

## OPINION

# Algorithms for survival: a comparative perspective on emotions

Dominik R. Bach and Peter Dayan

**Abstract** | The nature and neural implementation of emotions is the subject of vigorous debate. Here, we use Bayesian decision theory to address key complexities in this field and conceptualize emotions in terms of their relationship to survival-relevant behavioural choices. Decision theory indicates which behaviours are optimal in a given situation; however, the calculations required are radically intractable. We therefore conjecture that the brain uses a range of pre-programmed algorithms that provide approximate solutions. These solutions seem to produce specific behavioural manifestations of emotions and can also be associated with core affective dimensions. We identify principles according to which these algorithms are implemented in the brain and illustrate our approach by considering decision making in the face of proximal threat.

Emotions are ineluctably tied to our actions in and perceptions of the world. They organize and colour our behaviour, physiological states and conscious feelings. Perhaps less obviously, they are also a key part of our evolutionary heritage<sup>1</sup> and thus are putatively adaptive. However, empirical debates about emotions abound. This is partly because there are different views, based on divergent definitions of an emotion, that aim at explaining disjunctive sets of phenomena. For example, psychological approaches often put primacy on reported feelings such as fear, anger or happiness. These can be studied in relation to subjective experience in general (which is often collectively termed ‘affect’)<sup>2,3</sup> or in relation to other phenomena such as bodily changes, action tendencies or motivational measures<sup>4,5</sup>. Other approaches<sup>6,7</sup> focus on the facial, prosodic and bodily expression of emotions, partly motivated by comparisons across species<sup>1</sup>. Ethological and neuroscience researchers commonly investigate non-human behaviours labelled with terms such as ‘anxiety-like’ or ‘fear learning’ by way of analogy to humans, albeit noting that such cross-species relationships are not always transparent<sup>8–11</sup>.

It is thus no surprise that the theories that ensue also vary substantially, even to the extent that the very concept of emotion is used at distinct, and sometimes incommensurable, levels of analysis<sup>12</sup> (BOX 1). Emotion is sometimes conceived as being related to the putative goals of an agent (such as seeking information about potential threats when engaging in risk assessment<sup>13</sup>), sometimes to the psychological entities that are associated with observable phenomena (such as the notion of emotional states of fear or anger that cluster together distinct forms of responding to cues and situations<sup>14</sup>), and sometimes to the neural circuits that control behaviour (such as fear circuits<sup>15</sup>). Most often, however, the concept is used in a largely taxonomical manner: to categorize measurable phenomena.

Here, for conciliation, we seek to circumvent the quandaries that are associated with the definitions of emotion. Instead, acknowledging that we eschew qualia (the joyfulness of joy, the fearfulness of fear, and so on), we use decision theory to describe three facets of the determinants of behaviour in specific situations that lead to

phenomena that are often classed as being emotional. The first is a computational analysis (BOX 1) of the goals that humans and other animals pursue when making choices in natural environments and of the actions that may be needed to achieve such goals. The second is an algorithmic analysis (BOX 1) of the procedures that would allow an agent to decide on these actions. We describe specific exemplars of algorithms that seem to control phenomena that are often associated with emotions. The last is an implementational analysis (BOX 1) of the possible neural substrates of these decision-making algorithms. According to this framework, one or more neural controllers are engaged to decide singly or collectively upon a specific response. Sophistication within the controllers, and in their selection and reconciliation, may lead to a substantial heterogeneity in the output, including both phenomena associated with emotions and other overt and covert behaviours.

Bayesian decision theory (BDT) (BOX 1) provides a compelling computational-level prescription of adaptive behaviour. However, it suffers from statistical complexity in its requirement for a large amount of information in novel environments to produce good trajectories of choices and from calculational complexity in the assessment of the expected worth of those choices. We argue that the brain seems to have adopted two major simplifications to approximate optimal choice. Both simplifications are germane to emotions. The first simplification is to use partly pre-programmed algorithms to make these choices<sup>16</sup>. Here, we highlight the surprising richness of these algorithms, noting that they characteristically vary in at least three regards: the inputs that they consider, the extent to which they are plastic and the breadth of actions that they arbitrate. The second simplification is to combine multiple different sorts of algorithm, each of which excels in a different regime of training time and required speed<sup>17</sup>.

Although our approach applies equally to positive and negative circumstances, we mainly focus on decision making under circumstances involving proximal threat and use a decision-theoretic framework to arrange empirically known means to

## Box 1 | Levels of theoretical analysis

In computational neuroscience, it is common to distinguish different levels of analysis that go back to Marr<sup>21</sup>.

### Computational level

At the computational level<sup>21</sup>, theoretical analysis focuses on formalizing the problem that the nervous system has to solve and on finding an appropriate, often optimal or normative, solution. One optimal solution to any decision-making problem is given by Bayesian decision theory (BDT)<sup>119</sup>. According to this theory, agents should create and maintain a so-called belief state that summarizes the whole history of their past observations. To do so, they must use what is known as a generative model of the possible trajectories of environmental states and how those states generate sensory data (note that the 'environment' in this case encompasses the body of the agent). Agents should then make the choices that maximize average long-run benefit by computing an expectation over all possible present and future states along such trajectories. The long-run benefit is typically a weighted sum of the utilities of each possible outcome in the future, with more weight given to outcomes that occur sooner (temporal discounting). Specifying these outcome values is therefore a key ingredient of BDT. The BDT solution is a benchmark that no natural or artificial agent can surpass.

### Algorithmic level

The algorithmic level of analysis concerns how a given problem is solved. Various fields have suggested exact and approximate algorithmic approaches to BDT. These have been given names such as optimal control theory, dynamic programming and reinforcement learning<sup>119–121</sup>. Approximations are necessary because normative solutions are often analytically intractable and cannot even be computed numerically offline in an exact manner. Many neuroscientists use reinforcement learning theory as a formal framework for stating and solving the decision-making problems that they pose to their subjects.

### Implementational level

The implementational level of analysis considers the ways in which algorithms are realized in neural circuits. This spans descriptions on a macroscopic level (brain areas and large populations of neurons), on a mesoscopic level (modestly sized circuits of neurons subject to neuromodulatory influences) and on a microscopic level (within-neuron computations).

achieve survival-relevant output. Threat encompasses many phenomena that are associated with emotions and also raises specific concerns that are somewhat less well explored in the rich field of decision neuroscience.

In this Opinion article, we aim to address several key issues. First, it has been difficult to decide between related emotion theories that try to explain the same phenomena (as exemplified in REF. 4). A decision-theoretic analysis addresses this point by constraining the space of possible algorithms in terms of their efficacy. Second, there is little consensus as to whether emotional phenomena are the output of one or more dedicated mechanisms (for example, specific systems for appraising incoming sensory information<sup>14</sup>) or whether they are manifestations of the operation of more-general-purpose systems (which is how constructionist approaches view the generation of conscious feelings<sup>18</sup>). If there are indeed dedicated mechanisms, we do not know whether they are discrete, whether they are associated with common-sense categories of emotion (such as circuits directly realizing fear), or whether such mechanisms jointly or individually drive dimensional aspects of emotions<sup>14,19</sup>. We show how a

rapprochement between these positions can emerge from a decision-theoretic analysis. Last, we seek to provide clues as to the existence of meta-cognitive, and apparently low-dimensional, representations of affect<sup>2,20</sup>.

### Approximately optimal decisions

At an abstract computational level<sup>21</sup>, appropriate behaviour can be specified by BDT. This maps states of beliefs about the world to optimal choices (BOX 1). The decisions made by humans and other animals often come surprisingly close to those that would be optimal according to BDT in simple, short-run tasks<sup>22–24</sup>. However, apparently simple prescriptions of BDT beg some critical conceptual problems concerning utility functions, limited information and the specification of possible actions. BDT also faces substantial computational challenges in more-complicated problems; this focuses attention on approximations.

