# Affective Experience and Evidence for Animal Consciousness

**Patrick Butlin**

**Forthcoming in *Philosophical Topics*; please cite published version**

**Abstract**

Affective experience in non-human animals is of great interest for both theoretical and practical reasons. This paper highlights research by the psychologists Anthony Dickinson and Bernard Balleine which provides particularly good evidence of conscious affective experience in rats. This evidence is compelling because it implicates a sophisticated system for goal-directed action selection, and demonstrates a contrast between apparently conscious and unconscious evaluative representations with similar content. Meanwhile, the evidence provided by some well-known studies on pain in non-human animals is much less convincing. This comparison may offer lessons for the future study of animal consciousness.

## 1. Introduction

In some recent papers, Peter Godfrey-Smith (2017, 2019) has suggested that in evolutionary history consciousness may have taken two distinct original forms. These are perceptual experience, on one hand, and affective or evaluative experience, on the other. This paper assesses evidence for affective experience in non-human animals, focusing on a body of research by the psychologists Anthony Dickinson and Bernard Balleine which deserves to be better known. Affective experience is particularly significant for understanding animal welfare, and some scientists have made it the focus of their contributions to the study of animal consciousness (Cabanac et al. 2009, Denton 2006, Panksepp 2005).

Dickinson and Balleine's research concerns the role of conscious pleasure and displeasure in goal-directed action selection, a form of action selection which depends on animals' expectations about the consequences of their actions, and the values that they attribute to possible outcomes (Balleine & Dickinson 1998, Dickinson & Balleine 2000, 2009). Based on experiments on rats going back to the 1980s, Dickinson and Balleine argue for two bold claims. First, they claim that conscious experiences of pleasure and displeasure are necessary for information concerning a

range of bodily states to influence goal-directed action selection. These states include body temperature, nutrition levels, and markers of poisoning. They call this claim 'hedonic interface theory', describing hedonic experiences – conscious pleasure and displeasure – as an interface between 'motivation' and 'cognition'. In their terms, 'motivational' systems are those that access states of the body relatively directly, and use them to evaluate features of actions and outcomes. 'Cognition' is, in effect, the process of goal-directed action selection. Their view is that affective conscious experience carries information across a boundary between these two psychological domains. Second, in some papers they argue that acting as an interface of this kind is the function of consciousness more generally, and hence that the capacity for goal-directed control distinguishes conscious creatures from mere 'beast machines'.

Goal-directed action selection is a relatively sophisticated cognitive capacity, which many animals seem to lack, so Dickinson and Balleine's work represents a relatively conservative strand in research on animal consciousness. However, there is no anthropocentrism in their view: their claims are based on the varieties of instrumental learning and action selection which they have observed in animals.

Dickinson and Balleine draw these conclusions from experiments inspired by an experience of Dickinson's, which took place in Palermo in Sicily when he was a young man (described by Dickinson and Balleine 2009). One hot, dusty day Dickinson ate his first ever slice of watermelon and found it delicious. That evening he drank too much local wine and was sick. Then a few days later, hot and thirsty once more, he again sought out the watermelon stall, but was surprised to find that the fruit now tasted disgusting. Dickinson had undergone taste aversion conditioning (sometimes called the 'Garcia effect'); the novel taste of watermelon had become associated with nausea. What was particularly striking about this sequence of events was that even though the conditioning must have taken place at the time when he was unwell, Dickinson remained motivated to seek watermelon until he tasted it again.

One plausible way of thinking about what happened is that when he was sick, Dickinson's mind formed an unconscious representation of watermelon as undesirable, which caused the unpleasant quality of his later experience of tasting it. It was only through this conscious affective experience, however, that a new representation with similar content could be formed which would be accessible for goal-directed action selection. So conscious affective experience appeared to provide an interface between an associative motivational system and a cognitive system involved in rational choice. Dickinson and Balleine's experiments found that a process which is apparently very similar also takes place in rats: retasting is necessary for taste aversion

conditioning to affect goal-directed action selection. In reference to Dickinson's experience, these experiments are sometimes called the 'Palermo protocol'.

The claim I will argue for is that the results of the Palermo protocol provide particularly good evidence of affective conscious experience in non-human animals, in contrast to those of some other prominent studies. These studies include work on pain in hermit crabs (Elwood 2012) and zebrafish (Sneddon 2011, 2013) which have gained a good deal of attention among philosophers (being discussed by Godfrey-Smith 2016, Birch 2017 and Tye 2017). The reason why the latter experiments provide relatively weak evidence is that they show only that intuitively painful stimuli contribute to the selection of non-reflexive actions, whereas Dickinson and Balleine's experiments show that certain evaluative states in rats have the specific role of making information available for goal-directed action selection. I will argue that both the *making available* and the rational form of goal-directed action selection are significant.

The broader lesson of this point is not, however, that we have more reason to believe that rats have conscious experiences than that hermit crabs do (although I do think this is true). It is that by making studies on animals more elaborate in certain specific ways we can get better evidence concerning their conscious experiences. The improvements in the quality of evidence we obtain will be incremental, not revolutionary, but significant nonetheless.

The remainder of this paper has three main parts. In the next section, I contrast goal-directed action selection with other forms, and describe the Palermo-inspired experiments in detail. In section 3, I identify characteristics of these experiments that make them a good source of evidence of affective experience in non-human animals. Then in section 4, I turn to the experiments which I claim provide significantly less good evidence, and analyse what I take to be the salient differences.

## 2. The Palermo Protocol in Context

In this section, I first explain what Dickinson and Balleine mean by 'goal-directed' action selection; then describe the Palermo protocol; and finally present some other empirical evidence that provides important context.

