PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B BIOLOGICAL SCIENCES

Simple minds: a qualified defence of associative learning

Cecilia Heyes

Phil. Trans. R. Soc. B 2012 **367**, 2695-2703
doi: 10.1098/rstb.2012.0217

PHILOSOPHICAL TRANSACTIONS OF THE ROYAL SOCIETY B

CrossMark click for updates

Phil. Trans. R. Soc. B (2012) **367**, 2695–2703
doi:10.1098/rstb.2012.0217

Review

Simple minds: a qualified defence of associative learning

Cecilia Heyes*

All Souls College and Department of Experimental Psychology, University of Oxford, Oxford OX1 4AL, UK

Using cooperation in chimpanzees as a case study, this article argues that research on animal minds needs to steer a course between 'association-blindness'—the failure to consider associative learning as a candidate explanation for complex behaviour—and 'simple-mindedness'—the assumption that associative explanations trump more cognitive hypotheses. Association-blindness is challenged by the evidence that associative learning occurs in a wide range of taxa and functional contexts, and is a major force guiding the development of complex human behaviour. Furthermore, contrary to a common view, association-blindness is not entailed by the rejection of behaviourism. Simple-mindedness is founded on Morgan's canon, a methodological principle recommending 'lower' over 'higher' explanations for animal behaviour. Studies in the history and philosophy of science show that Morgan failed to offer an adequate justification for his canon, and subsequent attempts to justify the canon using evolutionary arguments and appeals to simplicity have not been successful. The weaknesses of association-blindness and simple-mindedness imply that there are no shortcuts to finding out about animal minds. To decide between associative and yet more cognitive explanations for animal behaviour, we have to spell them out in sufficient detail to allow differential predictions, and to test these predictions through observation and experiment.

Keywords: animal cognition; associative learning; chimpanzee; cooperation; Morgan's canon; parsimony

1. SIMPLE-MINDEDNESS AND ASSOCIATION-BLINDNESS

There is a current of opinion in the study of comparative cognition suggesting we should assume that animals have simple minds. This current takes a variety of forms. Sometimes there is an explicit appeal to Lloyd Morgan's canon, parsimony or Ockham's razor. These principles are usually taken to indicate that, when animal behaviour could be due to two different psychological processes, and these processes vary in complexity, it is more scientifically healthy or legitimate to assume that the behaviour is due to the less complex option. Proponents of the simple-mindedness tend to emphasize the importance of associative learning. In the most extreme cases, the claim that animals have simple minds amounts to the claim that associative learning is the only way in which animals can think about the world.

In marked contrast with the simple minds lobby, many contemporary scholars of comparative cognition appear to assume that animals do not engage in associative learning. This view is implied by the many recently published studies reporting that animals 'understand' a particular aspect of the world (e.g. causality, intentions, reciprocity or the needs of others), or 'have' a particular psychological attribute (e.g. theory of mind, shared intentionality or social motivation), without mentioning associative learning. These studies, which very often involve primates, are not designed to test an 'understanding' hypothesis against an associative learning hypothesis, and do not discuss the possibility that associative learning could produce the focal, intelligent behaviour.

In this article, I suggest that research on comparative cognition needs to steer a course between simple-mindedness and association-blindness. I offer a qualified defence of associative learning, arguing that it should be treated as a contender—a candidate explanation for intelligent behaviour in all animals—but not as a default winner. When behaviour can be explained by both associative and more cognitive mechanisms, we need to do more empirical work; we cannot assume that associative processes are responsible.

I begin by outlining a recently published experiment that reported some interesting prosocial behaviour in chimpanzees, and did not refer to the possibility that this behaviour could be due to associative learning [1]. I chose this experiment by Horner and colleagues as a focal example of association-blindness because it is excellent in all other respects. It addressed a timely and important question with meticulous care, obtained interesting results and was published in a high profile journal. After outlining the case study,

*cecilia.heyes@all-souls.ox.ac.uk

Electronic supplementary material is available at http://dx.doi.org/10.1098/rstb.2012.0217 or via http://rstb.royalsocietypublishing.org.

One contribution of 14 to a Discussion Meeting Issue 'Animal minds: from computation to evolution'.

I shall discuss three reasons why associative learning should be considered as a candidate explanation, not only for the behaviour reported in the case study, but in all other studies of intelligent behaviour in animals, including primates. These reasons draw on evidence that associative learning (i) occurs in a wide range of vertebrate and invertebrate taxa and across a variety of functional domains, (ii) has been conserved in primates including humans, and (iii) cannot be dismissed on historical grounds. In the final part of the article, I turn from association-blindness to simple-mindedness, arguing that, although they are important candidates, associative hypotheses do not have a privileged status in the explanation of animal behaviour.

2. CASE STUDY: PROSOCIAL BEHAVIOUR IN CHIMPANZEES

Prosocial behaviour is currently a hot topic in comparative cognition. In the past decade, increasing numbers of observational and experimental studies have investigated the extent to which other primates share the cognitive and motivational processes that underwrite human cooperation. In one of the latest studies of this kind [1], Horner and colleagues tested six female chimpanzees using a sophisticated version of the Prosocial Choice Test (PCT). In every trial of this test, the focal chimpanzee (the actor) was offered a bucket containing a jumble of plastic pipes, and allowed to remove one of them from the bucket. Half of these 'tokens' were of one colour, and the other half were of a different colour. If the actor selected a token of one colour (e.g. purple), the experimenter gave a food reward to the actor, and to another chimpanzee in an adjacent enclosure (the partner). If the actor selected a token of the other colour (e.g. green), the actor was given a food reward but the partner was not. The actors had been given the opportunity to learn the outcomes of their token choices immediately before the test. In this contingency training session, the procedure was the same as that used for testing, but the bucket contained just one token (e.g. green or purple) in each trial. During both contingency training and the PCT, actors and partners were able to see and hear one another through a wire mesh window between their enclosures. Slices of banana, wrapped in paper, were used as rewards. 'Unwrapping the paper made a loud noise (like eating bonbons), so that actors did not need to rely on vision alone to know whether the partner had been rewarded' [1, p. 13 850].