**Utility functions.** The first conceptual problem in BDT is a quantification of the costs and benefits that are associated with particular outcomes — this is called a utility function. Evolutionary precepts

suggest that the goal for an individual's preferences should be to prioritize reproductive fitness, including one's own and relatives' survival. However, practically, this metric is unusably long-term.

Behaviour thus seems to be influenced by a range of more-proximal homeostatic forces such as hunger, thirst and (an aversion to) pain. Each of these forces might generate its own utility contribution by quantifying the beneficial or deleterious nature of states or stimuli. If these different utility contributions can be closely approximated as independent and commensurable, then making an overall choice based on their sum would be appropriate. That is, an agent could generate single behaviours that arbitrated as best as possible between seemingly incompatible demands on ultimate reproduction merely by consulting this overall utility.

There is indeed evidence that utility contributions<sup>25</sup> and some forms of approximate overall utility<sup>26</sup> are realized in neural systems. However, it is also known that decision-making algorithms can generate appropriate behaviour without reference to any explicit utility computation. A famous finding in economics is that if an organism's behaviour satisfies some basic principles of rationality, such as consistency and transitivity, then there exists a utility function that is consistent with its choices<sup>27</sup>. Therefore, an organism's behaviour can seem as if it had been generated by a utility function, even if this utility function is purely virtual. Elucidating such cases experimentally poses an obvious challenge.

**Limited information.** The next conceptual problem arises when biological agents have very limited information about very complex environments and, at the same time, when exploratory actions are dangerous, for instance in the face of mortal threat, starvation or dehydration. There are particularly severe computational costs attached to the standard decision-theoretic approach of building hierarchical Bayesian models in which this ignorance about aspects of the model is treated as itself being just another form of uncertainty<sup>28</sup>. One apparent solution to this conundrum is pre-programming: we argue that there are restrictive prior distributions that specify what to expect in the environment and constrained policies that map observations to actions. The pre-specification and the constraints obviate the costs of learning and computation<sup>16,29</sup>.

**Action repertoire.** The final conceptual question relates to the set of actions that are available to the agent. In conventional applications of BDT, this set is of modest size and fully known to the agent. However, in natural environments, the range of possible effective actions can be overwhelming and is at least partly not known. To solve this problem, the agent could compute with a limited action menu that is pre-programmed and/or is a substantial target for transfer from previous learning.

A separate dimension of choice is when or how vigorously to act. A cost–benefit trade-off arises, with the energetic or inaccuracy cost of acting quickly balanced against the opportunity costs of acting slowly<sup>30–32</sup>. In benign environments, opportunity costs are rewards foregone while being slothful and are quantified according to the average reward rate in the environment. In threatening environments, acting slowly may increase exposure to threat. It has been suggested that these two sorts of opportunity cost can be unified by

treating averted potential punishments as being the equivalent of gained rewards<sup>32–35</sup>. Arousal has been interpreted as resulting from the prediction of a need for vigour<sup>36</sup> in terms of this unified opportunity cost. However, it is important to note that acting slowly might in some cases decrease exposure to threat, in which cases animals should either exhibit more haste than speed in active avoidance or engage in passive avoidance<sup>37</sup>.

Along with the conceptual problems described above, another problem for BDT is its formal intractability: the required computations can rarely be performed with viable amounts of time and/or require more storage than is realistically available. A number of generic approximations have therefore been proposed (BOX 2). As we describe below, specific exemplars of these approximations seem to govern behaviour under threat. It is important to note that these particular algorithms are not simple or transparent consequences of BDT itself.

## Box 2 | Types of controller

A controller is a system or device that selects or modulates internal or external actions. Controllers have algorithmic or mathematical descriptions in terms of things such as the constraints that they exactly or approximately enforce; they can also be implemented in neural tissue or in other substrates.

### Pavlovian versus instrumental control

Animal behaviour reflects the influence of different controllers with specific characteristics. Pavlovian control hard-wires certain pre-programmed behaviours to certain events, or to learned predictions thereof, without evaluating the consequences of the actions. In environments or circumstances that are suitably stable (that is, action–outcome contingencies that are expected to be constant over the organism’s whole life), there are advantages to this approach; however, in labile environments, animals must be more flexible. Instrumental control, even if it reflects certain initial biases, can learn to make choices on the basis of the contingency that is experienced between action and outcome, thus providing more flexibility. Pavlovian and instrumental control can be experimentally distinguished by exploiting cases in which hard-wired actions (such as pecking predictors of food pellets in pigeons<sup>122</sup> or rooting with objects associated with food in pigs<sup>123</sup>) are pitted against experimentally determined contingencies (such as denying or delaying rewards that are approached in this way).

### Model-based versus model-free control

At least two canonical methods have been described for making predictions when whole trajectories of future states and possible actions must be considered. Model-based reasoning<sup>120</sup> involves building a precise set of beliefs about the structure of the environment and the outcomes that it affords, and searching the model prospectively at the time of choice through a form of dynamic programming<sup>120</sup>. This has some attractive properties: for example, models are often relatively easy to learn, and choices can be appropriately sensitive to changes in the environment. However, building and searching such models can be ruinously expensive in terms of computation and working memory as the number of future possibilities escalates<sup>113</sup>. Thus, simplification is essential. One simplification is a more general form of model-based control algorithm<sup>70</sup>, in which action–outcome contingencies are assumed to be fixed. This is Pavlovian model-based control, which still involves a representation of a specific goal.

Model-free reasoning provides a radically different method of simplification: estimates or predictions of net long-run utility are learned by experience, based on nothing more than ‘cached’ observations of the utility itself via Pavlovian or instrumental learning rules, without building or using a model. The resulting values are intended to estimate the same quantities that model-based reasoning would produce: the summed expected utility of the future outcomes<sup>120</sup>. Given the way that these estimates are acquired, model-free predictions cannot change immediately if either the worth of the outcomes changes (for example, because of satiation) or the transitions leading to them are altered<sup>124</sup>. This characteristic fixedness allows model-based and model-free values to be discriminated experimentally.

The need to integrate model-based and model-free influences has been considered to be an example of a more general meta-control problem<sup>125–128</sup>, influenced by particular characteristics such as the relative uncertainties of the two sorts of controller<sup>17</sup> or the cost versus benefit of engaging in expensive model-based calculations to overcome potentially incorrect model-free ones<sup>129,130</sup>.

For completeness, we note that model-based instrumental control is sometimes equated with ‘rational’ or ‘non-emotional’ control and contrasted with ‘emotional’ model-free or even Pavlovian control<sup>131</sup>. However, such a characterization is not well supported by the evidence and interpretations that we have adduced.

### Control algorithms for survival

Control algorithms are characterizations of ways that an agent — a machine or an animal — can determine appropriate actions. Efficient control algorithms approximate BDT as closely as possible while minimizing computational costs. Such algorithms can be classified along two orthogonal fault-lines (BOX 2). One concerns action contingency and is associated with the distinction between Pavlovian and instrumental control<sup>38,39</sup>. The other concerns prospective versus retrospective prediction about the future and is associated with the distinction between model-based and model-free control<sup>17,40–42</sup>.

By considering how behaviour under threat is controlled, we can identify several principles. Perhaps the most important in this area is the pre-programming that we mentioned above. One instance of this is Pavlovian control, in which there is an ineluctable coupling of particular predictions to particular actions. However, there are at least three further aspects of pre-programming, all of which arise as limits to flexibility or as a lack of requirement for inference or learning. First, as exemplified in the next section, algorithms often take as input only a selected set of sensory cues and ignore others<sup>43</sup>. Pre-specifying the set that is considered circumvents the more general problem of inferring which cues are relevant<sup>44</sup>. Second is plasticity: the extent to which predictors of important outcomes can be learned *de novo*. Some systems cannot

learn at all and consequently can only operate in a purely pre-specified manner<sup>43</sup>; for others, plasticity is limited<sup>45,46</sup>. Third is that the menu of possible actions may be restricted to different degrees, pre-specifying which is ever even considered<sup>47</sup>. As we describe below, various behaviours seem to be controlled by distinct algorithms that have different pre-programming characteristics and may thus potentially represent separate controllers.