In goal-directed action selection (also known as goal-directed control), animals use representations of the values of outcomes and of contingencies between actions and outcomes, which may be learnt independently. These representations are combined to calculate the expected reward value of possible actions, roughly in accordance with the expected utility theory of rational choice. Goal-directed control contrasts with two other processes used by non-human

animals for action selection: habitual control, in which the values of actions themselves are learnt and employed in action selection; and Pavlovian control, in which innate behavioural responses such as approach, consumption and withdrawal are expressed, and may come to be elicited by new stimuli as a result of learning. In computational terms, the goal-directed and habitual forms of action selection are thought to implement particular forms of model-based and model-free reinforcement learning, respectively (Daw et al. 2005, Daw & O'Doherty 2013).

It is noteworthy that there are some behaviours that are naturally described as goal-directed, but which may not fit this definition. For example, imagine a foraging animal that detects the odour of a nutritious root, searches for the place where the odour is strongest, then digs up and eats the root. In this process the animal's behaviour may have been robustly oriented towards the food, its goal, and may be explained by the value of this goal. But in the technical sense in which I am using the term 'goal-directed', we cannot know whether the animal's behaviour was goal-directed on the basis of this observation alone. It is also possible that the behaviour was controlled solely by Pavlovian means, in which case the animal's approach and digging behaviour would be innate, and it would be unable to learn to refrain from performing it, even in an experiment in which refraining was rewarded and digging was not. True goal-directed behaviour, in contrast, is sensitive to the relationship between the action and the outcome (Daw & O'Doherty 2013). Because of the possibility of behaviours like this, the term 'goal-directed' is sometimes used in the animal consciousness literature in a less demanding sense than it is here.

Goal-directed control relies on representations of two kinds, which can be updated independently, so two corresponding experimental paradigms are particularly significant for distinguishing this form of control from others. One is *contingency degradation* (Hammond 1980, Dickinson et al. 1998), which tests whether an animal's performance of an action depends on its knowledge of the consequences. This might be done by making it the case that delivery of food no longer depends on an action's being performed, or requires the performance of a different action from previously. Changes in behaviour under such conditions are evidence of goal-directed control.

For our purposes, however, the more important paradigm is *outcome devaluation* (Adams & Dickinson 1981), in which the value that the experimental animal places on an outcome is manipulated.[1] This manipulation takes place away from the environment in which the action concerned can be performed, in order to ensure that it is the value placed on the *outcome* which is affected, not the value placed on the *action* (which would also affect habitual control). A typical

[1] Bowers (2016) argues that devaluation experiments could be useful in assessing whether fish feel conscious pain.

outcome devaluation experiment has three stages. In the first stage, animals are given the opportunity to perform an action, such as pressing a lever, to obtain a novel and rewarding food. In the second stage, the animal is offered the same food in a different environment, and is then made ill by an injection of lithium chloride (LiCl). Then in the third stage, the animal is returned to the environment where the action is available, and observed to see whether it is performed. In this stage the food reward is not delivered – the action is performed 'in extinction'.

In outcome devaluation experiments, a reduction in responding after devaluation (in comparison to controls) is taken to be evidence of goal-directed control. This is because test animals differ from controls only in that they have been exposed to a manipulation that would naturally lead them to place a lower value on the food reward. If such a manipulation affects their behaviour, they must employ representations not only of this value, but also of the causal relationship as they know it between the action and the delivery of this specific food. That is, they must represent both outcome values and action-outcome contingencies. Animals that have been overtrained (i.e. trained for too long in the first stage) do not reduce responding after devaluation, indicating that their actions have come under habitual control (Adams 1981), and similar results have also been observed in humans (Tricomi et al. 2009).

The experiments inspired by Dickinson's experience in Palermo focus on the second stage of outcome devaluation. In one study, Balleine and Dickinson (1991) found that rats that had been through the second stage as described above – that is, they had been given the food immediately followed by a lithium chloride injection – would not reduce responding compared to controls, unless they had also been given a further opportunity to consume the food, between the second and third stages. They also tested the importance of retasting in a different design, in which rats were trained to perform two actions for different foods, then immediately given an injection of LiCl. The rats were allowed to retaste one of the two foods, and then found to reduce performance, in extinction, of whichever action was associated with the retasted food. These experiments both indicate that the outcome value representations used in goal-directed control are updated by taste aversion conditioning only after retasting. In a further study, Balleine and colleagues (1995) found that when retasting took place under the influence of an anti-nausea drug the outcome devaluation effect was attenuated, further confirming the previous results, and suggesting that retasting causes devaluation by triggering an experience of nausea.

These experiments suggest that rats undergo an unpleasant conscious experience when they retaste foods that have previously been paired with LiCl injections, and this causes them to assign a new value to these foods for the purposes of goal-directed control. To put the point intuitively, the experiments show that rats have different behavioural dispositions before and

after retasting, and the natural explanation is that the food *tasted bad*. I will explain in detail why I take these experiments to provide particularly good evidence of conscious affective experience in non-human animals in the next section. Before moving on, however, I will comment on four further aspects of the relevant empirical evidence.

First, there is evidence that revaluation in some other circumstances also requires evaluatively-valenced conscious experience. This evidence comes from studies of the effects of hunger, satiety and thirst (which scientists working in this area call 'motivational states') on goal-directed action. One would naturally expect hungrier animals to be more inclined to perform actions that they expect to lead to food delivery, and less hungry animals to be less inclined to do so, without needing any particular training. Remarkably, however, hunger and similar states affect goal-directed action only when the animal has undergone *incentive learning*, in which it gains experience of consuming the specific food concerned in those states (Dickinson & Dawson 1988, Balleine 1992, Niv et al. 2006). Thus, for instance, if a rat has learnt to perform an action to receive a food when hungry, and later has the opportunity to perform that action when sated, it will not reduce responding unless it has had the opportunity to consume the food in this state. Rats have to learn that specific foods are less good when they are sated, in order for satiety to make them less motivated to pursue those foods. Conversely, hungrier rats do not increase responding compared to controls unless they have learnt that the specific foods involved are better when they are hungry. These results may be explained in the same way as those of the Palermo protocol: hunger, satiety and thirst do not directly influence goal-directed behaviour, but instead dispose animals to have more or less pleasant conscious experiences when they consume food or water. These experiences in turn cause outcome value representations to become contingent on physiological states. Similarly, taste aversion conditioning disposes animals to have less pleasant conscious experiences when they consume foods, which cause outcome devaluation. So studies of incentive learning provide further support for Dickinson and Balleine's interpretation of the Palermo experiments, because they identify other effects that can be explained by a similar process.