The study yielded two key findings: (i) In the PCT, the actors chose a dual-reward token, resulting in the provision of food to both actor and partner, in significantly more trials than they chose a single-reward token, resulting in the provision of food to the actor only. (ii) In control trials (where the actors were tested in the same way, but no partner was present in the adjacent enclosure), the actors selected dual-reward and single-reward tokens at random. (See electronic supplementary material for discussion of additional findings.)

In combination, these results were interpreted as evidence that chimpanzees are capable of 'prosocial choice'; as undermining claims that chimpanzees are 'marked by indifference to the welfare of others', and have 'a limited sensitivity to the needs of others'; and as 'indicating that the partners were not passive food recipients but understood the difference between selfish and prosocial token choices'. Couched in a positive way, these statements imply that the results of the study provide evidence that, at least as actors, chimpanzees know (cognition) and care (motivation) about the needs of conspecifics.

Here is another interpretation of the findings, inspired by research on associative learning: when a chimpanzee in this experiment got a food reward, she heard loud rustling as she unwrapped the treat, immediately before she slipped the juicy slice of banana into her mouth. Consequently, just as Pavlov's dogs got to like the sound of a bell, the chimpanzees got to like the sound of paper rustling. In other words, as a result of being presented with the banana, the sound of paper rustling became a 'conditioned reinforcer', a previously neutral event that has acquired reward value through Pavlovian conditioning. Now, single-token choices yielded one shot of this conditioned reinforcer—the sound of the actor unwrapping her own treat—but dual-token choices yielded two shots—the sound of both the actor and the partner unwrapping their treats. Dual-token choices were more richly rewarding for the actor than single-token choices, and therefore, via a process known as 'instrumental learning', the frequency of dual-token choices increased relative to the frequency of single-token choices. This associative explanation for the chimpanzees' prosocial behaviour is consistent with the fact that actors did not show a bias towards dual-token choices in control trials, when no partner was present. In these trials, which used a new pair of token colours, dual-token choices were not followed by the sound of a second treat being unwrapped.

On the associative learning account, the chimpanzees can still be said to have shown prosocial choice because the *effect* of their bias towards dual-token selection was to benefit their partners. However, the associative account implies that the chimpanzees' behaviour was controlled purely by the 'selfish' reward value of the events (paper rustling and banana consumption) that followed their choices. 'Selfish' is in cautionary quotes because, although this description may seem appropriate from the human perspective, the associative account does not imply that the actors represented their behaviour *as* selfish. Furthermore, and crucially, the associative account does not imply that the chimpanzees knew or cared about the needs of their partners. It implies that the effect of their choices was prosocial, but that the underlying representations and motivations were not.

I have suggested in this section that the prosocial behaviour reported by Horner and colleagues could have been due to associative learning. A critic might concede that this is possible in principle, but argue that the chances of it being true in practice are so remote that it would not have been productive for Horner *et al*. to discuss the associative learning hypothesis, or to design their experiment so that associative learning could be distinguished empirically from the hypothesis that chimpanzees know and care about

the needs of others. In §§3–5, I suggest that this line of argument is not persuasive, and that, given what we know about associative learning, it should be considered as a candidate explanation—a major contender—for all new observations of intelligent behaviour in animals.

3. ASSOCIATIVE LEARNING IS EVERYWHERE...

Associative learning is ubiquitous. Evidence of associative learning has been found in every major taxon where it has been sought, and in a huge range of functional contexts, from foraging, through predator avoidance, to mate choice and navigation.

The last major survey of the taxonomic distribution of associative learning was published by MacPhail in his book *Brain and intelligence in vertebrates* [2]. That survey focused on carefully controlled laboratory experiments, and applied exacting criteria to decide whether the members of a particular species were capable of associative learning. MacPhail first asked whether they had provided evidence of Pavlovian and instrumental conditioning, and then whether they had shown compound conditioning effects characteristic of associative learning in mammals. The compound conditioning effects examined by MacPhail [2] included overshadowing and blocking. In Pavlovian conditioning procedures, these effects occur when more than one conditioned stimulus (CS; e.g. a light and a noise) is presented in a contingent relationship, or 'paired', with the unconditioned stimulus (US; e.g. electric shock). Overshadowing refers to a reduction in conditioning to a CS when it is presented with another CS than when it is presented alone. Blocking refers to a reduction in conditioning of one CS as a result of its being presented with a second CS that had previously been paired with the US.

Using these exacting standards, MacPhail found evidence of associative learning, not only in mammals, but in all other major vertebrate taxa: fish, amphibians, reptiles and birds. The fish examples included lemon sharks, goldfish, Siamese fighting fish, golden shiner minnows, green sailfin mollies, guppies, Beau glory, carp and many more. Fewer well-controlled experiments have been conducted with amphibians and reptiles, but MacPhail found evidence of conditioning in leopard and green frogs; spadefoot, African clawed and Woodhouse's toads; salamanders and crested newts (amphibians); and in Bengal monitor, collared and tegu lizards; red-eared turtles; tuatara; indigo and garter snakes; alligators and crocodiles (reptiles). Most studies of conditioning in birds have used the pigeon, but MacPhail's survey also found evidence of associative learning in chickens, doves, quail, magpies and mynah birds.

A contemporary survey of vertebrates would add many new species to these lists. Of yet more interest, however, is recent work providing evidence of associative learning in invertebrate species. For example, demonstrations, not only of conditioning, but also of blocking, have been reported in insects [3], molluscs [4] and platyhelminthes [5].

To secure precise experimental control, many studies of associative learning use arbitrary or 'unnatural' stimuli and responses, such as electric shock and lever pressing. However, there is also a rich supply of more naturalistic studies indicating that associative learning is a motor of adaptive plasticity in a wide variety of species and behavioural contexts. For example, the flavour-aversion learning paradigm first developed by Garcia and colleagues [6], and now a standard laboratory procedure, uses distinctively flavoured foods as conditioned stimuli, rather than lights or tones, and poison as the US, rather than electric shock. The results obtained using this paradigm show that associative learning enables rodents and birds to enhance foraging efficiency and to promote survival by avoiding foods that are potentially toxic [7]. Naturalistic studies have also shown that associative learning enables animals to recognize and to avoid predators [8], to navigate [9], and to enhance the effectiveness of their territorial [10] and sexual behaviour [11].