**Consummatory actions.** Consummatory responses — instincts, or fixed action patterns — occur in the presence of evidently significant events, such as imminent or proximal threat. They seem to be substantially pre-programmed; however, they are not hard-wired to the extent that activation of an algorithm leads to the same action pattern every time.

Startling, for instance, is a stereotypical action pattern that is found in many species. It protects a subject from predator attack, is exclusively elicited by a selective set of sensory cues, cannot become associated with other sensory cues via learning and is apparently not altered by unfavourable outcomes<sup>43</sup>. Thus, it seems to be governed by a Pavlovian controller and to be strongly pre-programmed in all three domains described above. However, its magnitude seems to vary according to the prior probability of attack and opportunity costs<sup>48</sup>. Certain other protective actions seem to be more plastic than startling; for example, the eye-blink reflex to corneal air puff<sup>49</sup> can become associated with predictive cues through learning.

Other threat-related consummatory responses include the suite of behaviours that are often labelled as fight, flight and freeze responses<sup>50</sup>. The algorithm underlying these responses putatively infers the proximity of the threat that is a latent cause of the animal's observations (this is known as the 'defensive distance' (REFS 51,52) or 'predatory imminence' (REF. 53)) and makes delicate judgements between the response options. It is often implicitly assumed that this algorithm is Pavlovian and strongly pre-programmed in terms of the action repertoire.

In the absence of mortal threat, unexpected events may require sampling of information, and thus elicit a physiological orienting response<sup>54</sup> and inhibition of goal-oriented behaviour<sup>55</sup>. These responses can co-occur with feelings of surprise in humans<sup>55</sup>. However, the algorithms and implementations involved are less well understood.

On the appetitive side, in non-human species, the manipulation and handling of food, aspects of social interactions between peers and parenting and/or husbandry have been identified as Pavlovian consummatory actions that persist even in the absence of reinforcement. Famous examples include pecking in gull chicks<sup>56</sup>, courtship in sticklebacks<sup>57</sup>, egg moving in geese<sup>57</sup> and potentially elementary eating actions in wild gorillas<sup>58</sup>. The prevalence of such pre-programmed appetitive behaviours is not well researched in humans. They may occur, for example, in the context of affection between infants and parents or between sexual partners.

**Preparatory actions.** When significant events are not yet present but can be predicted from innate or learned precursors, preparatory controllers enter the frame. These often exhibit a substantial degree of plasticity. Predictions can be made in either a model-based or a model-free manner. Model-based predictions of forthcoming outcomes support specific forms of preparation; this could underlie particular bodily responses such as the conditioned protective eye blink<sup>49</sup> or limb withdrawal<sup>59</sup>. Such preparation could be functionally linked to the consummatory responses that the actual arrival of the outcomes would inspire. However, model-based predictions could potentially also support more-general preparatory actions such as approach, avoidance and inhibition. By contrast, model-free predictions are, by their very design, limited to the support of such general preparation because they marginalize away specific outcomes. This means that they can lead to what seems to be suboptimal or self-contradictory choices. For example, in situations in which the outcome is devalued, a subject may execute preparatory actions that get it to a state in which a consummatory response would be possible but then fail to emit that response<sup>40</sup>. Both model-based and model-free predictions could determine a unified opportunity cost of sloth<sup>32</sup>.

Fear responses provide well-known examples of behaviours that are subject to a preparatory controller. These responses include Pavlovian actions that allow the subject to prepare for specific threats<sup>50</sup> and that might arise discretely from model-based algorithms, together with relatively unspecific bodily arousal that could arise from either model-based or model-free control. It has been suggested that preparation for specific threats

arises from multiple separate neural controllers<sup>60</sup>. Precursors of threat are often learned through experience, thus requiring plasticity. This is apparent in cue-conditioned<sup>61</sup> and context-conditioned freezing<sup>62</sup>. Such learning occurs for various sensory stimuli across different modalities, although some stimulus–outcome combinations are apparently more readily learned than others<sup>45,63</sup>, suggesting that there is a pre-programmed restriction on plasticity.

Research on fear has also highlighted instrumental preparation for threat. Examples of this include conditioned active avoidance<sup>33</sup> and the 'escape from fear' paradigm, which involves the *de novo* acquisition of actions that avert predicted threat<sup>47</sup>. Some pre-programmed constraints are apparent in the action repertoire: for example, rats can apparently learn to rear to avoid a threat but not to nose poke<sup>47</sup>.

Finally, a large body of work has described instrumental controllers for obtaining distal reward<sup>64</sup>. This forms a crucial part of the behavioural repertoire for survival in the context of foraging<sup>65,66</sup>, possibly resonating with emotional phenomena such as enthusiasm.

### **Resolving conflict between controllers.**

There may be direct conflict between different controllers' prescriptions, for instance, between Pavlovian and instrumental mechanisms for achieving the same goal or between controllers advocating approach and avoidance (for example, when foraging in conditions of both hunger and threat<sup>52</sup>). In the latter case, the dedicated action pattern that is adopted to resolve such conflict has been termed 'anxiety-like' (REF. 8) and includes passive avoidance (that is, a complete lack of approach). In exploration or foraging paradigms, such avoidance gradually disappears over time<sup>67</sup>. A related response in humans is anxiety-like behavioural inhibition, which has been suggested to be partly under instrumental, and possibly model-based, control<sup>37,68</sup>.

Whenever controllers conflict, arbitration is necessary. One way this might happen is via some common currency reporting strength or importance on an absolute scale. Interestingly, there is an entire field in economics concerned with designing mechanisms that ensure that individual agents achieve common goals. It has been proposed to translate such approach to neuroscience, in our case, by regarding algorithms as individual agents<sup>69</sup>.

**Summary.** In sum, as described above, several control algorithms with distinct features jointly determine an animal's survival-relevant choices. The control of many consummatory behaviours seems to be Pavlovian but model based<sup>70</sup>: that is, it is associated with specific outcomes but does not consider whether the desired outcomes are actually achieved. Furthermore, there seem to be several distinct algorithms in control of these behaviours, characterized by further pre-programming of specific aspects. It is difficult to explain such distinct algorithms in the context of a general-purpose emotion controller, as suggested by some dimensional theories in emotion psychology<sup>14</sup>. Instead, they resonate to a degree with theories that posit the existence of sets of distinct emotions<sup>1,71,72</sup>. On the other hand, the distinct algorithms highlighted here do not map onto the classical emotion categories proposed by basic emotion theory<sup>7,73</sup> and its derivatives. For example, the phenomena classically labelled as 'fear' may involve parallel algorithms, including at least one that does not take previous action outcomes into account (Pavlovian) and one that does (instrumental). Thus, our analysis suggests the existence of discrete algorithmic categories that may not necessarily map neatly onto phenomenological boundaries.

If there is indeed a multiplicity of controllers that are incompletely aware of their own domains of applicability, arbitration may be necessary, which could rely on common currencies. Model-free controllers can, by design, not consider particular goals but only attach scalar quantities to environmental states or actions, as this underpins their formal simplicity. As such, the output of these controllers may be captured in a low-dimensional space with axes such as utility or valence (mediating approach or withdrawal) and arousal (mediating invigoration and inhibition).

