Human Minds

David Papineau

1. Introduction. Humans are part of the animal kingdom, but their minds differ from those of other animals. They are capable of many things that lie beyond the intellectual powers of the rest of the animal realm. In this paper, I want to ask what makes human minds distinctive. What accounts for the special powers that set humans aside from other animals?

Unfortunately, I shall not fare particularly well in answering this question. I shall explore some possible answers, but none will prove fully satisfactory. In effect, then, this paper will tell the story of a failure. Still, it is a story worth telling, for it is an interesting failure, I think, and one with significant morals for the study of human minds.

Before proceeding, let me put to one side one familiar answer to my question. Most people, if asked what distinguishes humans from animals, would probably answer—"language". Now, I certainly do not want to deny that our uniquely human facility with language plays some part in differentiating us intellectually from other animals. But it seems to me that, on its own, "language" does not add up to a satisfying answer to my question. For we still need to know what humans do with language. Does language yield distinctive human cognition because it enhances communication of facts, or because it facilitates social coordination, or because it allows records to be kept, or inferences to be drawn, or what?

Given some such hypothesis about the specific ability supported by language, it may turn out that language was constitutively necessary for that ability, in the sense that humans would not have had any distinctive such ability prior to the emergence of language. (For example, suppose that language was evolutionary significant specifically because it enhanced social coordination. Then one possibility is that no distinctive human powers of social coordination were available prior to the emergence of language.) On the other hand, it is also possible that the relevant ability preceded language, and that language evolved thereafter because it accentuated this ability. (On this scenario, distinctive human powers of social coordination would have come first, with language then being favoured by natural selection because it enhanced those powers.) Or, again, it may have been that the relevant ability co-evolved with language, with increased levels of one creating the evolutionary conditions for increased levels of the other, and vice versa.

However, we can ignore these alternatives here. For they all presuppose that there is some other ability distinctive to humans, apart from "language" itself, which explains the evolutionary significance of language. That is, language is important because it enables humans to do something else, be that social coordination, or inference-drawing, or whatever. My focus in this paper will be on this further distinctive ability, rather than the details of its evolutionary relationship with language.

Of course, it is not to be taken for granted that the intellectual contrast between humans and other animals should be explained by reference to the historical evolution of just one distinctive human ability.[1[1]] Maybe the evolution of a number of different abilities has contributed to the contrast (which different abilities could then have been evolutionarily related in various ways). Still, without denying this, I shall here set myself the limited task of identifying at least one ability which marks an evolutionary distinction between humans and other animals. We can worry about other similar abilities once we have succeeded in this limited task.

2. Means-End Reasoning.

In what follows I shall explore the idea that humans are distinguished from other animals by their powers of means-end reasoning. I shall consider various versions of this hypothesis, but the rough idea will be that animals are not capable of the kind of reasoned selection of means to desired ends that is found in humans.

I first became attracted to this idea as a result of thinking about 'Evolutionary Psychology'. Those who march under this banner ('Evolutionary Psychologists', with capitals, henceforth) embrace a number of commitments which go beyond the general idea that it is a good thing to bear evolutionary considerations in mind when thinking about human psychology (cf. Barkow, Cosmides and Tooby, 1992, Pinker, 1997). In particular, Evolutionary Psychologists advocate a strongly modular view of the human mind, viewing it as a battery of devices each devoted to some specific purpose, such as recognizing faces, selecting mates, detecting social cheats,

[1[1]] Nor is it to be taken for granted that any historically evolved differences between humans and other animals must be entirely genetic in nature. While there are undoubtedly important genetic differences between humans and other animals, the phenotypic intellectual powers that distinguish humans from other animals may well owe as much to non-genetic features of their cultural environment as to their genes. Cf. Deacon, 1997. (Note also that such non-genetic features can be vertically transmitted from parents to children, and thus subject to natural selection, in essentially the same way as genes are. Cf. Avital and Jablonka, 2000, Mameli, 2001, 2002.)

and so on. The standard metaphor is that of the human mind as a Swiss Army knife, containing a number tools each designed to perform some definite task.

However, this metaphor seems to rule out any account of how the overall selection of action is informed by the processing in the various specialized modules. It is noteworthy that humans seem able to reach decisions, form intentions, and make plans in a way that is influenced by a wide range of information about disparate subject matters. But how is this possible? Evolutionary Psychologists often seem blind to this issue. They often speak about people, and indeed animals, as 'deciding' what to do on the basis of the deliverances of their special-purpose modules (Cosmides and Tooby, 1992, pp 54, 113). But what system enables the deciding? Evolutionary Psychologists are generally suspicious of Jerry Fodor's 'central system', some non-modular part of the brain which in higher animals mediates intelligently between the deliverances of sensory input systems and behaviour (*op cit*, pp 49, 93). And perhaps they are right to reject this specific model for the intelligent guidance of behaviour. But, still, there must be some story to tell about the way human decision-making and planning can be informed by an open-ended range of judgements from disparate input modules.[2]

This line of thought suggests a possible answer to my original question. Maybe some power of integrated decision-making marks a division between humans and other animals. Perhaps other animals, unlike humans, have no way of integrating information from different sources and using it to make well-informed choices. That is, maybe the difference between human and animal cognition is that animals do not have the same intellectual wherewithal to select means to ends.

However, this thought is not easy to focus. It is not hard to see why. After all, nearly all animals have *some* ways of selecting suitable actions, some way of generating behaviour appropriate to their current circumstances on the basis of various kinds of sensory information. So some more precise specification of 'means-end reasoning' is needed, if we are to have any hope of showing that 'means-end reasoning' is peculiar to humans. 'Means-end reasoning' can't include *any* ways of gearing behaviour to circumstances, for even sea cucumbers have some of those. Rather, we need to specify a cognitive structure which selects actions in some particular sophisticated matter, and then argue that this specific mechanism is present in humans but not other animals.

In the main body of this paper I shall explore a sequence of hypotheses about such a specifically human cognitive structure. None of these hypotheses stands up. In each case it turns out that there is some well-attested species of animal behaviour that displays 'means-end reasoning' in precisely the specified sense.