In this section, I have tried to give some sense of the wide range of taxa and functional domains in which associative learning is known to occur. Although the survey is brief and very far from complete, it highlights the fact that conditioning has been found in every major vertebrate taxon, as well as several invertebrate groups, and is something that happens in the real world, not just in laboratories.

4. ...EVEN IN (HUMAN) PRIMATES

The evidence surveyed in the previous section suggests that associative learning is an important candidate explanation for adaptive behaviour in a wide range of taxa and functional contexts. However, by itself it does not imply that associative learning is an important contender to explain the prosocial behaviour of chimpanzees—behaviour of the kind observed in our case study [1]—or, more generally, to explain complex behaviour in primate species. To be persuaded of this, we need evidence that associative learning has been conserved in primates, and that it does not just control 'spit and twitches' [12], but contributes to the development of subtle, voluntary patterns of primate behaviour. A sample of this evidence is reviewed later. The examples are all taken from recent studies of associative learning in humans. This is necessary because very little contemporary research examines associative learning in other apes. Fortunately, in this case, the focus on humans is not a problem. If associative learning plays an important role in guiding complex human behaviour, in spite of the many ways in which our lives differ from those of all other animals, there is no reason to doubt that it remains a powerful force in shaping the behaviour of other primates.

Let us start with a couple of particularly striking examples where associative learning has been implicated in cognitive functions that many would regard as characteristically human. First, recent research on individual differences shows that, alongside working memory and processing speed, associative learning makes a substantial independent contribution to IQ. How well people perform on standardized tests of general intelligence is predicted by the efficiency of their associative learning [13]. Second, associative learning has recently been implicated in the development of

'sense of agency', the phenomenal experience of producing events through one's own intentional action. In a study measuring 'temporal binding'—a strong index of sense of agency—a blocking effect, characteristic of associative learning, was observed when people were exposed to flashes of colour on a computer screen that may or may not have been produced by their finger movements [14].

Another recent study implicated associative processes in human geometry learning [15]. In a computer-based task, participants were required to learn the location of a goal (food for three blind mice) within a room using geometric information available on the screen. Consistent with an associative model [16], when the goal location was defined by two shapes that differed in salience, the more salient shape overshadowed the less salient shape, and when the shape was pre-trained as a signal for the goal location, learning about the geometric cues of the other shape was blocked.

Turning to the social domain, a series of studies in my own laboratory has provided evidence that associative learning drives the development of mirror neurons, and of the capacity to imitate [17]. For example, using transcranial magnetic stimulation (TMS) and functional magnetic resonance imaging (fMRI), these studies have shown that the action of the mirror neuron system can be reversed—such that it is active during the observation of one action and the performance of a *different* action—by training on a non-matching sensorimotor contingency (e.g. foot movement stimulus > hand movement response, and vice versa [18]). Providing more specific evidence for the involvement of standard associative processes, behavioural experiments examining the effects of training on automatic imitation have shown that it is sensitive to contingency, and shows context effects characteristic of counter-conditioning [19,20].

Mathematical modelling of fMRI data has also shown that associative learning is important in a variety of human decision-making tasks. These studies typically identify associative learning through the footprint of 'prediction error' [21]. Following the Rescorla–Wagner model, most contemporary models of associative learning assume that there is a change in the strength of an association between two event representations (learning occurs) when the sequel or outcome of the first event differs from the predicted outcome, i.e. when there is a prediction error. Therefore, when people are engaged in decision-making, and the blood oxygen-dependent (BOLD) response in focal areas of the brain both correlates with behavioural responses, and fits a prediction error model better than alternative models, this shows that the decision-making is mediated by associative learning. Using this approach, brain imaging studies have demonstrated that in humans, as in other animals, associative learning plays a fundamental role in learning about relationships between actions and their outcomes [22]; in learning higher-order relationships between outcomes [23]; in incidental encoding of relationships among stimuli [24]; in updating perceptual representations [25]; and in tracking the value of social cues (i.e. advice) as well as asocial cues in decision-making [26].

Brain imaging studies of this kind follow and support a substantial body of behavioural experiments, initiated by Dickinson *et al.* [27], showing that associative learning contributes to human causality judgements. This on-going programme of research (see [28] for a recent review) has tackled two issues that have led some researchers to doubt that associative learning is an important determinant of human behaviour. The first issue relates to conscious awareness. In a widely cited paper, Brewer [29] defined conditioning as something that occurs in the absence of contingency awareness, and, finding no evidence of conditioning without contingency awareness in adult humans, declared that human behaviour is not susceptible to conditioning. One response to this challenge has been to find evidence of human conditioning in the absence of awareness; for example, in studies of difficult discriminations [30], or in patients under anaesthesia [31]. Another, more compelling response has been to question Brewer's premise: given that so little is known about the taxonomic distribution and functions of consciousness, why should we assume that the consequences of associative learning—knowledge of a contingency between events—is never available to conscious awareness?

The second and more substantial issue relates to the role of 'inferential' or 'propositional' processes, rather than that of associative processes, in producing basic (Pavlovian and instrumental) and complex (e.g. blocking, overshadowing) conditioning phenomena in human subjects (see [32], including commentaries, for a recent survey of this debate). The most important thing to note about this debate is that the majority of even the most enthusiastic contemporary supporters of associative learning would not deny that inferential processes play crucial roles in human cognition. They are subscribers to some kind of 'dual-process' theory (see [33,34] for further discussion) assuming that humans use both associative learning and inference processes to find out about the world. Furthermore, the majority would readily agree that, at least in humans under some circumstances, conditioning phenomena can be produced by inferential rather than associative processes. However, along with the brain imaging data discussed earlier, carefully designed experiments on human causality judgements have shown that, in many cases, complex human decision-making is controlled by associative learning. For example, this (highly technical) literature shows that blocking occurs, not only when people have a chance to reflect, but also in speeded, unintentional learning tasks [35]; that cue-outcome contingency has an impact on probability estimates even when variations in contingency do not affect the objective probabilities [36]; and that human conditioning shows trial order effects that would not occur if it was based on reasoning about events stored in memory [37].