As the appropriateness of control algorithms in a particular situation depends on the goals of the organism, substantial variability in their output is to be expected. It is therefore improbable that sharp boundaries can be drawn between phenomenological categories of behaviour as being associated with particular algorithms. Similarly, it may not be possible to enumerate precisely a particular set of algorithms on the basis of just behavioural evidence. Furthermore, organisms that occupy separate ecological niches may also use very distinct controllers.

### Neural circuits for survival

Armed with this basic architecture of control, we now turn our attention to the analysis of their neural implementation. As described above, we have functionally defined a collection of discrete, pre-programmed algorithms and have also identified dimensions such as (predicted) positive and negative utility that drive model-free control or others that might arbitrate between controllers. This discrete and/or dimensional duality is also evident in the neural systems that mediate these control algorithms.

**Multiple neural controllers.** We have proposed the existence of multiple discrete controllers with restricted action menus. Some algorithmically distinct controllers are implemented in close macroscopic proximity. For instance, the controllers for fight and/or flight and for different kinds of freezing behaviour may be anatomically closely related in subdivisions of the periaqueductal grey<sup>74,75</sup> and operate on the basis of the same sensory input. Utility functions that are associated with distinct controllers may be implemented in closely related and rather small neuron populations in the hypothalamus<sup>25</sup>.

In favour of macroscopically separated controllers, circumscribed brain lesions can have a profound and specific impact on emotional behaviour. For example, amygdala lesions impair what is termed cue-conditioned freezing<sup>76</sup> but seem to leave intact some innate anxiety-like behaviour in rodents<sup>77</sup>. The latter are reduced by hippocampal lesions<sup>77,78</sup>, which do not affect cue-conditioned freezing<sup>76</sup>. There are other examples of such specificities: for example, it has been proposed that learning appropriate preparatory actions to specific threats (which algorithmically requires model-based control) may require partly separate and independent neural systems<sup>60</sup>.

In addition, different Pavlovian actions seem to be under the influence of topographically defined regions of the nucleus accumbens. Chemical stimulation of neurons in different parts of this structure can lead to appetitively or aversively directed actions<sup>79,80</sup>, although the loci that relate to each type of action vary according to the familiarity of the context<sup>81</sup>. Such gross dynamic reorganization according to properties of the environment may be a strategy to induce long-term, but not hard-wired, pre-programming of neural decision controllers.

**Distributed neural controllers.** Despite the evidence outlined above, we believe that it is probably inaccurate to consider discrete neural controllers as isolated coherent units that can be defined by their histology, macroscopic structure or transmitter systems. Rather, functional control units that can be separated on an algorithmic level could correspond to distributed and redundant systems on an implementation level. Hierarchically organized controllers may also involve some separate and some shared structures.

For example, learning to predict a specific threat and to elicit an appropriate response to predictors (as in Pavlovian fear conditioning) can be abstractly described by a single decision-making algorithm. However, it seems probable that a considerable array of brain regions is involved<sup>10</sup>. This could include computation of evidence for threat in the amygdala, computation of meta-evidence on the current applicability of this prediction in particular environments in the prefrontal cortex (as occurs, for example, during extinction training<sup>82</sup>), and the additional involvement of sensory cortices for predictors with particular sensory properties<sup>83,84</sup>.

**Scalar representations.** There is also evidence for neural representations of some of the axes of dimensional systems. Neuroimaging studies have demonstrated widespread representation of scalar stimulus valence<sup>85–88</sup> and shared representation of diverse pleasures<sup>26</sup>; electrophysiological recordings show encoding of global utility in the orbitofrontal cortex<sup>89</sup> and of reward-prediction errors across various stimuli in phasic dopaminergic responses<sup>90</sup>. Model-free prediction and control, which lack specific goal directedness, have been ascribed to the central nucleus of the amygdala, the core of the accumbens and the dorsolateral striatum<sup>64,91–95</sup>. Furthermore, tonic dopaminergic responses seem to reflect average reward<sup>30,32</sup>. This duality of discrete and dimensional systems reflects our algorithmic notion that there are discrete controllers that use scalar functions, some of which are shared.

### Arbitration between controllers.

The crucial remaining implementational question concerns the neural basis of arbitration and interaction among the discrete controllers and, at a more systemic level, between model-based and model-free control.

One worked example of this question concerns top-down, model-based inhibition. For instance, in learned helplessness experiments, the overexuberant activity in the serotonergic raphe that is caused by repeated negative outcomes and drives helplessness is apparently suppressed via the medial prefrontal cortex in those subjects that are able to exert control<sup>96</sup>. Thus, one controller that helps to mediate passivity and behavioural inhibition (the raphe) is suppressed by another (the medial prefrontal cortex). Indeed, neuromodulators and neuropeptides<sup>97</sup> could provide a convenient way to communicate dimensional quantities such as utility or arousal globally, in keeping with widespread dopamine<sup>98</sup>, serotonin<sup>99</sup> and noradrenaline<sup>100</sup> projections. Circulating

hormones (for example, stress hormones<sup>101</sup>) could spread even broader influences over even longer timescales.

There is also evidence that instrumental inhibition of Pavlovian misbehaviour is accompanied by particular theta rhythms, which could be signatures or signals associated with regulation<sup>102</sup>. Relevant to this, it is known that controllers of fear and anxiety, which seem to exploit a common microcircuit for storing threat predictions<sup>103</sup>, are associated with amygdala oscillations in the same theta frequency range<sup>52,104–106</sup>.

## Feelings as actions

We have so far considered emotions from the outside looking in. One could adopt a more first-person view and ask about subjective feelings, which in humans often

occur in the absence of overt behaviour. These are, of course, the subject of entire subfields of psychology<sup>107</sup>, and so our aim is just to show how they might fit into the current picture. Importantly, although they are regarded by some as being crucial for the assignment of an emotional label, we here assume that feelings are not required to initiate immediate actions, a proposal that is in line with previous biological and psychological approaches<sup>1,12,108</sup>. This raises two central questions: what are feelings? And what is, if anything, their adaptive function?

In terms of their nature, feelings might be meta-cognitive representations of the inner workings of decision-making systems. They would thus be constructed as the output of more-basic psychological operations<sup>18</sup>. Given the many ways described above that scalar quantities (such as utility and vigour) provide a low-dimensional projection of the bulk of decision-making controllers, it is no surprise that our subjective sense and its verbalization hews substantially to the dimensions of valence and arousal<sup>2–20</sup>.

Various data suggest that experienced (even incidental) feelings influence future decisions, as well as immediate actions<sup>109</sup>. First, there is a suggestion that moods can be understood as long-run averages of short-lasting feelings and that these moods could themselves have an enduring impact on future decisions, acting as forms of generic environmental priors<sup>36,110,111</sup>. Second, although decisions are shaped by currently experienced feelings, they are also influenced by the feelings that are anticipated to occur after relevant outcomes<sup>112</sup>. Hence, feelings experienced in the past may provide sparse and efficient signals for future deliberation of decision outcomes and thus simplify model-based search<sup>113</sup> and/or memory lookup<sup>114</sup>. Such anticipated feelings may be rather abstract or may induce actual feelings<sup>115</sup>. As feelings are only incompletely able to represent the full workings of the various controllers, their influence may seem to be suboptimal or irrational (just as we argued for model-free controllers). Last, an adaptive function of conscious feeling may be to enable verbal communication that relies on conscious access to content. Communication is an aspect of emotions that we have not discussed in this Review (BOX 3).

Overall, the view outlined in this article provides a basis for the existence of dedicated feelings attached to emotional behaviour, something that is only incompletely paralleled in the conscious

## Glossary

### Decision theory

A computational-level theory for making choices given information about states and resulting utilities. Bayesian decision theory is a formally optimal (normative) decision theory.

### Algorithms

In this article, an algorithm denotes an abstract, self-contained set of operations or effective procedures that maps sensory input and internal state to external and internal actions.

### Controllers

In this article, a controller corresponds to a realized neural circuit that is capable of implementing one or several algorithms for choosing or emitting actions.