So in the end I shall fail to find a satisfactory answer to my original question. Still, this does not necessarily mean that the search will have been fruitless. Much can be learned by exploring hypotheses that eventually turn out to be empirically flawed, and I would say that the path I have taken does much to illuminate the range of cognitive structures available to humans and other animals. But you do not have to take my word for this. Let me fill in the story, and you can judge for yourself whether it is one that is worth telling.

3. *Inferential Limitations.* My first attempt to identify a distinctive mode of human means-end reasoning involved this hypothesis: non-human animals can't piece together representations of disparate causal facts to infer that some behaviour B is good for some outcome O, unless they or their ancestors have previously experienced Bs leading to Os.

Note that this is not to claim that non-human animals never use *any* causal representations of the form *B will produce O* in selecting behaviour. As I shall explain in a moment, I take there to be a good sense in which even very simple animals do that. Rather the claim is that non-human animals are incapable of combining *different* items of causal information to select *novel behaviour*, where this is defined as behaviour B which is done in pursuit of O even though neither the agent not its ancestors have ever experienced B as leading to O.

Let me elaborate. First let me explain why I take even very simple animals to use a kind of causal representation. This will then bring out why there might be a specific problem with novel behaviour.

In my view, animals use representations of causal facts to guide their behaviour as soon as their cognition is complicated enough to involve *drive* states. By a drive state I mean a state whose purpose is to get the animal to perform behaviours that are good for getting some specific outcome like food, say, or water, or sex, or avoiding danger, or so on. I take it that relatively simple animals, such as fish, have such states, in that they will only engage in feeding behaviour, say, when they are hungry. Suppose now that some such animal has some

[2] Note that my worry here is different from the complaint that Evolutionary Psychology lacks a mechanism to decide which module to activate in which circumstances. I see no reason why the brain should not be structured so that this problem takes care of itself (pace Fodor, 2000). My complaint is more specific: we need some system that will allow information from different modules to be combined in selecting behaviour. Rather than asking for something to control the modules, I'm in effect asking for an extra module, to do means-end reasoning. (Cf. Papineau, 2001, sects. 1 & 5.)

behaviour (B) which it is disposed to perform under a given conditions (C) if a drive directed at some outcome (O) is activated. Moreover, suppose that the animal is innately so disposed because its ancestors who did B in C succeeded thereby in getting O.

In such a case, I say, we should regard their drive as representing the outcome O. And correspondingly we should regard the innate disposition to do B in C given D as representing the causal fact that: behaviour B in condition C will produce outcome O. After all, by hypothesis the biological purposes of the drive state is to generate (behaviour which will lead to) the outcome O. In line with this, the behavioural disposition will serve its biological purpose insofar as it is indeed the case that behaviour B in condition C will produce outcome O.[3[3]]

Some readers may object that this latter information, that B in C will produce O, is at best represented procedurally, not declaratively. After all, the vehicle of the representation is only a disposition to behaviour, not any sentence-like object in some language of thought. However, I am uneasy about placing any weight here on the distinction between procedural and declarative representation. After all, dispositions to behaviour are not ethereal traits, but must have some physical basis: there must be physical differences between animals who have the disposition and those who lack it. Moreover, note that these physical features will enter into a kind of rudimentary practical inference when they interact with active drives to generate behaviour in a way that is appropriate to their putative representational contents: thus, the drive 'for O', plus a perception 'that C', will interact with the disposition embodying the information 'that B in C will lead to O', to generate the behaviour B. The disposition may not seem particularly sentence-like, but this doesn't stop it here operating in just the way a sentence-like representation would in generating a practical inference appropriate to its content.

So I have no qualms about speaking of representations of causal facts as soon as we have animals with drives and associated innate behavioural dispositions. However, while these causal representations will interact with drives and perceptions of current circumstances in rudimentary practical inferences, they won't necessarily enter into another kind of inference. Simple animals whose causal information is embodied only in innate behavioural dispositions won't be able to piece together separate items of such information to figure out any further links between means and ends.

Let me illustrate. Suppose that some primate is disposed to shake apple trees to dislodge the fruit when it is hungry, and also disposed to throw any handy apples at predators when threatened. This by itself won't be enough to enable it to figure out that it should shake the trees when it is threatened and no apples are to hand, because nothing in the cognitive structure specified will make a threatening predator, as opposed to hunger, a stimulus to shaking trees. It will have the information that 'shaking produces apples' and that 'throwing apples will repel predators', but won't be able to 'chain' these two general claims together to draw the relevant inference.

Of course, if some of its ancestors had genes which disposed them to shake the trees when predators appeared, then these genes would presumably have been selected, assuming those ancestors also had the disposition to throw the apples to repulse the predators. And this would then have instilled a further innate disposition in the primate, to shake the trees when threatened by predators. But the point remains that the two originally posited innate dispositions can be present without this further innate disposition, and then the organism won't be able to figure out the further implication. So here we have a precise sense in which organisms who embody general information about means to ends solely in their innate behavioural dispositions won't be able to perform novel behaviours. They won't perform B in pursuit of O in condition C unless their ancestors achieved O as a result of doing B in C and were genetically shaped accordingly. It's no good being innately disposed to shake the trees for apples, and being innately disposed to throw apples to repel predators, if your ancestors weren't also directly genetically selected shake the trees when threatened by predators.

Nor is the situation substantially altered if we switch from innate behavioural dispositions to those instilled by instrumental learning (that is, 'operant' or 'Skinnerian' conditioning). Here an organism may become disposed to do B in C in pursuit of O, not because B in C led to O in its ancestral past, but because B in C led to O in the individual organism's experience, and this reinforced its disposition to do B when C. (Gross, 1996, p 161.) Here the cause of the disposition is different—individual rather than ancestral experience—but the resulting structure remains just the same. The information that B in C will yield O will be embodied in the organism's disposition to do B when it has a drive for O and a perception of C. And, given that the information is embodied in this way, the organism won't be able to combine separate items of such information to figure out that some new behaviour is good for some result in some circumstances, when it hasn't itself experienced that behaviour as leading to that result in those circumstances. So, to adapt the above example, an organism that has been conditioned to shake apples trees for fruit when it is hungry, and has also been conditioned to throw apples at

[3[3]] In general I understand representation in 'teleosemantic' terms: the representational contents of cognitive states should be analysed in terms of the conditions required for them to serve their biological function. Cf. Millikan, 1984, 1989, Papineau, 1984, 1993. For an explanation of why representational content requires at least specialized drive states, see Papineau, 1998; and for more on the application of teleosemantics to behavioural dispositions, see Papineau, 2001, sects. 2 and 3.

predators when threatened, won't automatically shake the trees when threatened by predators, because shaking trees, as opposed to throwing apples, won't have been conditioned to the predator stimulus.