This short survey of recent research with human subjects makes clear that the mechanisms of associative learning have been conserved in primates, and that they contribute, not just to basic functions, but to complex decision-making and to high-level functions such as IQ, sense of agency, navigation and imitation. This evidence therefore suggests that associative learning is an important contender to explain complex behaviour in non-human primates, including the kind of prosocial behaviour reported in our case study.

5. BUT ISN'T THIS JUST BEHAVIOURISM?

Association-blindness—failure to consider associative learning as a candidate explanation for complex behaviour—could not be due solely to scepticism of the kind addressed in the previous two sections; to doubts about whether ancient, taxonomically general psychological processes contribute to animal cognition. This is made clear by a recent paper that calls for a 'bottom-up perspective on animal and human cognition' [38]. This paper explicitly advocates a research strategy in which complex cognition is explained by identifying the conserved, taxonomically general 'building blocks' from which it is constructed. But the paper makes no mention of one of the most important and well-understood building blocks of them all—associative learning. Why might this be?

A possible answer is suggested by the common, and typically pejorative, use of the term 'behaviourist' to describe associative explanations and the people who deal in them [39]. If associative learning were merely 'behaviourist', and if behaviourism had been shown to be false by the 'cognitive revolution' in psychology, then association-blindness would be justified. We could safely assume that associative learning is a fiction, something like phlogiston, dreamed up by previous generations of scientists to explain data that can now be explained in a much more rigorous and felicitous way. The trouble is that neither of the premises of this argument is sound. Associative learning is not merely behaviourist [12], and the cognitive revolution of the 1950s and 1960s revealed that behaviourism was wildly over-ambitious, not that it was false [40].

Contemporary research on associative learning does retain some elements of behaviourism. For example, it still makes use of operant chambers and other components of the technology developed by Skinner, and it commonly assumes that the events that enter into associations (stimuli and responses) are represented only in a 'thin' sense; by copies or traces of the sensory stimulation they produce, rather than in a way that makes them semantically evaluable—that can be characterized as true or false, correct or incorrect. This commitment to thin representation, although it dates back to the 'sense impressions' of the British Empiricists, is reminiscent of the operationalism popular among behaviourists. However, in most respects, contemporary associative learning theory is thoroughly cognitive, and has been for at least 40 years. For example, rather than defining learning as a change in behaviour, contemporary associative theory assumes that learning is something that happens in an animal's mind, and that changes in behaviour, rather than being constitutive of learning, are signs that learning is taking place. In other words, associative learning theory embraces evidence of 'behaviourally silent' learning [41]. Furthermore, memory is assumed to play a crucial role in associative learning [42], and most models assume that attention is both a major determinant of associative learning, and a process that can itself be modulated by associative learning [43]. The hybrid character of research on associative learning—the fact that it includes traces of behaviourism as well as solid cognitivist credentials—is unsurprising when one reflects on the broad sweep of its history. Yes, the study of associative learning was dominated by behaviourism from the 1920s to the 1950s, but it originated in the work of the British Empiricists some 300 years earlier; was converted into an experimental science by Pavlov in the pre-behaviourist era; was pursued by cognitivists such as Tolman even at the height of the behaviourist period; and has been making rapid progress throughout the 50 years since that vainglorious period ended.

The second premise—that behaviourism has been shown to be false—misrepresents what happened in the phase of psychology's history known as the 'cognitive revolution' [40]. The term 'cognitive revolution' implies that there was a Kuhnian scientific revolution, in which the old behaviourist 'paradigm' was swept away and replaced by a new, incommensurable cognitive 'paradigm'. However, a close examination of the events of the 1950s and 1960s reveals that the discovery of what were subsequently known as 'biological constraints on learning' [6], and increasing doubts about the capacity of associative learning theory to explain linguistic behaviour [44], stimulated rational changes in the content and ambition of associative learning theory. It led to the development of attentional theories of associative learning [45], and to broad acceptance that associative learning can explain some, but not all, behaviour.

So, association-blindness cannot be justified on historical grounds. For a period in its past, research on associative learning kept bad company—it hung out with people who wanted to explain all behaviour using highly impoverished theoretical resources—but the elements it retained from that behaviourist period are not the ones that history has shown to be false. It is likely that residual association-blindness—the kind that remains even when a sceptic is reminded of the evidence reviewed in §§3 and 4—is due, not to bad company, but to bad press and bad complexion. The bad press was a direct result of the misinterpretations of history discussed in this section. Regardless of the historical facts, generations of students have been told that associative learning equals behaviourism and that behaviourism is wrong-headed [39]. The bad complexion comes from the fact that research on associative learning can appear to be exclusive and forbidding. It is an area of cognitive science in which common sense or 'folk psychology' is a poor guide, and that generally values rigour of method over relevance of result. Partly as a consequence of this rigour, research on associative learning has accumulated a substantial body of formal theory, a long list of robust effects, and a highly technical vocabulary. These are scientific strengths, but they do not make research on associative learning user-friendly. Folk psychological explanations for animal behaviour—referring to what the animal 'understands'—are much easier even for specialists to generate and comprehend, and they send non-specialists—including journalists and professional editors of journals such as *Nature* and *Science*—a much more appealing message. The intellectually challenging character of associative learning theory is an understandable but not a good reason for association-blindness, especially when lucid introductions to the subject are readily available [7,46].

6. A CONTENDER, BUT NOT THE DEFAULT

I have argued that association-blindness cannot be justified; given what is known about the pervasive influence of associative learning, it should be considered as a candidate explanation for all new examples of intelligent behaviour in animals. Some researchers would like to go further, to argue that associative learning should be regarded, not just as a contender to explain intelligent behaviour, but as the default. According to these advocates of simple-mindedness, associative learning has an inherent edge. If a pattern of behaviour can be explained with reference to associative processes and in a yet more cognitive way—i.e. if the data are equally consistent with an associative and a 'super-cognitive' hypothesis—we should prefer the associative account. In this section, I explain briefly why I do not think this is right. (In common with Shettleworth [34], I take a 'Catholic' view of cognition [33]. On this view, associative processes are cognitive—they operate on (thin) representations—but not super-cognitive; they are not rational in Dickinson's sense [33]. Thus, Dickinson's 'cognitive' category is equivalent to my 'super-cognitive' category of explanation.)