### Constructionist approaches

A family of theoretical approaches that view subjectively experienced mental categories (such as feelings) as constructed representations of more-basic psychological operations, which are not consciously accessible.

### Utility functions

A real utility function quantifies how useful or dangerous certain outcomes are to an agent, in a given situation, and is realized in the output of actual neural circuits. A virtual utility function is an as-if construct that provides quantifications that are consistent with behavioural choices, but without necessarily underlying those choices.

### Consistency

Choice consistency, or independence, denotes that if A is preferred over B, then A + C is preferred over B + C, irrespective of what C is. This is a fundamental component of expected utility theory and of revealed choice theory.

### Transitivity

Assuming that A is preferred over B and that B is preferred over C, these preferences are said to be transitive if A is also preferred over C. This is a fundamental component of expected utility theory and of revealed choice theory.

### Pre-programming

In this article, pre-programming refers to any restriction on the workings of a controller that can be cast in Bayesian decision theory terms as an immutable, prior mapping of state or prediction to action, or utility function.

### Action contingency

The causal relationship between the execution of actions and the outcomes that result.

### Pavlovian

In this article, the term Pavlovian is used to denote an algorithm or a controller making a choice of actions that is insensitive to the actual consequences of those actions. Here, the term is not used to denote design characteristics of experiments (as is sometimes the case).

### Instrumental

In this article, the term instrumental refers to an algorithm or a controller making choices that are contingent on their past or predicted future consequences. Here, the term does not refer to design characteristics of experiments.

### Model-based

In this article, the term model-based is used to characterize algorithms that exploit a model of the structure of the environment and the outcomes that it affords to make long-run predictions about the future. Predictions need not be action contingent and thus can support either Pavlovian or instrumental controllers.

### Model-free

In this article, the term model-free is used to describe algorithms that learn to make long-run predictions by caching or saving experiences from the past, generally by enforcing self-consistency in successive outputs. Predictions are typically scalar, for instance, of summed future value and consequently do not encode the specific outcomes underpinning those values. Model-free predictions need not be action contingent and thus can support either Pavlovian or instrumental controllers.

### Appraisal theory

A family of emotion theories, all of which posit that manifestations of emotions (feelings, motivational processes, bodily reactions, and so on) are the output of a set of cognitive appraisals or encompass such appraisals. Theories differ widely according to the appraisals that they consider part of the set.

## Box 3 | Cooperation and competition

In contexts that include multiple agents, communicative actions — which are another output of emotional systems — require game-theoretic considerations of competition and cooperation<sup>132</sup>. Cooperation includes the possibility of learning from vicarious rewards and punishments supplied to others whose reactions we can observe — something that can, of course, be expensive for them.

When in competition, subjects have an incentive to produce fallacious external (but not internal) reactions, such as false emotions (or ‘cheap talk’). Thus, it could be useful to pretend to be distraught to get other people to help or (as in the game of chicken) to pretend to be angry (that is, to imply that one will perform self-harming or irrational actions such as fighting) to gain concessions or dominance. The possibility of deception provides one’s interlocutors with the incentive to detect and punish cheaters; evolutionary game theory provides some hints about how such long-run battles might end, helping to address the important questions as to how individual-level emotional characteristics maximize population-level fitness<sup>133,134</sup> and how heterogeneity across a population might result<sup>135</sup>.

Of course, although the internal representations that we use might be based on the same regularities of behaviour that external third parties also observe, even the meta-cognitively challenged people among us are able to access much richer, covert, physiological and neural signals about our own states, including a machinery for interoceptive inference<sup>136</sup>. Thus, we can have a more accurate and faithful model of ourselves than we can of others. Sadly, in an internal form of competition, when considering ourselves, we might be susceptible to self-serving editing (also known as Pavlovian pruning<sup>113</sup>) such that we could be less biased observers of others than of ourselves.

perception of other mental operations. The existence of such feelings would thus explain the lingering differentiation between cold and hot cognition in neuroscience research, even though such a distinction may not exist in terms of the mathematical or even neural structures of the inferences concerned<sup>116,117</sup>.

### Conclusion

Emotion is a vast and critical topic. We have tried to provide a formal foundation for a computationally oriented study of emotions. Our decision-theoretic approach resonates with a central tenet of appraisal theories of emotion: emotional phenomena are the output of a system for response optimization<sup>14</sup>, just like any other behaviour. We therefore analysed the goals of behaviour in biological environments, dissected emotions into associated actions and feelings, and characterized aspects of the particular decision-making algorithms that govern these actions. We exploited parallels with reward-based decision making in which the decision-theoretic analysis of model-based and model-free, as well as Pavlovian and instrumental, control has been more extensively examined. However, our focus on threat allowed us to highlight the crucial importance of pre-programming in controlling phenomena that are often associated with emotions. We discussed some of the evidence for multiple, discrete, neurally distinct decision-making systems that do not map onto classical phenomenological emotion categories, as well as for scalar systems that support dimensions of behaviour and possibly also feeling.

This extends and joins previous accounts that either assumed phenomenological emotion categories (such as basic emotions<sup>7,73</sup>) or proposed non-modular, dimensional systems that can contribute to more than one common-sense emotion category (as, for example, in many instances of appraisal theory or in constructed emotion theory<sup>18,118</sup>). Many computational, psychological and neural questions remain, and we hope to have furnished a useful framework for answering them.

*Dominik R. Bach is at the Division for Clinical Psychiatry Research, Psychiatric Hospital, University of Zurich; at the Neuroscience Centre Zurich, University of Zurich, 8032 Zurich, Switzerland; and at the Wellcome Trust Centre for Neuroimaging, University College London, London WC1N 3BG, UK.*

*Peter Dayan is at the Gatsby Computational Neuroscience Unit, University College London, London W1T 4JG, UK.*

Correspondence to D.R.B.

[dominik.bach@uzh.ch](mailto:dominik.bach@uzh.ch)

[doi:10.1038/nrn.2017.35](https://doi.org/10.1038/nrn.2017.35)