So, just as before, novel behaviour will be beyond the reach of the organism. True, instrumental conditioning can lead you to perform B in pursuit of some result O that none of your *ancestors* obtained from B. But this still requires that you yourself have previously obtained O after performing B. We still have no process that will lead you to perform B in pursuit of O when neither you nor your ancestors have experienced O following B.[4[4]]

Before proceeding, let me make one brief comment about conditioned learning. In what follows I shall refer at various points to instrumental and other kinds of associationist learning. I would like to make it clear that these references carry no implication that associationist learning is more important than genes in constructing cognitive systems in animals or even humans. For all I say in this paper, cognition may be largely hard-wired, and conditioning may do no more than fine-tune pathways laid down by genes. My interest in associationist conditioning here is largely hypothetical: to the extent that it does play a part, does it lead to new kinds of cognitive architecture? And the point I have just made is that it does not, at least as far as the impact of instrumental conditioning on novel behaviour goes.

4. *The Power of Classical Association*. So there is the initial thesis. Non-human animals are not capable of novel behaviours, that is, not capable of choosing a means to an end in some circumstance when neither they nor their ancestors have previously experienced that means as producing that result in that circumstance.

Unfortunately, the thesis can easily be shown to be false. Animals can embody causal information in what I shall 'classical associations', as well as in dispositions to behaviour, and when these classical associations are combined with behavioural dispositions, then the upshot can well be novel behaviour in the above sense.

By a classical association I mean a disposition to move from a particular judgement S to another particular judgement T. Thus an animal might be disposed to move from *a change in light intensity* to *an edge of an object*, or from *a moving shadow* to *a hawk is overhead*, or from *the sound of a bell* to *food is arriving*.

Such associations can be innate, or can derive from learning. In the latter case, the relevant mode of learning will be classical or 'Pavlovian' conditioning, rather than operant or 'Skinnerian' conditioning. I shall use a familiar example of Pavlovian conditioning to illustrate the way in which classical associations give rise to novel behaviour in the sense specified in the last section. Since pretty much all animals are capable of Pavlovian conditioning, this will show that novel behaviour in this sense is effectively universal in the animal realm.

Pavolvian conditioning does not involve the reinforcement of some behaviour by a reward, as in instrumental conditioning, but rather the association of two stimuli: animals who have experienced stimulus S being followed by stimulus T will come to respond behaviourally to stimulus S in ways they previously responded to stimulus T. (You can think of the association as ensuring that the registration of stimulus S will 'activate' the state which normally registers stimulus T, and thereby will stimulate any behaviour that was previously triggered by stimulus T.) For example, a dog who has experienced the sound of a bell being followed by the nearby presentation of food will come to respond to the bell in ways it previously responded to the sight of food. For example, the bell alone will now make it approach the expected site of the food when hungry, in the way the sight of food itself previously did. (Gross, 1996, p 157.)

This now immediately gives us an example of a novel behaviour in the relevant sense. Neither the dog nor any of its ancestors need previously have derived any advantage from approaching in response to a bell alone when hungry, yet classical conditioning will bring it about that the dog now does this.

It will be helpful to think about the process in representational terms. Suppose the animal starts out disposed to do B, in circumstances T, given a drive for O. (It is disposed to approach the food, given a drive to eat it.) Then, as argued in the last section, we can view the embodiment of this disposition as representing that *B in T will lead to O*. (Approaching food leads to eating.) Now suppose in addition that classical conditioning leads the dog to associate stimulus S with stimulus T, so that, when it registers S, this activates the state which normally registers T. We can think of the embodiment of this association as representing that *all Ss are Ts*. (Bells are followed by food.) Then we can view the new *behavioural* upshot of the classical conditioning, namely, the disposition to do B in the new circumstances S, given a drive for O (the dog now approaches when the bell sounds, given a drive to eat) as representing the fact that *B in S will lead to O* (*approaching when the bell sounds will lead to eating*). Moreover, we can regard this last claim as the conclusion of an inference from the two already attributed premises that *all Ss are Ts* and that *B in T will lead to O*. The organism puts together these two claims and draws the obvious inference that *B in S will lead to O*. It is thereby led to perform a novel behaviour—doing

[4[4]] Some readers may be wondering whether the phenomenon of 'secondary reinforcement' would produce the requisite novel tree-shaking behaviour. I shall discuss secondary reinforcement in section 5, and its relevance to the tree-shaking example in footnote 5.

B in S in pursuit of O—even though neither it nor its ancestors have ever done B in S before (the dog has never previously approached when hungry in response to the sound of a bell).

So this is certainly one sense in which non-human animals can perform novel actions. However, this line of reasoning suggests that there may be another species of novel action which may be beyond them. The 'inference' I have just described allows animals to move from B in T will lead to O to B in S will lead to O. But such inferences won't ever allow animals to figure out that some behaviour B is good for some result O unless they or their ancestors had previously experienced B as leading to O in some circumstances. Classical associations may allow them to transfer this knowledge from one circumstance to another, so to speak, but perhaps the underlying B-O means-end relation always needs to be grounded in direct individual or ancestral experience of B leading to O.

5. Acquired Desires. But this idea doesn't stand up either. Consider the phenomenon known as secondary reinforcement. Standard learning theory tells us that some circumstance P that is not initially rewarding to an animal can come to acquire a positive value as a result of experiences which lead the animals to associate P with something already rewarding. Put it in more familiar terms, the animal comes to desire things it experiences as precursors or means to things it already desires. For example, suppose that an animal habitually passes some landmark on its way to feeding. Then it will come to desire to pass the landmark in itself. Moreover, passing the landmark will come to function as a reward on its own, as will be shown by its ability to reinforce other behaviours, even when it is not followed by feeding. (Of course, continued experience of the landmark not being followed by food will reverse the process, and render the landmark neutral in affect once more.) (Gross, 1996, p 164.)