Supporters of simple-mindedness have appealed on various occasions to MacPhail's null hypothesis, Ockham's razor, parsimony, and Morgan's canon to justify their preference for associative explanations. However, the principle that is clearly most relevant to contests between associative and super-cognitive explanations is Morgan's canon [47]. MacPhail's null hypothesis is a potential tie-breaker when there are two candidate explanations (intellectual and contextual) for a behavioural difference between species, not when there are two candidate explanations (associative and super-cognitive) for a behaviour observed in a single species. Ockham's razor (typically rendered as 'entities must not be multiplied beyond necessity'), and the principle of parsimony, encourage explanations that postulate the minimum number of causal entities or free parameters. In contrast, Morgan's canon is concerned with *kinds* (e.g. associative and super-cognitive) rather than *numbers* of explanatory devices, and not with explanation in general, but with the explanation of animal behaviour in particular.

In 'possibly the most important single sentence in the history of the study of animal behaviour' [48, p. 9], or even in psychology as a whole [49], Morgan [50] rendered his canon:

> 'In no case may we interpret an action as the outcome of the exercise of a higher psychical faculty, if it can be interpreted as the outcome of the exercise of one which stands lower in the psychological scale.' (p. 53)

To see why many advocates of simple-mindedness have interpreted Morgan's canon as endorsing associative explanations, let us return to the case study presented in §2. In that study, Horner and colleagues found that chimpanzees selected dual-reward tokens—tokens that delivered a food reward to a partner, as well as to themselves—more often than they chose single-reward tokens—which delivered a food reward only to themselves—and interpreted this choice behaviour as evidence that chimpanzees know and care about the needs of others. In contrast, I pointed out that the chimpanzees may have chosen the dual-reward tokens more often than the single-reward tokens because the former choice was followed by two instances, rather than one instance, of a sound—paper rustling—that had acquired (selfish) reward value through Pavlovian conditioning.

It is often difficult to tell exactly what a super-cognitive explanation is claiming about animal minds. However, these hypotheses—the super-cognitive 'needs of others' and associative 'paper rustling' explanations—differ in at least two ways that seem to put the associative explanation lower on Morgan's psychological scale. First, the paper rustling explanation suggests that the chimpanzees' behaviour was based on associative processes, on the formation of excitatory and inhibitory links between representations, whereas the needs of others explanation implies that the behaviour was based on reasoning or inferential processes, involving the application of explicit rules to representations. Second, and relatedly, the associative account assumes that the representations supporting the focal behaviour were thin or concrete; just sense impressions, or memory traces, of the sound of rustling paper and the taste of juicy banana. In contrast, the needs of others explanation suggests that the behaviour was based on abstract or conceptual representations, with the structured or language-like quality necessary to support inference processes. In *The limits of animal intelligence*, Morgan proposed 'a threefold division [of mental activity] into instinct, intelligence, and reason' [51, p. 225], with instinct occupying the lowest rung of the ladder, intelligence guided by associative learning in the middle, and reason at the top [52]. Therefore, it seems likely that Morgan himself would have taken his canon to support the associative, paper rustling interpretation of the chimpanzees' prosocial behaviour, over the super-cognitive, needs of others explanation.

But even if associative explanations in general, and the paper rustling explanation in particular, would have received Morgan's imprimatur, there are at least three reasons to resist the dictates of his canon. They concern history, evolution and simplicity.

(a) *History*

Historians and philosophers of science have identified errors in Morgan's work. For example, Sober [53] has argued that Morgan founded his canon on the assumption that 'psychical faculties' evolve by the 'Method of Variation'. The Method of Variation was said by Morgan to imply a distribution of faculties across species in which 'any one of the faculties 1, 2 or 3, may in [species] b and c be either increased or reduced relative to its development in [species] a' [50, p. 57]. Misled by his own graphical representation of this statement, Morgan took it to imply that a species could have a lower faculty without a higher faculty, but not the reverse. If this were true, it would make possession of a lower faculty more probable than possession of a higher faculty, and therefore justify the canon. However, as Sober pointed out, the quoted formula allows both possibilities—a higher faculty without a lower one, as well as a lower faculty without a higher one—and is therefore 'too permissive to justify the canon' [53, p. 233].

(b) *Evolution*

A number of evolutionary considerations might be thought to support Morgan's canon. For example, it could be argued that associative processes are so widespread in the animal kingdom, and so powerful in delivering behavioural adaptation across a variety of functional domains, that they are likely to screen-off selection in favour of higher or super-cognitive processes. Therefore, according to this evolvability argument, super-cognitive processes will evolve only in a narrow range of ecological conditions, where they bestow a significant marginal benefit. Similarly, a cost argument would suggest that super-cognitive processes are less likely to be present in any given species than associative processes because they require larger brains, which are metabolically expensive. On a different tack, a phylogenetic argument might suggest that, if a super-cognitive process is absent in close taxonomic relatives, it is unlikely to be present in the focal species.

As Fitzpatrick [54] has pointed out, rather than providing blanket support for the theoretical conservatism of Morgan's canon, these evolutionary considerations cut both ways. Morgan's canon is a general methodological principle. Therefore, to justify the canon, we need reasons that apply to *all* species in *all* cases, but the evolvability, cost and phylogenetic arguments only give us reasons to favour associative over super-cognitive explanations for *some* behaviour in *some* species. In other cases they give us reason to favour super-cognitive over associative explanations. If the species in question occupies an ecological niche where a super-cognitive process would have a significant marginal benefit (e.g. theory of mind in an especially complex social environment), or has a large brain (e.g. high encephalization quotient), or has relatives that appear to be running the super-cognitive process (e.g. humans for other ape species), then the evolvability, cost and phylogenetic arguments militate in favour of super-cognitive explanation and against the dictates of Morgan's canon.

