) Associative-cybernetic model
Given this analysis, the issue is whether associationism is a representational theory. Up to now, I have certainly used the term ‘representation’ liberally in describing associative theory, and it certainly meets the criterion of having learning processes that bring associations into correspondence with reality. However, it is far from clear that the processes that deploy the content of association, excitation and inhibition are in any sense rational. At a psychological level, excitation and inhibition gain their explanatory power by analogy to physical processes and for this reason, associationism is often characterized as mechanistic. So the cognitive status of associationism is determined, at least in part, by whether the deployment of an association by the processes of excitation and inhibition constitutes a representational system of good standing. However, we have already seen that this is not the case. The associative processes yielding ‘rational’ retrospective revaluation also produced irrational mediated learning, which is not surprising given that the processes of excitation and inhibition are constrained by mechanistic principles rather than by those of rationality.

Typically, associative explanations view the learning processes as functioning with an unstructured network of nodes and connections. This is true of the associative accounts I considered for both retrospective revaluation and ‘syllogistic reasoning’ and the hybrid model of ‘common-cause reasoning’. In fact, the architectural constraints imposed upon an associative network usually amount to little more than assuming that the pathway from input to output passes through layers of so-called ‘hidden’ units with perhaps some recursive connections. However, such an impoverished architecture is neurobiologically unrealistic. It is known, for example, that goal-directed action of the type mediated by practical reasoning is controlled by complex feedback circuits and loops between cortical and subcortical brain structures [11] so that any associative processes functioning within or between these structures does so within a constrained systems architecture. Given this fact, the ability of associative processes to implement, rather than just finesse imperial cognition could arise from constraints imposed by a particular processing architecture.
To give this rather abstract claim substance, I shall sketch the associative-cybernetic model of goal-directed action of the type mediated by practical reasoning [33,34]. However, before describing the system in associative terms, I shall present a folk psychological gloss of how, according to this model, a rat comes to press a lever for food in a goal-directed manner having previously learned to do so. According to this account, the sight of the lever causes the rat to think about pressing it because this is what it has done previously. This thought in turn retrieves a memory of the food pellet that in the past has been delivered by pressing the lever. The current value of the pellet is then assessed and, if this evaluation is positive, the rat then acts on its thought of lever pressing. Figure 3 illustrates how this account may be implemented in an associative architecture and also the relationship between the functioning of this system and the components of practical reasoning.

A latent intention to lever-press comes about because the sight of the lever activates a perceptual unit in the habit memory, which in turn activates a response unit for pressing via a stimulus–response association established during previous training. Following only moderate training, however, the association is not sufficiently strong for the output of the habit memory to drive the motor unit that generates the behaviour of lever pressing reliably. However, the activation is sufficiently strong to generate lever pressing on some occasions, which then allows the rat to experience the contingency between this behaviour and the food reward.

Even though this activation by itself may not produce reliable lever pressing, it can generate the ‘intention’ to press. The generation of the intention involves the interaction between the habit memory and the associative memory, which is the repository of the associations between stimulus representations arising from experiencing relationships between stimuli (the stimulus inputs activating these representations are not illustrated in figure 3). On occasions when the rat presses the lever during training, activation of the lever-press unit in the habit memory will be paired with the stimulus feedback (proprioceptive, somesthetic, visual, auditory) from this action, a pairing that results in the formation of an association between the habit memory response unit and one in the associative memory representing the sensory feedback from pressing (association 1 in figure 3). As a consequence, this association can generate the ‘intention’ to lever-press in the form of an excited sensory representation of this action in the associative memory even on occasions when the activation level in the habit memory is not sufficient to produce overt behaviour.

The next step in the control loop involves an association in the associative memory between the response-feedback unit and one activated by the sensory characteristics of the food reward, which is brought about by experience with instrumental contingency during training. This connection in the associative memory instantiates the belief that lever pressing causes a particular food within the practical reasoning account (association 2 in figure 3). The final component of the cognitive account, the desire for the food reward, is embodied in the model by an association between the food unit in the associative memory and a reward unit in the incentive system (association 3 in figure 3). Again, this association has to be learned in that a desire for this particular food is acquired by experiencing that the food is rewarding [35].