Now, secondary reinforcement can bring it about that animals will perform novel behaviour in the strong sense specified at the end of the last section: that is, they will perform some B in pursuit of O even though neither they nor their ancestors have ever experienced O after doing B. (Let me call this 'strongly novel' behaviour henceforth.)

I can usefully illustrate the point by describing an experiment of Anthony Dickinson's (Dickinson and Dawson, 1988, 1989, Heyes and Dickinson, 1990). In the first stage, rats are trained while hungry but not thirsty, in an environment where they gain dry food pellets from pressing a lever, and a sucrose solution from pulling a chain. Both the pellets and the sucrose solution satisfy hunger. If the rats were thirsty, however, only the sucrose solution would satisfy their thirst.

This prompts an obvious question: what will the rats do if they are thirsty? Will they pull the chain which delivers the sucrose solution, rather than press the lever? In fact they won't do this straight off. But provided they are given an opportunity to drink the sucrose solution when they are thirsty, even in circumstances quite removed from the experimental apparatus, they will then differentially pull the chain when they are next placed in the apparatus when thirsty.

This is now strongly novel behaviour. The rats are pulling the chain in order to quench their thirst, even though neither they nor their ancestors have ever quenched their thirst by pulling the chain before.

Dickinson himself takes this experiment to show that rats are capable of genuine cognition, involving the manipulation of some kind of sentence-like representations, and thus are more than simple associationist systems. As he sees it, the rats must have acquired the information that chain-pulling leads to sucrose solution from their original training. Later they learned that sucrose solution quenches thirst. And then they put the two items of information together, to draw the inference that chain-pulling is the thing to do if you are thirsty.

I agree that the rats can usefully be viewed as performing this inference. However, I see no reason to conclude with Dickinson that this elevates the rats beyond associationist systems and into some separate realm of genuine representation and inference, involving the manipulation of sentence-like representations. It is true that the rats must somehow be able to remember, from their original period of training, that the chain-pulling leads specifically to the sucrose solution. Moreover, there was nothing differentially rewarding about the sucrose solution, as opposed to the food pellets, in that original period of training—both sucrose solution and food pellets alike satisfied hunger. This may indeed make it seem that the information that chain-pulling leads to sucrose solution must be stored in some non-dispositional sentence-like representation—after all, since the sucrose solution wasn't differentially rewarding, it is not clear how the information that chain-pulling leads to sucrose solution could have become embodied in some specific disposition to chain-pull in pursuit of sucrose solution.

However, recall the possibility of secondary reinforcement. Since the rats, in their original training, experience the sucrose solution as preceding hunger satisfaction, the sucrose solution will have become a secondary reinforcer. In the terms I used earlier, the rats will 'acquire a desire' for sucrose solution as such. Moreover, when this 'desire' is satisfied it will act as a reinforcer, and so the rats will have become disposed to perform behaviours when the sucrose desire is active which in their experience have led to sucrose solution—thus in the case at hand, they will become disposed by their original training to chain-pull when they are in the experimental apparatus and desire sucrose solution.

Then later, after being given sucrose solution when they are thirsty, they will associate sucrose solution with thirst satisfaction, and consequently be disposed to activate their desire for sucrose solution when they are thirsty. And then they can put this together with the prior disposition, instilled by their original training, to chain-pull when they desire sucrose solution. The overall result, then, is that they will chain-pull when they are thirsty, even though neither they nor their ancestors have ever quenched their thirst by chain-pulling before.[5[5]]

My analysis thus agrees with Dickinson in allowing that the rats are inferring the appropriateness of some behaviour (chain-pulling) to some end (thirst quenching) as a result of embodying the separate items of information that chain pulling will lead to sucrose solution, and sucrose solution quenches thirst. But I disagree with Dickinson's view that these items of information need to be embodied in some explicit sentence-like manner, open to general logical manipulation, as opposed being embodied in dispositions to behaviour which can be combined in the way sketched above. I am happy to view the rats as performing an inference. But they do this by deriving a complex disposition from the combination of two other dispositions, rather than by manipulating explicit sentence-like representations. Once they are disposed to desire sucrose when thirsty, and disposed to chain-pull when they desire sucrose, then they will derivatively chain-pull when thirsty, and therewith derive the conclusion that chain-pulling is a means to quenching thirst.[6[6]]

6 Observation versus Experience. Dickinson's experiment certainly shows that rats can perform strongly novel actions, that is, that they can do some B in pursuit of some O even though neither they nor their ancestors ever did B in pursuit of O before. But if I am right about the rats embodying the relevant information in dispositions to action, rather than in some sentence-like format, it remains possible that rats are limited in another way. Maybe they are incapable of learning from observation, as opposed to learning from experience. Indeed, perhaps this inability differentiates all other animals, and not just rats, from humans.

Let me explain. In the story I have just told, I credited the rats with various items of information to the effect that some action in some situation will lead to some result. Their potential for strongly novel actions then derived from their ability to piece such items of information together. They 'knew' that chain-pulling leads to sucrose solution, and that sucrose solution quenches thirst, so they were able to 'infer' that chain-pulling is a means to thirst quenching. But note that, in order to acquire the original items of means-end information, the rats needed to have performed the relevant action themselves, and needed themselves to have experienced the reward of the relevant result. The rats acquired the relevant information because they had experienced their own chain-pulling as leading to their getting sucrose solution, and their own consumption of sucrose solution as quenching their own thirst.

This means that, for all that has been said so far, the rats will have no way of observing some other animal performing some B and getting some result O, and on this basis acquiring the information that B leads to O. Still less will they be able to observe inanimate nature 'performing' some action B which leads to O, and thence inferring that B is a means to O.

I can usefully illustrate the point with an anecdote.[7[7]] The trainers of a troop of monkeys on a research station in Puerto Rico occasionally reward the monkeys by putting coconuts in the camp fire; the coconuts then burst open, making the tasty flesh available to the monkeys. However, the monkeys seem unable to learn from this that they can put the coconuts in the fire themselves. Moreover, even when one particular monkey somehow acquired the trick, the other monkeys seemed not to cotton on that they could do it themselves.