(c) *Simplicity*

It is no simple matter to explain why simplicity is an explanatory virtue. However, the idea is so pervasive in science and philosophy that it seems reasonable to suppose that Morgan's canon would be justified if it could be shown to be a special case—an animal behaviour-specific case—of the requirement to prefer simple explanations. Unfortunately, attempts to do this have not been successful. The first problem harks back to history: Morgan himself rejected the idea that simplicity is a proper criterion for theory choice in science, and argued that 'higher' explanations, because they are more anthropomorphic, are often simpler than 'lower' explanations [50, pp. 53–54]. The second, related and more general problem is that there are a great many kinds of simplicity, and, at best, Morgan's canon favours some kinds of simplicity at the cost of others [54]. An explanation can be simpler by virtue of: being easier to generate and understand (ease of use, the sense in which anthropomorphic explanations are simple); postulating a smaller number of causal entities or free parameters (ontological parsimony), iterations of a process (iteration parsimony), or evolutionary changes (phylogenetic parsimony); implying a lesser burden on limited resources (load, e.g. metabolic, memory); or by allowing similar observations to be explained by similar processes (uniformity, e.g. allowing superficially similar behaviour to be explained by the same processes across species) [54]. In any particular case, an explanation that is 'lower' in the context of Morgan's canon, and specifically an associative explanation, might score well on ontological parsimony, phylogenetic parsimony and uniformity, but poorly on ease of use, iteration parsimony and memory load. Therefore, to vindicate Morgan's canon, it would be necessary to show that it consistently favours a particular kind of explanatory simplicity, and to argue that this kind of simplicity is of overriding importance. Perhaps the best candidate for this queen of simplicities is ontological parsimony. However—and this is the third problem—Sober [55] has shown using 'model selection theory' that there is no reason to suppose that 'lower' explanations, in Morgan's sense, are generally more ontologically parsimonious than 'higher' explanations; surprisingly, even when a 'lower' explanation postulates only first-order intentionality (e.g. beliefs) and the corresponding 'higher' explanation postulates both first-order and second-order intentionality (e.g. beliefs about beliefs), the latter may have fewer free parameters.

In summary: advocates of simple-mindedness often assume that, when behaviour can be explained by associative processes and with reference to yet more cognitive processes, the associative account should be accepted. If any general methodological principle could justify this assumption, it is likely to be Morgan's canon, but Morgan's canon is not up to the task. Morgan's own justification for the canon is flawed, and subsequent attempts to support it with evolutionary arguments and by appeal to the virtues of simplicity have identified factors that sometimes favour associative explanations and sometimes favour super-cognitive explanations. They have not revealed considerations that consistently favour associative explanations; the kind necessary to support the use of Morgan's canon as a tie-breaker, and thereby the treatment of associative explanations as a default.

7. WHERE SHOULD WE GO FROM HERE?

I have argued that neither association-blindness nor simple-mindedness is a legitimate strategy when investigating animal minds. Viewed from a certain angle, these are both labour-saving devices; attempts to avoid (further) empirical work. If neither device is fit for purpose, we have to accept that in research on animal minds, as in other areas of science, the way to find out whether process X or process Y is producing a particular set of phenomena is to look for differential evidence. To find out whether behaviour is mediated by associative or super-cognitive processes we have to design and implement studies that test the two hypotheses against one another. More specifically, the two hypotheses need to be spelled out in sufficient detail to allow differential predictions—behavioural (or neural) effects that one would expect to see if the associative account is correct and not if the super-cognitive account is correct, and vice versa—and these predictions need to be tested through observation and experiment.

Unfortunately, just as there is no general principle that allows us to choose between associative and super-cognitive explanations without empirical work, there is no general formula telling us how to test these explanations against one another. However, Dickinson [33] & Shettleworth [34] provide many examples of the kind of empirical work that distinguishes associative from super-cognitive processes, and the nature of the task can be illustrated with a final look at the study by Horner *et al*. [1] on prosocial behaviour in chimpanzees. One way to pit the associative, paper rustling explanation of this behaviour against the super-cognitive, needs of others explanation, is to test whether the prosocial bias is stronger with some partners than with others. The associative account assumes that the actor's bias in favour of dual-reward tokens was supported solely by the sound of paper rustling—a cue that had been paired with the actor's own banana consumption—and therefore predicts that the magnitude of the bias will not vary with the relationship between actor and partner. In contrast, it seems reasonable to assume that, if chimpanzees know and care about the needs of others, they might know and care *more* about the needs of genetic relatives, close affiliates and individuals who recently did them a favour, than about the needs of unrelated, unfamiliar or unhelpful partners. Therefore, the super-cognitive explanation predicts that the prosocial bias will be greater in the former cases than in the latter. In effect, this test has already been conducted. Horner and colleagues did not make any reference to associative learning in their paper, but they provided support for the associative account by seeking and failing to find any effect of kinship, affiliation or reciprocity on the magnitude of the prosocial bias.

However, no single test is definitive. It is possible that the measure of prosocial bias used in the study by Horner and colleagues was not sensitive enough to detect subtle variations owing to the relationship between the actor and partner. Another way to test the associative account against the super-cognitive explanation would be to run 'ghost control' trials in which a partner is present in the cage adjacent to the actor, and the sound of unwrapping is heard from that cage, but the actor is able to see that the partner did not do the unwrapping or get the treat. Because the associative account assumes that the actor's bias was supported solely by the sound of paper rustling, it predicts that the bias would be sustained in these ghost control trials. In contrast, the super-cognitive account assumes that the sound of paper rustling was significant to the actors only as an indicator that the partner was receiving food, and therefore that her needs were being met. If this is correct, the dual-token bias should disappear in the ghost control condition. (See electronic supplementary material for further discussion.)

Whatever the outcome of this particular test, or of the current debate about cooperation in chimpanzees, the moral of my tale should by now be clear. If we really want to find out about animals' minds, we can afford neither to ignore associative learning, nor to assume that it reigns supreme. We need experiments—and then yet more experiments—to discover when Mother Nature has left the job to associative learning, and when she has devised a new super-cognitive gadget to support behavioural adaptation.

I am grateful to Martin Eimer, Bennett Galef, Mark Hazelgrove, John Pearce, Alex Thornton and an anonymous referee for their perceptive comments on an earlier draft of the manuscript.