Second, taste aversion conditioning has been demonstrated in many species (Lin et al. 2017), but many of these demonstrations took a relatively simple form which does not provide evidence of either goal-directed control or a requirement for retasting, so does not constitute evidence for consciousness of the kind at issue here. To take just two examples, slugs (*Limax maximus*; Gelperin 1975) and goldfish (Martin et al. 2011) have both been shown to significantly reduce consumption of foods with specific flavours which have been paired with noxious chemical stimuli. But in these experiments, instead of testing the performance of animals on an

instrumental action which had previously led to food delivery, the experimenters simply provided the foods, and allowed the animals to consume them. Behaviour could therefore have been controlled by Pavlovian action selection, and no test of the importance of retasting was possible, because the behaviour tested itself involved consumption. Meanwhile, there is surprisingly little conclusive evidence (that I am aware of) concerning the range of species that is capable of goal-directed action selection. In birds, a study on western scrub jays by Clayton and Dickinson (1999) showed sensitivity to outcome devaluation, using specific satiety, in searching for previously cached food; and there is some inconclusive evidence suggesting that pigeons are sensitive to outcome values in performing behaviours learnt by imitation (Saggerson et al. 2005, McGregor et al. 2006). These points give us reason to believe that the capacity may exist in some bird species. Perhaps the most relevant study on fish used a devaluation procedure to test whether rainbow trout were capable of learning a stimulus-outcome association between a green light and the delivery, a few seconds after the light had gone off, of food pellets (Nordgreen et al. 2010). This study examined behaviour that was likely to be the product of Pavlovian control, however, and I am not aware of experiments on fish which have attempted to show specifically goal-directed control.

Third, although both taste aversion conditioning (Bernstein & Webster 1980, Klosterhalfen et al. 2000) and outcome devaluation (Tricomi et al. 2009) have been demonstrated in humans, to my knowledge these paradigms have not been combined in such a way as to demonstrate that humans, like rats, require retasting for taste aversion conditioning to lead to outcome devaluation. In taste aversion conditioning studies, participants are given the opportunity to retaste foods which have earlier been paired with illness, and are found to consume less of them; but their willingness to consume the foods prior to retasting has not been tested. In outcome devaluation studies, devaluation of foods for human participants has been achieved by encouraging consumption of those foods until specific satiety is reached, rather than by administering drugs or other procedures which cause illness, and reaching specific satiety inevitably involves tasting foods after they have been largely devalued. So as far as I have been able to tell, no studies that perfectly parallel the Palermo protocol have been conducted with human subjects.

Fourth, and finally, there is evidence that retasting is not necessary for taste aversion conditioning, and incentive learning is not necessary for hunger and thirst to affect behaviour, in cases where action is Pavlovian or habitual. In particular, Best and colleagues (1989) found that rats would less readily approach a box where food had been delivered after that food had been devalued with LiCl, without retasting. Approach behaviour is thought to be under Pavlovian

control, so this result suggests that taste aversion conditioning has a direct effect on Pavlovian action selection, with no need for an unpleasant conscious experience of consuming the food concerned. Experiments on sensitivity to hunger and thirst have shown a similarly direct influence on both Pavlovian and habitual control (Balleine 2000, Niv et al. 2006). The significance of these points is that, if the Palermo protocol is taken to provide good evidence of conscious experience, the connection thus drawn is specifically between consciousness and goal-directed control.

**3. The Experiments as Evidence for Consciousness**

I now turn to philosophical analysis of the Palermo protocol: what is it about these experiments, exactly, that makes them compelling evidence of affective conscious experience? I will discuss three features, in increasing order of importance. First, though, I will give some more detail about the hypothesis that I take the experiments to support.

The experiments provide evidence that when rats retaste foods that have been previously paired with LiCl injections, this causes them to learn to place a lower value on those foods in future. At the point of retasting, the rats therefore appear to undergo states which both represent the food's identity – or perhaps some taste, texture and odour properties – and have further features which cause devaluation. Exactly what these further features are is uncertain, but they may include any or all of: representation of the onset of nausea; representation of negative or lower-than-expected value; or, if the states are conscious, a quality of felt unpleasantness. My claim is that the experiments provide evidence that these states are conscious, and I will continue to refer to them as *affective* states, but for present purposes they could equally well be described as 'evaluative' or 'hedonic'. The point is that they have a quality that carries immediate implications for subjective value. I do not claim that the experiments provide evidence that any other states involved in the processes of taste aversion conditioning or goal-directed action selection are conscious, although some may be. I should also clarify that by 'conscious' I mean *phenomenally* conscious (Block 1995), so the suggestion is that rats undergo phenomenally conscious affective experiences.

The reason that the Palermo-inspired experiments provide good evidence of consciousness is that the affective states posited to explain their results are available for use in goal-directed action selection in a way that contrasts with other states that rats use for storing evaluative information. In these respects, their function role is similar to the functional role which characterises conscious states in humans. This latter role has not yet been fully or conclusively identified, but

we do understand it sufficiently to be able to make some assessment of the quality of evidence that the present experiments provide. In particular, the most vigorous point of debate for some years has been whether consciousness requires that content is represented in a global workspace, and thus made accessible for a wide variety of uses, including rational action selection. This debate pits workspace theorists (such as Dehaene & Naccache 2001 and Naccache 2018) against opponents who argue for a more liberal view of the functional role of consciousness, which includes certain forms of perceptual-format representation outside the workspace (e.g. Block 2007, Lamme 2010). But there is very little debate about the *sufficiency* of accessibility via a global workspace for phenomenal consciousness, so a state's having a functional role in a non-human animal akin to presence in the human workspace is good evidence that it is conscious. My argument therefore works by analogy: certain affective states in rats are likely to be conscious because they are similar in important respects to conscious states in humans. One reason why this form of argument is justified is that in broader respects rats and humans are very similar things indeed; they are animals with comparatively recent common ancestors (estimated at 87 million years ago; Springer et al. 2003).