The final component of the model is an output from the reward unit that supplies simultaneous activation to all motor units. Although this positive feedback activation is indiscriminate, it selectively summates with the feed-forward activation from the habit memory on the motor unit for lever pressing because the feed-forward and feedback activation have a common origin in the response unit in the habit memory. This coincident activation of the motor unit will thereby enhance the likelihood that the rat will press the lever in preference to all the other behaviours potentially supported by the environmental context. It is this temporally determined summation that implements the process of practical inference.
There are a number of points to note about this account of the practical reasoning underlying goal-directed action. First, because of the indiscriminate nature of the excitatory feedback, the potential actions excited by environmental stimuli have to be ‘reasoned’ about serially. This is because the desire for the goal is bound with the appropriate action only by the temporal coincidence of action-specific feed-forward excitation from the habit memory with the general feedback activation from the incentive system, which was itself initiated by the activation in the habit memory. Second, the practical reasoning is not explicitly implemented in the model but rather emerges from the way in which the architecture of the model constrains the interaction between its associative components. It is this feature of the model that allows it to explain not only the rational, goal-directed action but also some of the non-rational aspects of the rat’s behaviour. For example, it may well be that lever pressing increases the rat’s heart rate in anticipation of the receipt of food, which could be explained by assuming that activation of the reward unit in the incentive system also excites the heart (figure 3). However, this heart rate response, although possibly biologically adaptive, is not psychologically rational in the way that lever pressing is. The elevated heart rate is unrelated to the fulfillment of the rat’s desire for food pellets—pharmacologically blocking the heart rate increase would not affect whether the food pellets were delivered and therefore the desire fulfilled. So the lever-press → food pellet and the food pellet → reward unit associations do not function as a ‘belief’ and a ‘desire’ in the causation of the heart rate increase in the way they do in the control of the goal-directed action, lever pressing. This point emphasizes the fact that one cannot characterize an association as a ‘belief’ or ‘desire’ without reference to the process that deploys it. One and the same lever-press → food pellet association is a belief with respect to the act of lever pressing but not with respect to the heart rate response. In the case of lever pressing, the system that deploys the association in the service of the goal-directed action constrains the functioning of the association within a processing architecture that implements practical reasoning. No equivalent constraint is evident, however, in the way in which the association mediates the heart rate increase.

The same point can be made in respect of the intention to lever-press, which is generated in the model by activation of the respective unit in the habit memory. However, this is only the case when this activation initiates and, in turn, is modulated by the positive feedback loop through the associative memory and incentive system. It is well established that extensive training of instrumental behaviour can lead to a transition from goal-directed to habitual control. For example, whether or not devaluing food pellets has an immediate impact on the propensity to lever-press depends upon the amount of training. Whereas the devaluation reduces the propensity to press after limited training, more extended training renders this action impervious to the current value of the reward [36]. Pressing then appears to be simply elicited by the lever stimulus independently of the current value of the reward and functions as a stimulus–response habit. In the associative-cybernetic model, this change in control comes about because the output of the reward unit not only supplies a general activation to all motor units but also a reinforcement signal to the stimulus–response associations in the habit memory that increases the strength of any concurrently active association. As a result, the lever → press association, which initially functions to initiate an intention to press the lever, gets strengthened with training to such an extent that it can reliably activate the respective motor unit independently of any positive feedback. When this happens, the whole process of practical reasoning is short-circuited and the rat presses the lever independently of any belief it might have about which action causes the food pellets and its desire for the pellets. Under these circumstances, an excited lever → press association with the habit memory does not function to initiate an intention to lever-press, at least within a cognitive account this behaviour, but rather as a direct elicitor of a press response.

I have belaboured the exposition of the associative-cybernetic model in the hope that it provides a worked example of how associative mechanisms can manifest rational cognition once they are constrained by an appropriate processing architecture. Moreover, the use of an associative processing currency allows for the integration of the cognitive and non-cognitive aspects of a dual psychology.

4. CONCLUSIONS

It makes little sense to argue about whether associative processes per se are cognitive. Certain forms of associative processes, such as the stimulus–response learning mediating habits and probably even the association between signal and outcome underlying Pavlovian conditioning [37], are not because these processes do not conform to the canons of psychological rationality. However, constraining associative learning with a particular processing architecture can enable these processes to meet these criteria. It may well be thought that embedding associative processes within a complex system architecture sacrifices one of the important virtues of associationism, what Heyes [2] call ‘simple-mindedness’. As she herself points out, however, the virtue of simple-mindedness is not as well grounded as is often thought and may well be worth sacrificing for the integration of cognitive and non-cognitive psychologies.

There is, of course, little that is novel about my general thesis. The study of human cognition is replete with highly structured, complex processing models, which provide possible platforms for developing theories of animal cognition. Indeed, Van der Vaart et al. [38] have done just that by adapting Anderson’s ACT-R theory of human cognition [39] to provide a cognitive account of various aspects of corvid caching. All that has been added here is the suggestion that such components can function on associative principles and that one and the same association can function in both a cognitive and non-cognitive mode. So, to return to Cheke et al.’s clever jays that may have manifested domain-specific knowledge about physical causation that transcends the general causal knowledge of instrumental learning, my case is that a plausible theoretical task for students of animal
cognition is to develop a processing architecture that represents and deploys such knowledge by using the component processes supplied by associative theory.

REFERENCES