REFERENCES

1 Horner, V., Carter, J. D., Suchak, M. & de Waal, F. 2011 Spontaneous prosocial choice by chimpanzees. *Proc. Natl Acad. Sci. USA* **108**, 13 847–13 851. (doi:10.1073/pnas.1111088108)
2 Macphail, E. M. 1982 *Brain and intelligence in vertebrates*. Oxford, UK: Clarendon Press.
3 Chandra, S. B. C., Wright, G. A. & Smith, B. H. 2010 Latent inhibition in the honey bee, *Apis mellifera*: is it a unitary phenomenon? *Anim. Cogn.* **13**, 805–815. (doi:10.1007/s10071-010-0329-6)
4 Acebes, F., Solar, P., Carnero, S. & Loy, I. 2009 Blocking of conditioning of tentacle lowering in the snail (*Helix aspersa*). *Q. J. Exp. Psychol.* **62**, 1315–1327. (doi:10.1080/17470210802483545)
5 Thompson, R. & McConnell, J. 1955 Classical conditioning in the planarian, *Dugesia dorotocephala*. *J. Comp. Physiol. Psychol.* **48**, 65–68. (doi:10.1037/h0041147)
6 Garcia, J., Ervin, F. R. & Koelling, R. A. 1966 Learning with prolonged delay of reinforcement. *Psychon. Sci.* **5**, 121–122.
7 Shettleworth, S. J. 2010 *Cognition, evolution and behavior*. Oxford, UK: Oxford University Press.
8 Wisenden, B. D., Chivers, D. P. & Smith, R. J. F. 1997 Learned recognition of predation risk by *Enallagma* damselfly larvae (Odonata, Zygoptera) on the basis of chemical cues. *J. Chem. Ecol.* **23**, 137–151. (doi:10.1023/B:JOEC.0000006350.66424.3d)
9 Roberts, A. D. L. & Pearce, J. M. 1999 Blocking in the Morris swimming pool. *J. Exp. Psychol. Anim. Behav. Process.* **25**, 225–235. (doi:10.1037/0097-7403.25.2.225)
10 Hollis, K. L., Dumas, M. J., Singh, P. & Fackelman, P. 1995 Pavlovian conditioning of aggressive behavior in blue gourami fish (*Trichogaster trichopterus*): winners become winners and losers stay losers. *J. Comp. Psychol.* **109**, 123–133. (doi:10.1037/0735-7036.109.2.123)
11 Adkins-Regan, E. & MacKillop, E. A. 2003 Japanese quail (*Coturnix japonica*) inseminations are more likely to fertilize eggs in a context predicting mating opportunities. *Proc. R. Soc. Lond. B* **270**, 1685–1689. (doi:10.1098/rspb.2003.2421)
12 Rescorla, R. A. 1988 Pavlovian conditioning: it's not what you think it is. *Am. Psychol.* **43**, 151–160. (doi:10.1037/0003-066X.43.3.151)
13 Kaufman, S. B., DeYoung, C. G., Gray, J. R., Brown, J. & Mackintosh, N. 2009 Associative learning predicts intelligence above and beyond working memory and processing speed. *Intelligence* **37**, 374–382. (doi:10.1016/j.intell.2009.03.004)
14 Moore, J. W., Dickinson, A. & Fletcher, P. C. 2011 Sense of agency, associative learning, and schizotypy. *Conscious. Cogn.* **20**, 792–800. (doi:10.1016/j.concog.2011.01.002)
15 Prados, J. 2011 Blocking and overshadowing in human geometry learning. *J. Exp. Psychol. Anim. Behav. Process.* **37**, 121–126. (doi:10.1037/a0020715)
16 Miller, N. Y. & Shettleworth, S. J. 2007 Learning about environmental geometry: an associative model. *J. Exp. Psychol. Anim. Behav. Process.* **33**, 191–212. (doi:10.1037/0097-7403.33.3.191)
17 Heyes, C. 2010 Where do mirror neurons come from? *Neurosci. Biobehav. Rev.* **34**, 575–583. (doi:10.1016/j.neubiorev.2009.11.007)
18 Catmur, C., Mars, R. B., Rushworth, M. F. & Heyes, C. 2011 Making mirrors: premotor cortex stimulation