More specifically, the affective experiences posited to explain the Palermo results have three significant features. These are: that they contribute to action *selection*, as opposed to action guidance more generally; that the species of action selection to which they contribute is *goal-directed control*, a distinctively rational form; and that they contribute by *making available* information which was already stored in the rats' minds. I will discuss these three features in turn.

The first feature is that the affective experiences contribute to action selection, as opposed to some other aspect of action guidance, such as motor control. That is, they contribute to the rats' choices to perform or refrain from performing certain actions, as opposed to making a difference to how these actions are performed. This distinction matters because, as Bayne (2013) explains, there is compelling evidence that motor control and the manner of performance of intentional action can be influenced by unconscious states in humans. For example, studies of patients with impairments of the ventral visual processing stream suggest that they can use information processed by the dorsal stream to guide action, even though this information is not presented to them in conscious experience (Carey et al. 1995), but the specific form of guidance implicated in these experiments is motor control for action execution, not action selection. In contrast, Milner and Goodale (2008) argue that planning and action selection relies on representations in the conscious ventral stream.

The second feature, which raises more vexed and complex issues, is that the affective experiences in question contribute to goal-directed action selection, and that this form of action selection is distinctively rational. I will first argue that the involvement of goal-directed action selection in the Palermo experiments, as opposed to another form, allows them to provide more powerful evidence for consciousness; and that goal-directed control has important characteristics of rational choice. Then I will turn to the trickiest point in this area, which is the relationship between rationality and consciousness.

One reason why the involvement of goal-directed control is significant is that there is evidence from a very different source that rats undergo conscious experiences when using this system. As Redish (2016) describes, rats running mazes are often observed to stop at junctions and look in turn towards each of the onward paths, before moving on. By recording activity in the hippocampus, Redish and colleagues have found that place cells associated with different locations in the maze fire sequentially during these pauses, apparently tracing out possible future paths. This activity happens serially, so it is naturally interpreted as the neural signature of conscious, imaginative prospection. It is also particularly associated with goal-directed control; the behaviour decreases as stable environments become increasingly familiar, allowing control to be handed over to the habit system, and increases when reward contingencies change. Furthermore, Redish argues in detail that this process should be thought of as one of deliberation, and if this interpretation is correct, it entails that goal-directed control features one of the hallmarks of rational choice.

In addition to this, goal-directed action selection has other characteristics of rational choice. Most fundamentally, it relies on expectations about the specific consequences of possible actions. Animals using goal-directed control can be said to be responsive to reasons, at least in the sense that their actions are governed by representations of facts that are apt to rationalise them. For example, that pressing a lever will lead to the delivery of peanuts is undoubtedly a reason to perform this action for an animal that desires peanuts. Whether this means that animals using goal-directed action selection *act for reasons* is a more difficult question, because theories of action for reasons vary considerably. On Dretske's (1988) account, for instance, goal-directed control would suffice, but more recent accounts, especially in moral psychology, have tended to be significantly more demanding (see Schlosser 2012 for references and a typical example). However, these more demanding theories impose extra requirements on top of what I take to the basic element in action for reasons, which that the agent's behaviour is caused by representations of facts which are apt to rationalise the actions thus produced, via the operation

of a mechanism which is sensitive to such rationalising relationships. My claim is just that this basic element of rationality is in place.

The goal-directed system is also highly general, in the sense that in principle it can learn to select for or against the performance of any action that the animal has the strength and co-ordination to perform – it is not as though lever-pressing is likely to have any particular ecological significance for rats, and rats can also learn to omit lever-presses when the relationship between this action and food delivery changes from positive to negative (Dickinson et al. 1998). And a further respect in which the system is rational is that when new information is made available in the right format, the system is immediately responsive to this information, even if it was learnt outside the context in which the action is performed – this is shown by outcome devaluation experiments.

The fact that the Palermo experiments implicate goal-directed action selection is also significant because even in humans, actions under habitual control may be selected and initiated unconsciously (Evans & Stanovich 2013, Wood & Rünger 2016), although this process may subject to some form of conscious monitoring, and habitual actions typically generate conscious experiences of various kinds as they are performed. If humans can select and execute habitual actions unconsciously, it is very likely that non-human animals do too, and consequently the fact that they employ a further, more sophisticated system for action selection makes an important contribution to the case for consciousness in rats.

The claim that exhibiting the capacity for rational agency shows that an animal is conscious is a controversial one. On one side, Bayne (2013) defends the closely related view that *intentional* agency is a legitimate marker of consciousness, and the attractions of the claim must explain why Block (2002) defined access consciousness as requiring that representations are poised for use in rational control of action. But opposing this, Seth (2009) argues that consciousness is 'neither necessary nor sufficient for rational action'. Seth appeals primarily to results from social psychology, which have also led some authors to the view that conscious experiences associated with agency are a mere interpretation of unconscious processes which actually control action (e.g. Wegner 2002, Bargh 2005). On this view, even though consciousness is associated with rational choice in humans, it might be thought that rational agency could be achieved without consciousness in other animals.

Two factors make it more difficult in general to infer consciousness from rational behaviour, but neither of these tells strongly against the present case. The first factor is that rationality and consciousness are conceptually entangled. Smithies (2011) argues that rationality requires consciousness on the grounds that rational action must be motivated by introspectively-

accessible considerations and attributable to a conscious, reasoning subject. Clark and Kiverstein (2007), following Evans (1982), suggest the converse view: that consciousness requires rationality, because only states that are accessible to reasoning subjects can be conscious. Both claims are made on conceptual grounds, and Smithies' view in particular presents an obstacle, because it implies that rationality cannot be identified empirically any more easily than consciousness itself. However, although it might be a mistake to rely on Smithies' claim in an argument from empirical observations of apparently rational behaviour to the conclusion that an animal is conscious, it does not prevent us from considering the evidential value of aspects of rationality which are less thoroughly entangled with consciousness. My suggestion is that because goal-directed action selection is prospective, deliberative, general and responsive its presence in rats provides evidence of the capacity for conscious experience.