Given the points made in this paper so far, this needn't seem so surprising. So far I have considered cases where animals acquire the information that B will lead to O because they (or their ancestors) have themselves performed B and themselves later received O. But the mechanisms behind this will be blind to the observation of another animal doing B and getting O. After all, there is nothing rewarding, or otherwise advantageous, to the observer in seeing another animal enjoying outcome O. And even if the observer does get to

[5[5]] Consider the earlier example of an animal conditioned to shake apple trees when hungry, and to throw apples when threatened by predators. In section 3 I argued that this alone wouldn't make it shake the trees when threatened. But if its hunger gives it a secondary desire to have apples to hand, the lesson of Dickinson's rats will apply here to: the experiences that conditioned the primate to tree-shake-when-hungry will also dispose it to tree-shake-when-it-desires-to-have-apples-to-hand. And if the appearance of predators also triggers the secondary desire to have apples to hand, the appearance of predators will derivatively trigger tree-shaking.

[6[6]] Given the structural similarity we have observed so far between learned and genetically fixed behaviours, some readers may be wondering whether there is a innate analogue of the process by which dispositions resulting from secondary reinforcement give rise to novel behaviour. We will indeed find such an analogue, provided we are prepared to posit sufficiently fine-grained innate desires. Imagine that a primate is innately disposed to desire apples when threatened by predators, solely because of ancestral events involving predators, and innately disposed to shake trees when it desires apples, solely because of ancestral events involving food deprivation. Then this could lead it to shake the tree when threatened, even though none of its ancestors ever had ever done this before.

[7[7]] I was told this story by Ned Block, who in turn acquired it from Marc Hauser. It won't matter too much if some of the signal has been lost in the transmission, since I intend the anecdote only to illuminate the logic of my analysis, not to provide empirical backing.

enjoy the reward—it shares the coconut flesh, say—this still won't do the trick. For this reward won't reinforce the behaviour B—placing coconuts in the fire—since the observer hasn't itself performed this behaviour. The observer didn't place a coconut in the fire prior to the reward—it was just sitting there watching.

It is true that observation of another animal doing B and getting O can give rise to classical conditioning. The sight of the other animal doing B can come to make the observer anticipate O. In the coconut example, the observers may come to respond to the sight of the coconut going onto the fire with their pre-existing responses to food, such as salivating and approaching. But this will do nothing to get the observers doing B themselves. The classical association will make you salivate and approach when you see another animal putting a coconut in the fire—it won't get you putting the coconut in the fire in the first place.

So here is another possible way in which human intellects may outstrip those of other animals. Perhaps animals are unable to learn about means to ends from observation. Seeing another animal doing B as a means to O won't help them to do B in pursuit of O. Yet humans clearly can learn in this observational way. Indeed humans can draw such lessons from inanimate nature, as well as from animate agents. (If I saw a coconut landing in a fire by chance after falling from a tree, and then bursting, I would infer that I myself can also burst coconuts by putting them in fires.)

In a moment I shall consider whether this ability to learn from observation does indeed mark a difference between human and animal cognition. But first it will be helpful to make some related points.

7 Mimicry, True Imitation and Empathy. Some readers may be wondering how the issue of learning from observation relates to the topic of animal imitation. There is no doubt that animals often learn behaviour from other conspecifics. One oft-cited example is the rapid spread in the 1940s among British blue tits of the ability to peck open the tops on milk bottles to get at the cream inside. Potato-washing in the sea by Japanese macaques is another frequently mentioned case. Again, patterns of tool-use among both chimpanzees and crows are known to vary between populations within species, suggesting that these behaviours too are copied from conspecifics.

There is no question of engaging with extensive literature on animal imitation in this paper. (For a survey, see Shettleworth, 1998, ch. 10.) Let me content myself by making what I take to be two uncontentious points.

First, while there is no question that patterns of behaviour can spread from some animals to others, as in the examples just mentioned, it is a further issue whether such 'social learning' requires any specific imitative abilities. Thus, one possible explanation for the standard examples is simply that animals tend to follow each other around. Because of this, when animals who are expert in some behaviour go to the sites (milk bottles, sea shores) where they can practice their craft, novices will follow them, and thus be led to those special places where ordinary trial-and-error instrumental learning can instil the relevant behaviour. Without the experts to lead them, they wouldn't be in the right places for their ordinary behavioural experimentation to yield the relevant rewards. Again, another obvious explanation for some examples of 'social learning' is simply that animals can learn from others where certain things are. If I see an expert roll over a log to find grubs, then I will become aware that grubs lie under logs, and thereafter use my pre-existing abilities to remove obstacles to uncover the grubs myself. (Cf. Tomasello, 2000.)

Second, even when there is evidence for specific imitative abilities, these not involve any appreciation of causal links between behaviour and outcome. Let us define mimicry as a tendency for an animal to repeat behaviour that it observes in another conspecific. Clearly an animal might be capable of mimicry, even if it is not able to appreciate what the behaviour in question is good for. It would then do B simply because it had observed another animal doing B, and not because it appreciated that B would lead to some O. It would be 'parroting', so to speak—it would simply be copying the behaviour, without understanding its significance.. There is a large amount of evidence that some animals are capable of mimicry in this sense. [Ref?] But, as just observed, this won't amount to learning from observation in the sense of learning that there is a connection between B and some attractive further result O. Mimicry per se may connect your own behaviour with the observation of others performing the same behaviour, but it won't connect your behaviour with any intended outcomes.

Henceforth let me adopt the phrase 'true imitation' for the more sophisticated ability to learn, from observing other animals, that some behaviour B is connected with outcome O. As I observed at the end of the last section, it is clear that humans have this ability, even if other animals do not. So let me offer one speculation about the mechanism behind this ability. (This speculation can be detached from the rest of my argument, but I think it is of some interest in its own right.)

My speculation is that true imitation arises once an 'empathetic faculty' is added to a capacity for parrot-like mimicry. Suppose that, when you observe someone else getting something that you yourself desire, you undergo some vicarious satisfaction as a result of the observation. For example, when you are hungry and see someone else eating, you simulate their hunger satisfaction with a 'faint' version of your own.