enhances mirror and counter-mirror motor facilitation. *J. Cogn. Neurosci.* **23**, 2352–2362. (doi:10.1162/jocn.2010.21590)
19 Cook, R., Press, C., Dickinson, A. & Heyes, C. 2010 Acquisition of automatic imitation is sensitive to sensorimotor contingency. *J. Exp. Psychol. Hum. Percep. Perform.* **36**, 840–852. (doi:10.1037/a0019256)
20 Cook, R., Dickinson, A. & Heyes, C. M. In press. Contextual modulation of mirror and counter-mirror sensorimotor associations. *J. Exp. Psychol. Gen.*
21 Schultz, W. & Dickinson, A. 2000 Neuronal coding of prediction errors. *Annu. Rev. Neurosci.* **23**, 473–500. (doi:10.1146/annurev.neuro.23.1.473)
22 Jensen, J., Smith, A. J., Willeit, M., Crawley, A. P., Mikulis, D. J., Vitcu, I. & Kapur, S. 2007 Separate brain regions code for salience versus valence during reward prediction in humans. *Hum. Brain Mapp.* **28**, 294–302. (doi:10.1002/hbm.20274)
23 Wunderlich, K., Symmonds, M., Bossaerts, P. & Dolan, R. J. 2011 Hedging your bets by learning reward correlations in the human brain. *Neuron* **71**, 1141–1152. (doi:10.1016/j.neuron.2011.07.025)
24 den Ouden, H. E. M., Friston, K. J., Daw, N. D., McIntosh, A. R. & Stephan, K. E. 2009 A dual role for prediction error in associative learning. *Cereb. Cortex* **19**, 1175–1185. (doi:10.1093/cercor/bhn161)
25 Li, W., Howard, J. D., Parrish, T. B. & Gottfried, J. A. 2008 Aversive learning enhances perceptual and cortical discrimination of indiscriminable odor cues. *Science* **319**, 1842–1845. (doi:10.1126/science.1152837)
26 Behrens, T. E. J., Hunt, L. T., Woolrich, M. W. & Rushworth, M. F. S. 2008 Associative learning of social value. *Nature* **456**, 245–249. (doi:10.1038/nature07538)
27 Dickinson, A., Shanks, D. & Evenden, J. 1984 Judgement of act-outcome contingency: the role of selective attribution. *Q. J. Exp. Psychol.* **36**, 29–50.
28 Shanks, D. R. 2010 Learning: from association to cognition. *Annu. Rev. Psychol.* **61**, 273–301. (doi:10.1146/annurev.psych.093008.100519)
29 Brewer, W. F. 1974 There is no convincing evidence for operant or classical conditioning in adult humans. In *Cognition and the symbolic processes* (eds W. B. Weimer & D. S. Palermo), pp. 1–42. Leicester, UK: British Psychological Society.
30 Schultz, D. H. & Helmstetter, F. J. 2010 Classical conditioning of autonomic fear responses is independent of contingency awareness. *J. Exp. Psychol. Anim. Behav. Process.* **36**, 495–500. (doi:10.1037/a0020263)
31 Iselin-Chaves, I. A., Willems, S. J., Jermann, F. C., Forster, A., Adam, S. R. & Van der Linden, M. 2005 Investigation of implicit memory during isoflurane anesthesia for elective surgery using the process dissociation procedure. *Anesthesiology* **103**, 925–933. (doi:10.1097/00000542-200511000-00005)
32 Mitchell, C. J., De Houwer, J. & Lovibond, P. F. 2009 The propositional nature of human associative learning. *Behav. Brain Sci.* **32**, 183–98. (doi:10.1017/S0140525X09000855)
33 Dickinson, A. 2012 Associative learning and animal cognition. *Phil. Trans. R. Soc. B* **367**, 2733–2742. (doi:10.1098/rstb.2012.0220)
34 Shettleworth, S. J. 2012 Modularity, comparative cognition, and human uniqueness. *Phil. Trans. R. Soc. B* **367**, 2794–2802. (doi:10.1098/rstb.2012.0211)
35 Endo, N. & Takeda, Y. 2004 Selective learning of spatial configuration and object identity in visual search. *Atten. Percep. Psychophys.* **66**, 293–302. (doi:10.3758/BF03194880)
36 Shanks, D. R. 2007 Associationism and cognition: human contingency learning at 25. *Q. J. Exp. Psychol.* **60**, 291–309. (doi:10.1080/17470210601000581)
37 Dwyer, D. M., Le Pelley, M. E., George, D. N., Haselgrove, M. & Honey, R. C. 2009 Straw-men and selective citation are needed to argue that associative-link formation makes no contribution to human learning. *Behav. Brain Sci.* **32**, 206–207. (doi:10.1017/S0140525X09000946)
38 de Waal, F. & Ferrari, P. F. 2010 Towards a bottom-up perspective on animal and human cognition. *Trends Cogn. Sci.* **14**, 201–207. (doi:10.1016/j.tics.2010.03.003)
39 Byrne, R. W. & Bates, L. A. 2006 Why are animals cognitive? *Curr. Biol.* **16**, R445–R448. (doi:10.1016/j.cub.2006.05.040)
40 Greenwood, J. D. 1999 Understanding the 'cognitive revolution' in psychology. *J. Hist. Behav. Sci.* **35**, 1–22. (doi:10.1002/(SICI)1520-6696(199924)35:1<1::AID-JHBS1>3.0.CO;2-4)
41 Dickinson, A. 1980 *Contemporary animal learning theory.* Cambridge, UK: Cambridge University Press.
42 Wagner, A. R. 1981 SOP: a model of automatic memory processing in animal behavior. *Inform. Process. Anim. Mem. Mech.* **85**, 5–47.
43 Pearce, J. M. & Hall, G. 1980 A model for Pavlovian learning: variations in the effectiveness of conditioned but not of unconditioned stimuli. *Psychol. Rev.* **87**, 532–552. (doi:10.1037/0033-295X.87.6.532)
44 Chomsky, N. 1959 *A review of BF Skinner's verbal behavior*, pp. 48–66. (Reprinted in Ned Block, *Readings in the philosophy of psychology.* Cambridge, MA: Harvard University Press, 1980.)
45 Mackintosh, N. J. 1974 *The psychology of animal learning.* New York, NY: Academic Press.
46 Pearce, J. M. 2008 *Animal learning and cognition: an introduction,* 3rd edn. Hove, UK: Psychology Press.
47 Dwyer, D. M. & Burgess, K. V. 2011 Rational accounts of animal behaviour? Lessons from C. Lloyd Morgan's canon. *Int. J. Comp. Psychol.* **24**, 349–364.
48 Galef Jr, B. G. 1996 Historical origins. In *Foundations of animal behavior: classic papers with commentaries* (eds L. D. Houck & L. C. Drickamer), p. 5. Chicago, IL: Chicago University Press.
49 Frans 'de', W. 2002 *The ape and the sushi master.* New York, NY: Basic Books.
50 Morgan, C. L. 1909 *An introduction to comparative psychology.* London, UK: Walter Scott Publishing. (Original date of publication, 1894).
51 Morgan, C. L. 1892 The limits of animal intelligence. In *Int. Congress Exp. Psychol., 2nd Session, London, 1892*, pp. 44–48.
52 Allen-Hermanson, S. 2005 Morgan's canon revisited. *Philos. Sci.* **72**, 608–631. (doi:10.1086/505187)
53 Sober, E. 1998 Morgan's canon. In *The evolution of mind* (eds C. Allen & D. Cummins), pp. 224–242. Oxford, UK: Oxford University Press.
54 Fitzpatrick, S. 2008 Doing away with Morgan's canon. *Mind Lang.* **23**, 224–246. (doi:10.1111/j.1468-0017.2007.00338.x)
55 Sober, E. 2009 Parsimony and models of animal minds. In *Philosophy of animal minds* (ed. R. Lurz), pp. 237–257. Cambridge, UK: Cambridge University Press.