The second factor is the evidence from social psychology which Seth, Wegner, Bargh and others use to argue for the possibility of unconscious rational agency. Unfortunately, some of the results that these authors rely on have not been reproduced in recent replication attempts. For example, Seth (2009) appeals to results purporting to show that participants performed better in complex decision tasks, such as choosing between different models of car, when prevented from reasoning about them consciously (Dijksterhuis 2004, Dijksterhuis et al. 2006). A large-scale replication attempt and meta-analysis found that there was no such advantage to unconscious thought (Nieuwenstein et al. 2015).

However, one strand of social psychology which is of particular relevance here is research on unconscious goal priming (Aarts & Custers 2012). Studies in this area suggest that participants' motivation to pursue goals can be influenced without their knowledge by techniques such as the use of goal-associated words in prior tasks. Some of the behaviour thus prompted involves practised routines, suggesting that goal-priming works by triggering the habit system, but Aarts and Custers also argue that flexible behaviour may be employed in the pursuit of unconscious goals. It might be suggested, therefore, that human goals can be acquired and changed unconsciously, and hence that something similar could be taking place in outcome devaluation in rats. The reason this argument is unsuccessful is that the analogy between goal priming and outcome devaluation is really quite weak. In the Palermo protocol, the valence of the value placed on an outcome is reversed from positive to negative, through an experience of interaction with that very outcome. In contrast, experiments on goal priming typically seem to involve the activation of representations of ends or forms of behaviour which are already taken to be worthwhile, and there is evidence that such automatic triggering of goals relies on congruence with conscious desires and intentions (Sheeran et al. 2005). Furthermore, Bayne (2013) points

out that in some studies, the interpretation of primes as influencing which goals the participants pursue is doubtful – instead, what seems to be influenced is the manner of pursuit. For example, one well-known study found that participants primed with words related to achievement and striving persisted longer and performed better on puzzles than those exposed to neutral primes (Bargh et al. 2001), and this result can arguably be interpreted either way. So although there may well be many unconscious influences on goal-directed action, this line of evidence does little to undermine the evidence for consciousness from the Palermo results.

Finally, the third feature which makes the Palermo protocol particularly good evidence of consciousness is that the hypothesised affective experiences *make information available* to goal-directed control, which was already represented in the rats' minds. More specifically, perception of the foodstuff involved makes evaluative information about it available. So the Palermo-inspired experiments show us a contrast between two ways in which information can be represented in rats' minds, of which only one is accessible to rational action selection, with representations of this privileged form being generated by perception, but not by the form of inference involved in taste aversion conditioning itself. This is a particularly important point, as I will argue further in the next section: the Palermo protocol shows us a difference in status between two representations with similar content, and we have grounds to assimilate this to the conscious/unconscious distinction, since it determines the availability of this content to rational action selection.

It is also important, however, that the apparently conscious representation is supported by a perceptual state, because in humans perceptual imagination plays a particularly important role in evaluating potential goals. When faced with difficult decisions, humans often imagine alternative possible futures, in order to learn more about their own preferences (Gilbert & Wilson 2007, Nanay 2016). This process brings stored evaluative information to consciousness. This is particularly noteworthy because Seth (2009) argues that there is a growing consensus that the function of consciousness is integration: making information from one part of the mind available to others. This integration is made possible by the global workspace, and part of the function of the global workspace is to facilitate mental simulation and internal evaluation (Dehaene & Naccache 2001, Hesslow 2002, Revonsuo 2005; all cited in Seth 2009). If making stored or implicit evaluative information available through mental simulation in a perceptual format is a paradigmatically conscious process, then making such information available through perception itself is also likely to involve conscious experience.

The three features of the Palermo protocol which I have discussed in this section can be seen as providing evidence for affective consciousness through either of two routes.[2] First, one might take the view that since rational choice seems to draw on conscious mental states in humans, the presence of goal-directed action selection in rats is evidence that they too are conscious; then building on this, that the point about retasting is evidence that affective experiences are among the conscious inputs to the process. On this view the analogy between goal-directed control and rational choice would be crucial, with the Palermo protocol providing only an interesting extra detail. Second, however, an alternative perspective is that the Palermo protocol matters because it shows that some information must be represented in a special, accessible form in order to influence goal-directed action selection, and this adds weight to the case for associating goal-directedness with consciousness in the first place. Dickinson and Balleine (2009) seem to favour this second perspective, and I agree that it offers a fuller appreciation of the significance of their results.

## 4. Less Compelling Evidence

So far I have described the Palermo protocol and argued that its results provide particularly good evidence for affective conscious experience in non-human animals. In this section I add further substance to this claim by contrasting the Palermo protocol with three other studies, each of which has been taken to provide important evidence that the species involved are capable of experiencing conscious pain.

The three experiments I will discuss are listed by Godfrey-Smith (2016) as evidence of conscious pain in non-human animals, because they show responses to damage which are 'more than reflexes'. They are as follows:

*i. Sneddon on zebrafish:* Sneddon (2011, 2013) describes an experiment in which zebrafish were placed in an environment allowing access to two chambers. One of these chambers was 'enriched' with gravel, a plant, and a view of other zebrafish in a neighbouring tank, while the other was bare. In control conditions, the fish spent almost all of their time in the enriched chamber, but when they were injected with a noxious chemical, and an analgesic was dissolved in the water of the bare chamber, they spent more time in that bare chamber.