Now put this empathetic faculty together with a capacity for mimicry. Take a case where you desire O, and observe someone else doing B and getting O. Your basic tendency to mimicry inclines you to do B. Your empathetic faculty then gives rise to a faint simulation of the satisfaction you would derive from O. This vicarious reward will then reinforce, via normal instrumental conditioning, your tendency to do B when you desire O. The result is thus that your observation of B leading to O leads to your becoming disposed to do B when you desire O. So this gives us a mechanism whereby the observation of some other animal getting O from B can give rise to your acquiring the information that B leads to O. As before, this information will be embodied in a disposition to do B when you desire O, but now we have an account of how this disposition can be instilled by observation rather than by first-hand experience.[8[8]]

At this point, let me make some brief observations about imagination and means-end reasoning. It is a familiar thought that the ability to connect previously unperformed behaviours with intended outcomes is somehow facilitated by sensory imagination—we figure out that B is a means to O by imagining B being followed by O. This use of imagination might seem to offer a more basic and general mechanism for innovatory means-end reasoning than that provided by imitative learning from observation. However, I think that this puts the cart before the horse. As I see it, the power of imagination to inform means-end reasoning depends on imitative learning, rather than vice versa.

Let me explain. I take it that sensory imagination activates some of the same parts of the sensory cortex as would be activated by genuine observation of a similar scenario. When I imagine seeing a red square, this activates some of the same parts of my visual cortex as would be activated if I were really looking at a red square. We might here recall Hume's terminology, according to which sensory imagination is a 'faint replica' of the real thing.

However, if this is the right picture of sensory imagination, then it is unclear how imagining a means-end sequence can be a more basic route to action than actually observing it. If really seeing someone doing B and getting O isn't enough to get you doing it yourself, then 'faintly seeing' an imagined person doing the same seems even less likely to do the trick, for just the same reasons.

Of course, once true imitation does emerge, then we can expect sensory imagination to inform means-end understanding, though not as some separate mechanism, but as a corollary of true imitation. The basic mechanism behind true imitation, as I have told the story, is that actual observing a conspecific doing B and getting O can lead, via mimicry and vicarious reinforcement, to you yourself doing B in pursuit of O. However, if sensory imagination is a 'faint version' of actual observation, then we would expect it to produce the same result for similar reasons. In effect, you will be led to imitate the imagined person's pursuit of O by B. Thus, visually imagining someone doing B will lead, via the tendency to mimicry, to a disposition to do B yourself; and then imagining the visualized person receiving O will lead, via empathy, to your own vicarious satisfaction—and thus you will acquire a disposition to do B in pursuit of O via instrumental conditioning, as before.

8 Japanese Quails. The overall story I have told so far implies one definite prediction. Non-human animals who learn by observing other animals will be 'insensitive to demonstrator reward'. They will be capable of 'mimicking' the behaviour of conspecifics, but will do so with no appreciation of outcomes, and so will not learn differentially depending on whether or not their demonstrator's behaviour leads to some rewarding outcome. According to my latest hypothesis about the distinctive feature of human cognition, only humans can truly imitate, in the sense of copying an action just in case you have observed it leading to some result that you yourself desire.

A wide range of empirical data are consistent with the prediction of animal insensitivity to demonstrator reward. Thus Sara Shettleworth, in her comprehensive text Cognition, Evolution and Behaviour (1998) says '. . . whether or not the observer must also see the demonstrator obtain a reinforcer . . . is a question that has hardly been tackled' (p 473), and again '. . . the role of demonstrator reward has been little studied' (p 473).

However, a particular series of recent studies by Thomas Zentall and his associates shows clearly that there is at least one animal species that are sensitive to demonstrator reward—Japanese quails. Akins and Zentall (1998) trained demonstrator Japanese quails to either peck at or step on a treadle. They then allowed other Japanese quails to observe this behaviour. Their findings were that the observer quails copied the demonstrator's behaviour only if they also observed the demonstrator receiving a food reward for the behaviour.

[8[8]] There are obvious affinities between the idea that true imitation derives from the empathetic ability to experience vicarious satisfaction and recent work on 'understanding of mind', particularly simulationist accounts thereof (see Davies and Stone, 1995a and 1995b, Carruthers and Smith, 1996). Tomasello (2000) also suggests that the capacity for true imitation depends on understanding of mind, but for rather different reasons from mine: he does not view the widespread absence of true imitation as due to the inability of standard associationist mechanisms to allow observational learning of means-end connections; relatedly, he takes understanding of mind to be important simply because it allows observers to appreciate what their demonstrators intend, not because it yields empathy.

Interestingly, a further study (Dorrance and Zentall, 2001) showed that this effect required the observers to be hungry when they observed the demonstrator's behaviour. It wasn't enough that they be hungry when they were later placed in the apparatus and given the opportunity to peck at or step on the treadle. It turned out that even hungry observer quails wouldn't display the relevant behaviour at this later stage, if they hadn't also been hungry at the earlier observational stage.

These experiments clearly indicate that Japanese quail are capable of true imitation of the kind I have hypothesized to be peculiar to humans.

Moreover, the second study by Dorrance and Zentall suggests that quails' imitative powers may hinge on just the kind of empathetic identification with the demonstrator that I speculated may be the basic mechanism behind human imitation. This would explain the striking fact that the quails won't imitate unless they are hungry at the time of observation. At first sight, this can seem puzzling: why can't the quails just store the observationally-derived information that pecking at the treadle, say, yields food, and then use this later when they are hungry? Why should they need to be hungry at the time of observation in order to acquire the information? However, if the route from observation to behaviour proceeds via reinforcement of mimicking tendencies by empathetic reward, as outlined in the last section, then the Dorrance and Zentall finding becomes unpuzzling. The observer quails won't feel any empathetic reward at the sight of another feeding, unless they themselves are hungry.

Of course, other explanations remain possible. Maybe the function of observer hunger is simply to make the observers interested in matters to do with food. Perhaps they don't pay attention to what the demonstrator is up to, if they aren't hungry. If this is right, then perhaps there is some quite different mechanism behind the quails' sophisticated imitative abilities, nothing to do with the empathetic reinforcement model sketched in the last section.