Published online 31 Mar 2017

- Darwin, C. *The Expression of the Emotions in Man and Animals* Vol. 1965 (Univ. of Chicago Press, 1872).
- Russell, J. A. A circumplex model of affect. *J. Pers. Soc. Psychol.* **39**, 1161–1178 (1980).
- Kuppens, P., Tuerlinckx, F., Russell, J. A. & Barrett, L. F. The relation between valence and arousal in subjective experience. *Psychol. Bull.* **139**, 917–940 (2013).
- Scherer, K. R., Schorr, A. & Johnstone, T. *Appraisal Processes in Emotion: Theory, Methods, Research* (Oxford Univ. Press, 2001).
- Oatley, K. & Johnson-Laird, P. N. Cognitive approaches to emotions. *Trends Cogn. Sci.* **18**, 134–140 (2014).
- Ekman, P. & Oster, H. Facial expressions of emotion. *Annu. Rev. Psychol.* **30**, 527–554 (1979).
- Izard, C. E. Basic emotions, relations among emotions, and emotion–cognition relations. *Psychol. Rev.* **99**, 561–565 (1992).
- Calhoun, G. G. & Tye, K. M. Resolving the neural circuits of anxiety. *Nat. Neurosci.* **18**, 1394–1404 (2015).
- Burgdorf, J. & Panksepp, J. The neurobiology of positive emotions. *Neurosci. Biobehav. Rev.* **30**, 173–187 (2006).
- Herry, C. & Johansen, J. P. Encoding of fear learning and memory in distributed neuronal circuits. *Nat. Neurosci.* **17**, 1644–1654 (2014).
- Stephan, K. E. *et al.* Charting the landscape of priority problems in psychiatry, part 1: classification and diagnosis. *Lancet Psychiatry* **3**, 77–83 (2016).
- LeDoux, J. E. Coming to terms with fear. *Proc. Natl Acad. Sci. USA* **111**, 2871–2878 (2014).
- Blanchard, D. C., Griebel, G., Pobbe, R. & Blanchard, R. J. Risk assessment as an evolved threat detection and analysis process. *Neurosci. Biobehav. Rev.* **35**, 991–998 (2011).
- Scherer, K. R. in *Appraisal Processes in Emotion* (eds Scherer, K. R., Schorr, A. & Johnstone, T.) 92–120 (Oxford Univ. Press, 2001).
- Wilensky, A. E., Schafe, G. E., Kristensen, M. P. & LeDoux, J. E. Rethinking the fear circuit: the central nucleus of the amygdala is required for the acquisition, consolidation, and expression of Pavlovian fear conditioning. *J. Neurosci.* **26**, 12387–12396 (2006).
- Dayan, P., Niv, Y., Seymour, B. & Daw, N. D. The misbehavior of value and the discipline of the will. *Neural Netw.* **19**, 1153–1160 (2006).
- Daw, N. D., Niv, Y. & Dayan, P. Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nat. Neurosci.* **8**, 1704–1711 (2005).
- Lindquist, K. A. & Barrett, L. F. A functional architecture of the human brain: emerging insights from the science of emotion. *Trends Cogn. Sci.* **16**, 533–540 (2012).
- Jack, R. E., Garrod, O. G. & Schyns, P. G. Dynamic facial expressions of emotion transmit an evolving hierarchy of signals over time. *Curr. Biol.* **24**, 187–192 (2014).
- Nowlis, H. H. & Nowlis, V. The description and analysis of mood. *Ann. NY Acad. Sci.* **65**, 345–355 (1956).
- Marr, D. C. & Poggio, T. From understanding computation to understanding neural circuitry. *Neurosci. Res. Program Bull.* **15**, 470–491 (1977).
- Kording, K. P. & Wolpert, D. M. Bayesian decision theory in sensorimotor control. *Trends Cogn. Sci.* **10**, 319–326 (2006).
- Bach, D. R. & Dolan, R. J. Knowing how much you don’t know: a neural organization of uncertainty estimates. *Nat. Rev. Neurosci.* **13**, 572–586 (2012).
- Pouget, A., Beck, J. M., Ma, W. J. & Latham, P. E. Probabilistic brains: knowns and unknowns. *Nat. Neurosci.* **16**, 1170–1178 (2013).
- Graebner, A. K., Iyer, M. & Carter, M. E. Understanding how discrete populations of hypothalamic neurons orchestrate complicated behavioral states. *Front. Syst. Neurosci.* **9**, 111 (2015).
- Berridge, K. C. & Kringelbach, M. L. Pleasure systems in the brain. *Neuron* **86**, 646–664 (2015).
- Samuelson, P. A. Note on the pure theory of consumers’ behaviour. *Economica* **5**, 61–71 (1938).
- Blondel, V. D. & Tsitsiklis, J. N. A survey of computational complexity results in systems and control. *Automatica* **36**, 1249–1274 (2000).
- Hinton, G. E. & Nowlan, S. J. How learning can guide evolution. *Complex Syst.* **1**, 495–502 (1987).
- Niv, Y., Daw, N. D., Joel, D. & Dayan, P. Tonic dopamine: opportunity costs and the control of response vigor. *Psychopharmacology* **191**, 507–520 (2007).
- Choi, J. E., Waswani, P. A. & Shadmehr, R. Vigor of movements and the cost of time in decision making. *J. Neurosci.* **34**, 1212–1223 (2014).
- Dayan, P. Instrumental vigour in punishment and reward. *Eur. J. Neurosci.* **35**, 1152–1168 (2012).
- Mowrer, O. H. 2-Factor learning theory — summary and comment. *Psychol. Rev.* **58**, 350–354 (1951).
- Maia, T. V. Two-factor theory, the actor–critic model, and conditioned avoidance. *Learn. Behav.* **38**, 50–67 (2010).
- Lloyd, K. & Dayan, P. Safety out of control: dopamine and defence. *Behav. Brain Funct.* **12**, 15 (2016).
- Trimmer, P. C., Paul, E. S., Mendl, M. T., McNamara, J. M. & Houston, A. I. On the evolution and optimality of mood States. *Behav. Sci. (Basel)* **3**, 501–521 (2013).