*ii. Danbury et al. on chickens:* Danbury et al. (2000) studied chickens which had been trained to discriminate between different coloured feeds, one of which contained an analgesic, and found that lame chickens consumed significantly more of the drugged feed than uninjured birds.

[2] Thanks to an anonymous referee for encouraging me to clarify this point.

*iii. Elwood on hermit crabs:* Elwood (2012) describes a number of experiments on hermit crabs which were given electric shocks while in their shells. These experiments found that crabs occupying lower-quality shells would leave them in response to lower-voltage shocks, compared to those in higher-quality shells, and that when crabs were in similar shells, the voltage required to induce them to leave was affected by an odour in the water indicating the presence of a predator.

These experiments are taken to provide relatively strong evidence of pain – that is, of conscious affective experience – for different reasons. All three show non-reflex behaviour which is affected by, or in response to, injuries or noxious stimuli. The studies on zebrafish and chickens show preferences in favour of analgesics, which is taken to be evidence that the animals are in aversive motivational states which are ameliorated by these drugs. The studies on zebrafish and hermit crabs both show motivational trade-offs: in the zebrafish study the benefit of analgesia outweighs that of the otherwise-preferred environment, and in the hermit crab experiments the intensity of a noxious stimulus seems to be weighed against the costs of leaving a shell. Ginsburg and Jablonka (2019) might argue that the study on chickens is the most compelling, because it involves what they call 'unlimited associative learning', while the other two studies do not require learning of any form.

However, these three studies lack the latter two of the three features that make the evidence from the Palermo protocol compelling. That is, they do not implicate goal-directed action selection, or any comparably rational process, and they also do not demonstrate a contrast between accessible and inaccessible forms in which evaluative information can be represented. This is not intended as a criticism of the studies; their methods may be among the best ways available to elicit evidence of affective experience in the animals involved, especially if they are not capable of goal-directed action selection. But to think about how willing we should be to infer from these studies that the animals experience conscious pain, it makes sense to compare the evidence they provide to that which is available from different studies, on animals with different cognitive capacities.

Returning to the two absent features, the behaviours exhibited in the three experiments I am focusing on can all be explained as products of the Pavlovian action selection system. The zebrafish experiment involves moving from one place to another; the chicken experiment concerns feeding; and the hermit crab study involves abandoning shells. All three of these behaviours are plausibly innate for the species involved, and there is nothing about the details of the experiments that indicates that the goal-directed system would be required. In particular, motivational trade-offs are compatible with Pavlovian control, because some Pavlovian actions

are not mere reflexes, but are sensitive to the values associated with different stimuli. In the hermit crab study, for instance, which gives us the clearest example of a trade-off, all that is required is the use a system which can weigh the strength of two or more competing action tendencies, each of which is a consequence of directly-perceptible features of the immediate environment.

The fact that learning is required in the chicken experiment does entail a step up in the sophistication of the behavioural control required, but a natural Pavlovian explanation of the observed behaviour is that the birds learnt a stimulus-outcome association between feed colour and reduced sensation of bodily injury, which led to a preference for the drugged feed. This does not require either that the chickens grasped the instrumental relationship between eating and reduced sensation, or that this sensation was conscious. On the former point, animals exhibit behaviours such as approach and pecking towards stimuli which are associated with rewards even when these actions are consistently counterproductive (Dayan et al. 2006), so since this study involves only feeding, it does not demonstrate instrumental reasoning. On the latter, Danbury et al. note that chickens have also been shown to be able to self-select diets which provide sufficient protein and ascorbic acid, and presumably we should not assume that to be capable of this they must have conscious experiences of deficits in these nutrients.

Turning now to the second feature, what is noteworthy about the three experiments that I am considering in this section is that none of them show a contrast between inaccessible and accessible forms of representation, such as that seen in the Palermo protocol. In the zebrafish and hermit crab experiments animals respond in the moment to signals concerning their environments or internal states: the zebrafish move from the tank in which they receive a stronger nociceptive signal, and remain in the tank in which this is weaker; and the hermit crabs leaves their shells when they detect a strong shock, and remain when they detect only a weaker one. In the chicken study we know that a change in the strength of a signal of bodily injury contributes to a learning process, but this is not contrasted with a representation which fails to do so. These points are important because the Palermo protocol prompts us to explain what effect retasting had on the way in which evaluative information about the food concerned was represented in the rats' minds, which allowed it to contribute to subsequent choice, and a natural explanation is that retasting rendered this content conscious. But in the present experiments no such explanation is called for.

It may be objected that zebrafish, hermit crabs and chickens almost certainly do store information in ways that make it inaccessible to many of their cognitive processes, and that much or all of this storage is likely to be unconscious, so the contrast between conscious and

unconscious representation is also demonstrated in these animals, insofar as the three experiments provide evidence for consciousness. The thought would be that evidence of unconscious representation is easy to come by, so the contrast brought out in the Palermo experiments does not contribute to their evidentiary value. However, this objection is mistaken. If the three experiments each provides good evidence for consciousness, and we know as a general principle that zebrafish, chickens and hermit crabs also have unconscious information-storing states, then we can infer a contrast between conscious and unconscious states in these animals. That is, insofar as we have evidence for consciousness, this is also evidence for a contrast; if there is no consciousness, there is no contrast. But this is the opposite of the situation that the Palermo protocol puts us in, which is that we have evidence for a contrast between two kinds of information-bearing states, and this contributes to our evidence for consciousness. So the contrast brought out in the Palermo protocol does give us a further reason to believe that rats have conscious experiences, and the lack of a reason of this form makes the evidence from the other three studies less compelling.

**4. Conclusion**

I have argued that the Palermo protocol provides better evidence for conscious affective experience in non-human animals than the experiments by Sneddon, Elwood, and Danbury et al.. The latter experiments have been noted more often by philosophers in recent years, but this may be because they are themselves more recent, or because the current tendency among philosophers is to think that consciousness in other mammals is unremarkable. The more general lesson that I would like to emphasise, however, is that in investigating non-human consciousness we will benefit from careful attention to the range of mechanisms and processes for action selection that animals display. By deploying the distinction between Pavlovian, habitual and goal-directed forms of action selection, and being sensitive to the specific roles that particular states play in these systems, we can develop a more nuanced understanding of the functional role of conscious states in humans. We can then use this understanding to make better-informed judgments about the significance of particular behaviours in non-human animals.