Alternatively, perhaps reinforcement is involved, but not in a way that involves empathy. Consider this possibility. Quails are social creatures, and so in the normal course of events will often have observed others eating while they themselves are feeding. Because of this, the sight of another quail feeding could come to function as a secondary reinforcer—after all, this visual stimulus will characteristically have been experienced as preceding hunger satisfaction. This secondary reinforcer could then combine with basic mimicry to explain the quails' imitative abilities: their observation of the demonstrator will trigger their mimicking tendencies—and then these tendencies will be secondarily reinforced by the sight of the demonstrator eating. Moreover, this story also promises to explain why the learners have to be hungry when observing. If the sight of others feeding derives its status as a secondary reinforcer from experience associating it with hunger satisfaction, then it can be expected to function as a secondary reinforcer only when the observer is hungy.9[9]

Let me not continue. The precise mechanism behind the quails' abilities is clearly an empirical matter, to be decided by further experimental investigation, not by speculation. (I leave it as an exercise for readers to design experiments to decide between the three mechanisms suggested above.)

In any case, Japanese quail provide a counter-example to the hypothesis that only humans are capable of true imitation. True, once we discover the quails' mechanism, it may turn out that their imitative ability is relatively superficial, resting on some idiosyncratic quirk of their psychology, such as secondary reinforcement by observations of others eating, in which case it may be possible to argue that some more powerful species of empathy-involving imitation is peculiar to humans after all. Alternatively, however, it may be that just the same empathy-involving mechanism underlies true imitation in both humans and Japanese quails, and so presumably in many other species too, in which case the distinctive features of human cognition must lie quite elsewhere.

Still, as I said, these are empirical matters, and I do not propose to offer any further hostages to empirical fortune. None of my hypotheses about the special power of human cognition have stood up to the empirical data, and at this stage I have no further replacements to offer. Rather, I would like to conclude by drawing three general morals from my frustrated search for the key to human cognition.

9 General Morals

First Moral: The Significance of Observation. I hope I have persuaded readers that the ability to learn from observation is important, whether or not it is anything to do with the distinctive features of human cognition. By 'learning from observation' here I mean specifically the ability to acquire information about potential means-end connections by observing another organism getting the end from the means, as opposed to performing the means and enjoying the end yourself.

In the course of this paper I have showed how standard mechanisms of associationist learning, namely, instrumental, classical and secondary conditioning, can generate various species of novelty, informing organisms that given behaviours will lead to given ends in given circumstances, even when neither the organisms nor their ancestors have experienced those behaviours as leading to those ends in those circumstances. However, all such

9[9] This possible explanation was suggested to me in conversation by Cecilia Heyes.

associationist conclusions must be derived from pieces of information which are based on individual or ancestral experience. They can only deal with connections between actions previously performed at first hand and results previously experienced at first hand (even in cases where those specific results haven't previously followed from those specific actions, as with Dickinson's rats). Standard associationist mechanisms therefore offer no way of learning means-end connections from external observation rather than first-hand experience.

So, however the trick is done, the ability to learn about means-end connections from observation rather than experience marks a significant advance in cognition. Of course, there are many other facets to advanced means-end reasoning in humans. I earlier touched on a possible role for sensory imagination in making connections between potential means and ends. I also mentioned the ability to learn about potential means-end connections by observing inanimate nature, as opposed to observing other organisms. Moreover, once language emerges, the representation of such causal connections will be open to unlimited logical manipulation, which will vastly enhance the ability of agents to figure out novel behavioural routes to their ends. They will also be able to formulate complex plans, perhaps facilitated by an ability to commit themselves to fixed intentions in advance.

However, I think it would be a mistake to think of these developments as eliminating older systems of behavioural control and replacing them with something quite different. In general, evolution doesn't work like that. Rather, each new development must build on pre-existing systems, adding some modification which yields some immediate selective advantage. Given this, we shouldn't expect advanced forms of means-end reasoning to direct behaviour via completely novel mechanisms. Rather, they will feed into prior systems of behavioural control, giving us new ways of adjusting the structures of behavioural dispositions that these older systems worked with.

From this perspective, we can view means-end reasoning as being built up step by step from the kind of basic cognitive architecture produced by innate structures and associationist mechanisms. Each new step provides some extra way of shaping that architecture. My conjecture is that an absolutely crucial step was the ability to acquire new behavioural dispositions directly from external observation, rather than from first-hand experience. Natural selection and associationist learning can give rise to many powerful and novel behavioural strategies, as I hope I have shown, but they do not lead easily to learning from observation. I may be wrong in my speculations about how this barrier was overcome, and it may have little to do with the distinctive features of human cognition, but I hypothesize that it was a crucial evolutionary development in any case.

Second Moral: The Prevalence of Representation. Much recent thinking about cognition presupposes a sharp dichotomy between computational (propositional, conceptual) cognition, which is presumed to allow general logical operations over sentence-like representations, and mechanistic (associationist, non-conceptual) psychology, which involves no representation and hence no inference as such. This division is upheld by a wide range of theorists, including those who differ on exactly whether they would place the divide (cf. Fodor, 2000, Sterelny, 2000). For example, it is upheld both by thinkers in the animal learning tradition, most of whom would restrict genuine representation to higher mammals, if not to humans, and also by committed computationalists, most of whom would hold that computation and representation is widespread throughout the animal kingdom.

I hope that this paper has done something to show that this sharp dichotomy is misconceived. In earlier sections I showed that we can properly attribute representational contents as soon as organisms are complex enough to have specialized drives which interact with their perceptions and behavioural dispositions. Moreover, many of the interactions between these states can properly be viewed as inferences which appropriately generate further contentful states. So, from the perspective of this paper, representation by no means requires sentence-like vehicles processed as in a digital computer, but will be present as soon as we have the kinds of dispositional architectures produced by associationist learning or analogous processes of natural selection.