37. Bach, D. R. Anxiety-like behavioural inhibition is normative under environmental threat-reward correlations. *PLoS Comput. Biol.* **11**, e1004646 (2015).
38. Mackintosh, N. J. *Conditioning and Associative Learning* (Oxford Univ. Press, 1983).
39. Dickinson, A. *Contemporary Animal Learning Theory* Vol. 1 (CUP Archive, 1980).
40. Dickinson, A. & Balleine, B. in *Steven's Handbook of Experimental Psychology: Learning, Motivation and Emotion* Vol. 3 (eds Pashler, H. & Gallistel, R.) 497–534 (2002).
41. Doya, K. What are the computations of the cerebellum, the basal ganglia and the cerebral cortex? *Neural Netw.* **12**, 961–974 (1999).
42. Adams, C. D. & Dickinson, A. in *Information Processing in Animals: Memory Mechanisms* (eds Spear, N. E. & Miller, R. E.) 143–165 (Psychology Press, 1981).
43. Yeomans, J. S., Li, L., Scott, B. W. & Frankland, P. W. Tactile, acoustic and vestibular systems sum to elicit the startle reflex. *Neurosci. Biobehav. Rev.* **26**, 1–11 (2002).
44. Wilson, R. C. & Niv, Y. Inferring relevance in a changing world. *Front. Hum. Neurosci.* **5**, 189 (2011).
45. Garcia, J., McGowan, B. K., Ervin, F. R. & Koelling, R. A. Cues: their relative effectiveness as a function of the reinforcer. *Science* **160**, 794–795 (1968).
46. Rozin, P., Gruss, L. & Berk, G. Reversal of innate aversions: attempts to induce a preference for chili peppers in rats. *J. Comp. Physiol. Psychol.* **93**, 1001–1014 (1979).
47. Cain, C. K. & LeDoux, J. E. Escape from fear: a detailed behavioral analysis of two atypical responses reinforced by CS termination. *J. Exp. Psychol. Anim. Behav. Process.* **33**, 451–463 (2007).
48. Bach, D. R. A cost minimisation and Bayesian inference model predicts startle reflex modulation across species. *J. Theor. Biol.* **370**, 53–60 (2015).
49. Kehoe, E. J. & Macrae, M. in *A Neuroscientist's Guide to Classical Conditioning* 171–231 (Springer, 2002).
50. Blanchard, D. C. & Blanchard, R. J. Ethoexperimental approaches to the biology of emotion. *Annu. Rev. Psychol.* **39**, 43–68 (1988).
51. McNaughton, N. & Corr, P. J. A two-dimensional neuropsychology of defense: fear/anxiety and defensive distance. *Neurosci. Biobehav. Rev.* **28**, 285–305 (2004).
52. Gray, J. A. & McNaughton, N. *The Neuropsychology of Anxiety: an Enquiry into the Functions of the Septohippocampal System* Vol. 2 (Oxford Univ. Press, 2000).
53. Fanselow, M. S. & Lester, L. S. in *Evolution and Learning* (eds Bolles, R. C. & Beecher, M. D.) 185–212 (Psychology Press, 1988).
54. Siddle, D. A. Orienting, habituation, and resource allocation: an associative analysis. *Psychophysiology* **28**, 245–259 (1991).
55. Schutzwohl, A. Surprise and schema strength. *J. Exp. Psychol. Learn. Mem. Cogn.* **24**, 1182–1199 (1998).
56. Hailman, J. P. *The Ontogeny of an Instinct: the Pecking Response in Chicks of the Laughing Gull (*Larus atricilla* L.) and Related Species* (Brill, 1967).
57. Tinbergen, N. *The Study of Instinct* (Oxford Univ. Press, 1951).
58. Byrne, R. W. & Byrne, J. M. E. Complex leaf-gathering skills of mountain gorillas (*Corilla g. beringei*): variability and standardization. *Am. J. Primatol.* **31**, 241–261 (1993).
59. Zhang, S., Mano, H., Ganesh, G., Robbins, T. & Seymour, B. Dissociable learning processes underlie human pain conditioning. *Curr. Biol.* **26**, 52–58 (2016).
60. Gross, C. T. & Canteras, N. S. The many paths to fear. *Nat. Rev. Neurosci.* **13**, 651–658 (2012).
61. LeDoux, J. E., Sakaguchi, A., Iwata, J. & Reis, D. J. Auditory emotional memories: establishment by projections from the medial geniculate nucleus to the posterior neostriatum and/or dorsal amygdala. *Ann. NY Acad. Sci.* **444**, 463–464 (1985).
62. Kim, J. J. & Fanselow, M. S. Modality-specific retrograde amnesia of fear. *Science* **256**, 675–677 (1992).
63. Ohman, A. & Mineka, S. Fears, phobias, and preparedness: toward an evolved module of fear and fear learning. *Psychol. Rev.* **108**, 483–522 (2001).
64. Dolan, R. J. & Dayan, P. Goals and habits in the brain. *Neuron* **80**, 312–325 (2013).
65. Korn, C. W. & Bach, D. R. Maintaining homeostasis by decision-making. *PLoS Comput. Biol.* **11**, e1004301 (2015).
66. Rangel, A. Regulation of dietary choice by the decision-making circuitry. *Nat. Neurosci.* **16**, 1717–1724 (2013).
67. Foniio, E., Benjamini, Y. & Golani, I. Freedom of movement and the stability of its unfolding in free exploration of mice. *Proc. Natl Acad. Sci. USA* **106**, 21335–21340 (2009).
68. Bach, D. R. The cognitive architecture of anxiety-like behavioral inhibition. *J. Exp. Psychol. Hum. Percept. Perform.* **43**, 18–29 (2017).
69. Alonso, R., Brocas, I. & Carrillo, J. D. Resource allocation in the brain. *Rev. Econom. Studies* **81**, 501–534 (2014).
70. Dayan, P. & Berridge, K. C. Model-based and model-free Pavlovian reward learning: reevaluation, revision, and revelation. *Cogn. Affect Behav. Neurosci.* **14**, 473–492 (2014).
71. Tomkins, S. S. & McCarter, R. What and where are the primary affects? Some evidence for a theory. *Percept. Mot. Skills* **18**, 119–158 (1964).
72. Ekman, P., Friesen, W. V. & Ellsworth, P. *Emotion in the Human Face: Guide-Lines for Research and an Integration of Findings* (Pergamon Press, 1972).
73. Ekman, P. Are there basic emotions? *Psychol. Rev.* **99**, 550–553 (1992).
74. Brandao, M. L., Zanoveli, J. M., Ruiz-Martinez, R. C., Oliveira, L. C. & Landeira-Fernandez, J. Different patterns of freezing behavior organized in the periaqueductal gray of rats: association with different types of anxiety. *Behav. Brain Res.* **188**, 1–13 (2008).
75. Keay, K. A., Clement, C. I., Owler, B., Depaulis, A. & Bandler, R. Convergence of deep somatic and visceral nociceptive information onto a discrete ventrolateral midbrain periaqueductal gray region. *Neuroscience* **61**, 727–732 (1994).
76. Phillips, R. G. & LeDoux, J. E. Differential contribution of amygdala and hippocampus to cued and contextual fear conditioning. *Behav. Neurosci.* **106**, 274–285 (1992).
77. McHugh, S. B., Deacon, R. M., Rawlins, J. N. & Banner, D. M. Amygdala and ventral hippocampus contribute differentially to mechanisms of fear and anxiety. *Behav. Neurosci.* **118**, 63–78 (2004).
78. Bach, D. R. et al. Human hippocampus arbitrates approach–avoidance conflict. *Curr. Biol.* **24**, 541–547 (2014).
79. Reynolds, S. M. & Berridge, K. C. Fear and feeding in the nucleus accumbens shell: rostrocaudal segregation of GABA-elicited defensive behavior versus eating behavior. *J. Neurosci.* **21**, 3261–3270 (2001).
80. Reynolds, S. M. & Berridge, K. C. Positive and negative motivation in nucleus accumbens shell: bivalent rostrocaudal gradients for GABA-elicited eating, taste “liking”/“disliking” reactions, place preference/avoidance, and fear. *J. Neurosci.* **22**, 7308–7320 (2002).
81. Reynolds, S. M. & Berridge, K. C. Emotional environments retune the valence of appetitive versus fearful functions in nucleus accumbens. *Nat. Neurosci.* **11**, 423–425 (2008).
82. Sharpe, M. J. & Killcross, S. The prelimbic cortex uses higher-order cues to modulate both the acquisition and expression of conditioned fear. *Front. Syst. Neurosci.* **8**, 235 (2014).
83. Ohl, F. W., Wetzel, W., Wagner, T., Rech, A. & Scheich, H. Bilateral ablation of auditory cortex in Mongolian gerbil affects discrimination of frequency modulated tones but not of pure tones. *Learn. Mem.* **6**, 347–362 (1999).
84. Letzkus, J. J. et al. A disinhibitory microcircuit for associative fear learning in the auditory cortex. *Nature* **480**, 331–335 (2011).
85. Winston, J. S., Gottfried, J. A., Kilner, J. M. & Dolan, R. J. Integrated neural representations of odor intensity and affective valence in human amygdala. *J. Neurosci.* **25**, 8903–8907 (2005).
86. Lewis, P., Critchley, H., Rotshtein, P. & Dolan, R. Neural correlates of processing valence and arousal in affective words. *Cereb. Cortex* **17**, 742–748 (2007).
87. Chikazoe, J., Lee, D. H., Kriegeskorte, N. & Anderson, A. K. Population coding of affect across stimuli, modalities and individuals. *Nat. Neurosci.* **17**, 1114–1122 (2014).
88. Lindquist, K. A., Satpute, A. B., Wager, T. D., Weber, J. & Barrett, L. F. The brain basis of positive and negative affect: evidence from a meta-analysis of the human neuroimaging literature. *Cereb. Cortex* **26**, 1910–1922 (2016).
89. Critchley, H. D. & Rolls, E. T. Hunger and satiety modify the responses of olfactory and visual neurons in the primate orbitofrontal cortex. *J. Neurophysiol.* **75**, 1673–1686 (1996).
90. Schultz, W., Dayan, P. & Montague, P. R. A neural substrate of prediction and reward. *Science* **275**, 1593–1599 (1997).
91. Killcross, S. & Coutureau, E. Coordination of actions and habits in the medial prefrontal cortex of rats. *Cereb. Cortex* **13**, 400–408 (2003).
92. Balleine, B. W. Neural bases of food-seeking: affect, arousal and reward in corticostriatal limbic circuits. *Physiol. Behav.* **86**, 717–730 (2005).
93. Bassareo, V. & Di Chiara, G. Differential responsiveness of dopamine transmission to food-stimuli in nucleus accumbens shell/core compartments. *Neuroscience* **89**, 637–641 (1999).
94. Corbit, L. H. & Balleine, B. W. The general and outcome-specific forms of Pavlovian-instrumental transfer are differentially mediated by the nucleus accumbens core and shell. *J. Neurosci.* **31**, 11786–11794 (2011).
95. Liljeholm, M. & O’Doherty, J. P. Contributions of the striatum to learning, motivation, and performance: an associative account. *Trends Cogn. Sci.* **16**, 467–475 (2012).
96. Amat, J., Paul, E., Zarza, C., Watkins, L. R. & Maier, S. F. Previous experience with behavioral control over stress blocks the behavioral and dorsal raphe nucleus activating effects of later uncontrollable stress: role of the ventral medial prefrontal cortex. *J. Neurosci.* **26**, 13264–13272 (2006).
97. Dayan, P. Twenty-five lessons from computational neuromodulation. *Neuron* **76**, 240–256 (2012).
98. Fallon, J. H. Topographic organization of ascending dopaminergic projections. *Ann. NY Acad. Sci.* **537**, 1–9 (1988).
99. Hale, M. W. & Lowry, C. A. Functional topography of midbrain and pontine serotonergic systems: implications for synaptic regulation of serotonergic circuits. *Psychopharmacology* **213**, 243–264 (2011).
100. Berridge, C. W. & Waterhouse, B. D. The locus coeruleus-noradrenergic system: modulation of behavioral state and state-dependent cognitive processes. *Brain Res. Brain Res. Rev.* **42**, 33–84 (2003).
101. Liang, K. C. et al. Corticotropin-releasing factor: long-lasting facilitation of the acoustic startle reflex. *J. Neurosci.* **12**, 2303–2312 (1992).
102. Cavanagh, J. F., Eisenberg, I., Guitart-Masip, M., Huys, Q. & Frank, M. J. Frontal theta overrides pavlovian learning biases. *J. Neurosci.* **33**, 8541–8548 (2013).
103. Tovote, P., Fadok, J. P. & Luthi, A. Neuronal circuits for fear and anxiety. *Nat. Rev. Neurosci.* **16**, 317–331 (2015).
104. Likhtik, E. et al. Prefrontal entrainment of amygdala activity signals safety in learned fear and innate anxiety. *Nat. Neurosci.* **17**, 106–113 (2014).
105. Adhikari, A., Topiwala, M. A. & Gordon, J. A. Synchronized activity between the ventral hippocampus and the medial prefrontal cortex during anxiety. *Neuron* **65**, 257–269 (2010).
106. Adhikari, A., Topiwala, M. A. & Gordon, J. A. Single units in the medial prefrontal cortex with anxiety-related firing patterns are preferentially influenced by ventral hippocampal activity. *Neuron* **71**, 898–910 (2011).
107. Roseman, I. J. & Smith, C. A. in *Appraisal Processes in Emotion* Ch. 1 (eds Scherer, K. R., Schorr, A. & Johnstone, T.) 3–19 (Oxford Univ. Press, 2001).
108. Frijda, N. H. *The Laws of Emotion* (Lawrence Erlbaum Associates, 2007).
109. Dolan, R. J. Emotion, cognition, and behavior. *Science* **298**, 1191–1194 (2002).
110. Mendl, M., Burman, O. H. & Paul, E. S. An integrative and functional framework for the study of animal emotion and mood. *Proc. Biol. Sci.* **277**, 2895–2904 (2010).
111. Eldar, E., Rutledge, R. B., Dolan, R. J. & Niv, Y. Mood as representation of momentum. *Trends Cogn. Sci.* **20**, 15–24 (2016).
112. DeWall, C. N., Baumeister, R. F., Chester, D. S. & Bushman, B. J. How often does currently felt emotion predict social behavior and judgment? A meta-analytic test of two theories. *Emot. Rev.* **8**, 136–143 (2016).
113. Huys, Q. J. et al. Bonsai trees in your head: how the pavlovian system sculpts goal-directed choices by pruning decision trees. *PLoS Comput. Biol.* **8**, e1002410 (2012).