**References**


Aarts, H. & R. Custers. 2012. Unconscious goal pursuit: nonconscious goal regulation and motivation. In Ryan, ed., *The Oxford Handbook of Human Motivation*.

Adams, C. D. 1981. Variations in the sensitivity of instrumental responding to reinforcer devaluation. *Quarterly Journal of Experimental Psychology* 34B: 77-98.

Adams, C. D. & A. Dickinson. 1981. Instrumental responding following reinforcer devaluation. *Quarterly Journal of Experimental Psychology* 33B: 109-122.

Balleine, B. 1992. Instrumental performance following a shift in primary motivation depends on incentive learning. *Journal of Experimental Psychology: Animal Behaviour Process* 18: 236-250.

Balleine, B. 2000. Incentive processes in instrumental conditioning. In Mowrer & Klein, eds., *Handbook of Contemporary Learning Theories*.

Balleine, B. & A. Dickinson. 1991. Instrumental performance following reinforcer devaluation depends upon incentive learning. *Quarterly Journal of Experimental Psychology* 43B: 211-231.

Balleine, B. & A. Dickinson. 1998. Consciousness – the interface between affect and cognition. In Cornwell, ed., *Consciousness and Human Identity*.

Balleine, B., C. Garner & A. Dickinson. 1995. Instrumental outcome devaluation is attenuated by the anti-emetic ondansetron. *Quarterly Journal of Experimental Psychology* 48B: 235-251.

Bargh, J. 2005. Bypassing the will: toward demystifying the nonconscious control of social behavior. In Hassin, Uleman & Bargh, eds., *The New Unconscious*.

Bargh, J., P. M. Gollwitzer, A. Y. Lee-Chai, K. Barndollar & R. Trötschel. 2001. The automated will: nonconscious activation and pursuit of behavioral goals. *Journal of Personality and Social Psychology* 81: 1014-1027.

Bayne, T. 2013. Agency as a marker of consciousness. In Clark, Kiverstein & Vierkant, eds., *Decomposing the Will*.

Bernstein, I. L. & M. M. Webster. 1980. Learned taste aversions in humans. *Physiology and Behavior* 25: 363-366.

Best, M. R., S. F. Davis & C. A. Grover. 1989. Straight alley extinction performance is disrupted by taste-aversion conditioning. *Learning and Motivation* 20: 385-372.

Birch, J. 2017. Animal sentience and the precautionary principle. *Animal Sentience* 16 (1).

Block, N. 1995. On a confusion about a function of consciousness. *Behavioral and Brain Sciences* 18: 227-287.

Block, N. 2002. Concepts of consciousness. In Chalmers, ed., *Philosophy of Mind: Classical and Contemporary Readings*.

Block, N. 2007. Consciousness, accessibility and the mesh between psychology and neuroscience. *Behavioral and Brain Sciences* 30: 481-548.

Bowers, R. 2016. Devaluation as a strategy to assess empirically whether fish feel. *Animal Sentience* 3 (43).

Cabanac, M., A. J. Cabanac & A. Parent. 2009. The emergence of consciousness in phylogeny. *Behavioural Brain Research* 198: 267-272.

Carey, D. P., M. Harvey & A. D. Milner. 2005. Visuomotor sensitivity for shape and orientation in a patient with visual form agnosia. *Neuropsychologia* 34: 329-338.

Clark, A. & J. Kiverstein. 2007. Experience and agency: slipping the mesh. *Behavioral and Brain Sciences* 30: 502-503.

Clayton, N. & A. Dickinson. 1999. Memory for the contents of caches by Scrub Jays. *Journal of Experimental Psychology: Animal Behaviour Processes* 25: 82-91.

Danbury, T., C. Weeks, A. Waterman-Pearson, S. C. Kestin & J. P. Chambers. 2000. Self-selection of the analgesic drug carprofen by lame broiler chickens. *Veterinary Record* 146: 307-311.

Daw, N., Niv, Y. & P. Dayan. Uncertainty-based competition between prefrontal and dorsolateral striatal system for behavioral control. *Nature Neuroscience* 8(12): 1704-1711.

Daw, N. & J. O'Doherty. 2013. 'Multiple systems for value learning'. In Glimcher & Fehr, eds., *Neuroeconomics: Decision-Making and the Brain* (2nd edition).

Dayan, P., Y. Niv, B. Seymour & N. D. Daw. 2006. The misbehavior of value and the discipline of the will. *Neural Networks* 19: 1153-1160.

Dehaene, S. & L. Naccache. 2001. Towards a cognitive neuroscience of conscious: basic evidence and a workspace framework. *Cognition* 79: 1-37.

Denton, D. 2006. *The Primordial Emotions: The Dawning of Consciousness*.

Dickinson, A. & B. Balleine. 2000. Causal cognition and goal-directed action. In Heyes & Huber, eds., *The Evolution of Cognition*.

Dickinson, A. & B. Balleine. 2009. Hedonics: The cognitive-motivational interface. In Kringelbach & Berridge, eds., *Pleasures of the Brain*.

Dickinson, A. & G. Dawson. 1988. Motivational control of instrumental performance: the role of prior experience of the reinforcer. *Quarterly Journal of Experimental Psychology* 40B: 113-34.

Dickinson, A., S. Squire, Z. Varga & J. W. Smith. 1998. Omission learning after instrumental pretraining. *Quarterly Journal of Experimental Psychology* 51B: 271-286.

Dijksterhuis, A. 2004. Think different: the merits of unconscious thought in preference development and decision making. *Journal of Personality and Social Psychology* 87: 586-598.