At the same time, we should recognize that certain kinds of cognitive architecture, even those that sustain representation, are limited in the range of inferences that they can perform. Thus, my earlier apples-and-predators example showed how an organism can behaviourally embody the information that B will lead to O, and the information that D will lead to B, and yet not be able to infer that D will lead to O. Again, the monkeys-and-coconuts example showed how an organism might represent that B will be followed C in the form of a classical association between perceptions of B and C, and yet not be able to translate this into the practical conclusion that it should itself perform B when it wants C.

Some readers may feel inclined to respond that these inferential limitations only show that we do not yet have genuine representation, since representation by definition involves sentence-like vehicles which are open to a full range of logical manipulations. But I think that this is quite the wrong answer. If this paper has done anything, I hope it has shown how much sophisticated cognition can be performed by architectures which are very different from generalized theorem-provers. I also hope to have shown how some of the initial inferential limitations of such architectures can be overcome by adding further specialized architectures, which will no doubt leave us with further inferential limitations in turn. We will have no chance of understanding this cumulative process if we insist that there is no true representation in the absence of full inferential generality.

Third Moral: The Importance of Evolution. Finally, I hope that this paper has illustrated one uncontroversial way in which evolutionary considerations are important for the understanding of human psychology. Sceptics about

‘Evolutionary Psychology’ (with capital letters) often complain that its appeals to evolution are nothing more than ‘Just So Stories’—ungrounded speculations about historical antecedents which differ from Kipling’s fables only in not being funny. And there is some substance to this charge, given that the self-styled ‘Evolutionary Psychologists’ have a quite specific conception of the way in which evolution can illuminate psychology.

When Evolutionary Psychologists talk about the evolution of cognitive faculties in the ‘EEA’ (the ‘Environment of Evolutionary Adaptation’) they generally have in mind the differentiation of human cognition from that of other animals over the last 5 million years or so. Unfortunately, however, there are precocious little hard data by which to evaluate theories about the evolutionary pressures responsible for such differentiation. We don’t have much more than a few fossilized scraps of tooth and bone to constrain the imaginative reconstruction of stone-age scenarios which might have favoured human intelligence.

But there is a quite different way in which evolutionary considerations can illuminate human psychology. This focuses, not on the last 5 million years, but on what went before. After all, if we knew clearly how animal cognition works, then that would place immense constraints on possible theories of human psychology. Any distinctively human capacities would have to be ones that could plausibly have evolved within the last 5 million years. Equally importantly, they would have to be ones that natural selection could advantageously have added at each stage to what was already there. If only we could work out what was already there 5 million years ago, this would tell us a huge amount about the possibilities for human cognition.

This is an obvious enough point, but it is worth emphasizing. It would be a pity if justified doubts about ‘Evolution Psychology’ made us forget that we evolved fairly recently from other animals, and so stopped us using our knowledge of this to inform us about human minds.

References

Akins, C. and Zentall, T. 1998. ‘Imitation in Japanese quail: the role of reinforcement of demonstrator responding’. Psychonomic Bulletin and Review, 5, 694-7.

Avital. E. and Jablonka, E. 2000. Animal Traditions: Behavioural Inheritance in Evolution. Cambridge: Cambridge University Press.

Barkow, J., Cosmides, L. and Tooby J. 1992. The Adapted Mind. Oxford: Oxford University Press.

Carruthers, P. and Smith, P., eds. 1996. Theories of Theories of Mind. Cambridge: Cambridge University Press.

Cosmides, L. and Tooby, J. 1992. ‘The Psychological Foundations of Culture’, in Barkow, J., Cosmides, L. and Tooby J., 1992.

Davies, M. and Stone, T., eds. 1995a. Mental Simulation. Oxford: Blackwell.

Davies, M. and Stone, T., eds. 1995b. Folk Psychology. Oxford: Blackwell.

Deacon, T. 1997. The Symbolic Species. London: Allen Lane.

Dickinson, A. and Balleine, B. 2000. ‘Causal Cognition and Goal-Directed Action’, in Heyes, C. and Huber, L. eds, The Evolution of Cognition. Cambridge, Mass: MIT Press

Dickinson A. and Dawson, G. 1988. ‘Motivational Control of Instrumental Performance: The Role of Prior Experience of the Reinforcer’. Quarterly Journal of Experimental Psychology, 40B, 113-34.

Dickinson A. and Dawson, G. 1989. ‘Incentive Learning and the Motivational Control of Instrumental Performance’. Quarterly Journal of Experimental Psychology, 41B, 99-112.

Dorrance, B. and Zentall, T. 2001. ‘Imitative learning in Japanese quail depends on the motivational state of the observer at the time of observation’. Journal of Comparative Psychology, 115, 62-7.

Fodor, J. 2000. The Mind Doesn’t Work that Way. Cambridge, Mass: MIT Press

Gross, R. 1996. Psychology, Third Edition. London: Hodder and Stoughton.

Heyes, C. and Dickinson, A. 1990. ‘The Intentionality of Animal Action’. Mind and Language, 5.

Mameli, G. 2001. ‘Mindreading, Mindshaping and Evolution,’ Biology and Philosophy 16, 567-628.

Mameli, G. 2002. ‘Learning, Evolution and the Icing on the Cake,’ Biology and Philosophy 17, 141-153.

Millikan, R. 1984. Language, Thought, and other Biological Categories. Cambridge, Mass: MIT Press.

Millikan, R. 1989. 'Biosemantics'. Journal of Philosophy 86.

Papineau, D. 1984. 'Representation and Explanation'. Philosophy of Science 51.

Papineau, D. 1993. Philosophical Naturalism. Oxford: Blackwell.

Papineau, D. 1998. 'Teleosemantics and Indeterminacy', Australasian Journal of Philosophy 76, 1-14.

Papineau, D. 2001. 'The Evolution of Means-End Reasoning', in Walsh, D. ed. Naturalism Evolution and Mind, Cambridge: Cambridge University Press.

Pinker, S. 1997. How the Mind Works. London: Allen Lane

Shettleworth, S. 1998. Cognition, Evolution and Behavior. Oxford: Oxford University Press.

Sterelny, K. 2000. The Evolution of Agency and Other Essays. Cambridge: Cambridge University Press.

Tomasello, M. 2000. The Cultural Origins of Human Cognition. Cambridge, Mass: Harvard University Press.