114. Redish, A. D. Vicarious trial and error. *Nat. Rev. Neurosci.* **17**, 147–159 (2016).
115. Loewenstein, G. & Lerner, J. S. in *Handbook of Affective Sciences* (eds Davidson, R. J., Scherer, K. R. & Goldsmith, H. H.) (Oxford University Press, 2003).
116. Pessoa, L. & Adolphs, R. Emotion processing and the amygdala: from a 'low road' to 'many roads' of evaluating biological significance. *Nat. Rev. Neurosci.* **11**, 775–783 (2010).
117. Pessoa, L. On the relationship between emotion and cognition. *Nat. Rev. Neurosci.* **9**, 148–158 (2008).
118. Barrett, L. F. The theory of constructed emotion: an active inference account of interoception and categorization. *Soc. Cogn. Affect Neurosci.* **12**, 1–23 (2016).
119. Berger, J. O. *Statistical Decision Theory and Bayesian Analysis* (Springer Science & Business Media, 2013).
120. Sutton, R. S. & Barto, A. G. *Reinforcement Learning: an Introduction* (MIT Press, 1998).
121. Bertsekas, D. P. *Dynamic Programming and Optimal Control* Vol. 1 (Athena Scientific Belmont, 2005).
122. Schwartz, B. & Williams, D. R. The role of the response-reinforcer contingency in negative automaintenance. *J. Exp. Anal. Behav.* **17**, 351–357 (1972).
123. Breland, K. & Breland, M. The misbehavior of organisms. *Am. Psychol.* **16**, 681–684 (1961).
124. Dickinson, A. & Balleine, B. Motivational control of goal-directed action. *Anim. Learn. Behav.* **22**, 1–18 (1994).
125. Boureau, Y. L., Sokol-Hessner, P. & Daw, N. D. Deciding how to decide: self-control and meta-decision making. *Trends Cogn. Sci.* **19**, 700–710 (2015).
126. Dayan, P. How to set the switches on this thing. *Curr. Opin. Neurobiol.* **22**, 1068–1074 (2012).
127. Alexander, W. H. & Brown, J. W. Computational models of performance monitoring and cognitive control. *Top. Cogn. Sci.* **2**, 658–677 (2010).
128. Botvinick, M. & Weinstein, A. Model-based hierarchical reinforcement learning and human action control. *Phil. Trans. R. Soc. B Biol. Sci.* **369**, 20130480 (2014).
129. Keramati, M., Dezfouli, A. & Piray, P. Speed/accuracy trade-off between the habitual and the goal-directed processes. *PLoS Comput. Biol.* **7**, e1002055 (2011).
130. Pezzulo, G., Rigoli, F. & Chersi, F. The mixed instrumental controller: using value of information to combine habitual choice and mental simulation. *Front. Psychol.* **4**, 92 (2013).
131. Kahneman, D. *Thinking, Fast and Slow* (Farrar, 2011).
132. Camerer, C. *Behavioral Game Theory: Experiments in Strategic Interaction* (Princeton Univ. Press, 2003).
133. Devaine, M., Hollard, G. & Daunizeau, J. Theory of mind: did evolution fool us? *PLoS ONE* **9**, e87619 (2014).
134. Johnson, D. D. & Fowler, J. H. The evolution of overconfidence. *Nature* **477**, 317–320 (2011).
135. Giske, J. *et al.* Effects of the emotion system on adaptive behavior. *Am. Nat.* **182**, 689–703 (2013).
136. Seth, A. K. Interoceptive inference, emotion, and the embodied self. *Trends Cogn. Sci.* **17**, 565–573 (2013).

#### Acknowledgements

The authors thank G. Castegnetti, L. Harris, C. Korn, M. Mendi, H. Nakahara, L. Paul, the reviewers and many others for inspiring discussions during the writing of this article. This work was supported by the University of Zurich (to D.R.B.), the Gatsby Charitable Foundation (to P.D.) and a grant from the UK National Centre for the Replacement Refinement and Reduction of Animals in Research (K/00008X/1; to M. Mendi, E. Paul and P.D.). The Wellcome Trust Centre for Neuroimaging is supported by core funding from the Wellcome Trust (091593/Z/10/Z).

#### Competing interests statement

The authors declare no competing interests.