Dijksterhuis, A., M. W. Bos, L. F. Nordgren & R. B. van Baaren. 2006. On making the right choice: the deliberation-without-attention effect. *Science* 311: 1005-7.

Dretske, F. 1988. *Explaining Behavior: Reasons in a World of Causes*.

Elwood, R. W. 2012. Evidence for pain in decapod crustaceans. *Animal Welfare* 21: 23-27.

Evans, J. & K. E. Stanovich. 2013. Dual-process theories of higher cognition: advancing the debate. *Perspectives on Psychological Science* 8: 223-241.

Evans, G. 1982. *The Varieties of Reference*.

Gelperin, A. 1975. Rapid food-aversion learning by a terrestrial mollusk. *Science* 189 (4202): 567-570.

Gilbert, D. T. & T. D. Wilson. 2007. Prospection: experiencing the future. *Science* 317: 1351-1354.

Ginsburg, S. & E. Jablonka. 2019. *The Evolution of the Sensitive Soul: Learning and the Origins of Consciousness*.

Godfrey-Smith, P. 2016. Animal evolution and the origins of experience. In Livingstone Smith, ed., *How Biology Shapes Philosophy*.

Godfrey-Smith, P. 2017. The evolution of consciousness in phylogenetic context. In Andrews & Beck, eds., *The Routledge Handbook of Animal Minds*.

Godfrey-Smith, P. 2019. Evolving across the explanatory gap. *Philosophy, Theory and Practice in Biology* 11:1.

Hammond, L. J. 1980. The effect of contingency upon the appetitive conditioning of free-operant behavior. *Journal of the Experimental Analysis of Behavior* 34(3): 297-304.

Hesslow, G. 2002. Conscious thought as simulation of behaviour and perception. *Trends in Cognitive Science* 6: 242-247.

Klosterhalfen, S., Rüttgers, A., Krumrey, E., Otto, B., Stockhorst, U., Riepl, R. L., … Enck, P. 2000. Pavlovian conditioning of taste aversion using a motion sickness paradigm. *Psychosomatic Medicine.* 62: 671–677.

Lamme, V. 2010. How neuroscience will change our view on consciousness. *Cognitive Neuroscience* 1: 204-220.

Lin, J.-Y., J. Arthurs & S. Reilly. 2017. Conditioned taste aversions: from poisons to pain to drugs of abuse. *Psychonomic Bulletin and Review* 24: 335-351.

Martin, I., A. Gomez, C. Salas, A. Puerto & F. Rodriguez. 2011. Dorsomedial pallium lesions impair taste aversion learning in goldfish. *Neurobiology of Learning and Memory* 96: 297-305.

McGregor, A., A. Saggerson, J. Pearce & C. Heyes. 2006. Blind imitation in pigeons, *Columba livia*. *Animal Behaviour* 72: 287-296.

Milner, A. D. & M. A. Goodale. 2008. Two visual systems reviewed. *Neuropsychologia* 46: 774-785.

Naccache, L. 2018. Why and how access consciousness can account for phenomenal consciousness. *Philosophical Transactions of the Royal Society B: Biological Sciences* 373: 20170357.

Nordgreen, J., A. M. Janczak, A. L. Hovland, B. Ranheim & T. E. Horsberg. Trace classical conditioning in rainbow trout (*Onchorhynchus mykiss*): what do they learn? *Animal Cognition* 13: 303-309.

Panksepp, J. 2005. Affective consciousness: core emotional feelings in animals and humans. *Consciousness and Cognition* 14: 30-80.

Redish, A. 2016. Vicarious trial and error. *Nature Reviews Neuroscience* 17: 147-159.

Revonsuo, A. 2005. *Inner Presence: Consciousness as a Biological Phenomenon.*

Nanay, B. 2016. The role of imagination in decision-making. *Mind & Language* 31: 126-142.

Nieuwenstein, M. R., T. Wierenga, R. D. Morey, J. M. Wicherts, T. N. Blom, E. J. Wagenmakers & H. van Rijn. 2015. On making the right choice: a meta-analysis and large-scale replication attempt of the unconscious thought advantage. *Judgment and Decision-Making* 10: 1-17.

Niv, Y., P. Dayan & D. Joel 2006. The effects of motivation on extensively trained behaviour. *Leibniz Technical Report*, Hebrew University 2006-6.

Saggerson, A., D. George & R. C. Honey. 2005. Imitative learning of stimulus-response and response-outcome associations in pigeons. *Journal of Experimental Psychology: Animal Behaviour Processes* 31(3): 289-300.

Schlosser, M. 2012. Taking something as a reason for action. *Philosophical Papers* 41(2): 267-304.

Seth, A. 2009. Functions of consciousness. In Banks, ed., *Elsevier Encyclopedia of Consciousness.*

Sheeran, P., T. L. Webb & P. M. Gollwitzer. 2005. The interplay between goal intentions and implementation intentions. *Personality and Social Psychology Bulletin* 31: 87-98.

Smithies, D. 2011. Attention is rational-access consciousness. In Mole, Smithies & Wu, eds., *Attention: Philosophical and Psychological Essays.*

Sneddon, L. 2011. Pain perception in fish. *Journal of Consciousness Studies* 18: 209-229.

Sneddon, L. 2013. Do painful sensations and fear exist in fish? In *Animal Suffering: From Science to Law.*

Springer, M., W. Murphy, E. Eizirik & S. O'Brien. 2003. Placental mammal diversification and the Cretaceous-Tertiary boundary. *Proceedings of the National Academcy of Sciences* 100(3): 1056-1061.

Tricomi, E., B. W. Balleine & J. P. O'Doherty. 2009. A specific role for posterior dorsolateral striatum in human habit learning. *European Journal of Neuroscience* 29: 2225-2232.
Tye, M. 2017. *Tense Bees and Shell-Shocked Crabs*.
Wegner, D. 2002. *The Illusion of Conscious Will*.
Wood, W. & D. Rünger. 2016. Psychology of habit. *Annual Review of Psychology* 67: 289-314.