Opinion

Algorithms for survival: a comparative perspective on emotions

Dominik R. Bach[1-3] *& Peter Dayan*[4]

[1]Department of Psychiatry, Psychotherapy, and Psychosomatics, University of Zurich, Switzerland

[2]Neuroscience Centre Zurich, University of Zurich, 8057 Zurich, Switzerland

[3]Wellcome Trust Centre for Neuroimaging, University College London, London WC1N 3BG, UK

[4]Gatsby Computational Neuroscience Unit, University College London, London W1T 4JG, UK

Correspondence to D.R.B. email: dominik.bach@uzh.ch

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 employs a range of pre-programmed algorithms that provide approximate solutions. These solutions appear 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.

Introduction

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[1], 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")[2,3], or in relation to other phenomena such as bodily changes, action tendencies, or motivational measures[4,5]. Other approaches[6,7] focus on the facial, prosodic and bodily expression of emotions, partly motivated by comparisons across species[1]. 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[8-11].

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 (Box 1)[12]. Sometimes emotion is conceived as being related to the putative goals of an agent (such as seeking information about potential threats when engaging in risk assessment[13]); sometimes to the psychological entities associated with observable phenomena (such as the notion of emotional states of fear and anger that cluster together distinct forms of responding to cues and situations[14]); and sometimes to the neural circuits controlling behaviour (such as fear circuits[15]). Most often, however, the concept is used in a largely taxonomical manner: to categorise measurable phenomena.

Here, for conciliation, we seek to circumvent the quandaries associated with definitions of emotion. Instead, acknowledging that we eschew qualia (the joyfulness of joy or the fearfulness of fear and the like), 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 which actions may achieve such goals. The second is an algorithmic analysis (Box 1) of the procedures that would enable an agent to decide on these actions. We describe specific exemplars of algorithms that appear to control phenomena 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 which 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 also 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 calculational complexity in the assessment of the expected worth of those choices. We argue that the brain appears 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[16]; we highlight their surprising richness noting that they characteristically vary in at least three regards: the inputs they consider, the extent to which they are plastic, and the breadth of actions 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[17].

Although our approach applies equally to positive and negative circumstances, we mainly focus on decision-making under circumstances involving proximal threat, using a decision-theoretic framework to arrange empirically-known means to achieve survival-relevant output. Threat encompasses many phenomena 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[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[14]) or whether they are manifestations of the operation of more general-purpose systems (which is how **constructionist** approaches view the generation of conscious feelings[18]). If there are indeed dedicated mechanisms, we do not know whether they are discrete, or 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[14,19]. We show how a rapprochement between these positions can emerge from a decision-theoretic analysis. Finally, we seek to provide clues as to the existence of meta-cognitive, and apparently low dimensional, representations of affect[2,20].

[H1] Approximately optimal decisions

At an abstract computational level[21], 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[22-24]. However, BDT's apparently simple prescriptions beg a number of 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.

[H3] Utility functions

The first conceptual problem in BDT is a quantification of the costs and benefits associated with particular outcomes - this is called a utility function. Evolutionary precepts suggest the goal for an individual's preferences should be to prioritize reproductive fitness, including one's own and one's relatives' survival. Practically, however, this metric is unusably long-term. Behaviour thus appears to be influenced by a range of more proximal homeostatic

forces such as hunger, thirst, and (an aversion to) pain. Each such force 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[25] and some forms of approximate overall utility[26] 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[27]. Therefore, an organism's behaviour can appear 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.

[H3] Limited information

The next conceptual problem arises when biological agents have very limited information about very complex environments, and at the same time 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[28]. 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[16,29].

[H3] 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 unknown. 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[30-32]. In benign environments, opportunity costs are rewards foregone whilst 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[32-35]. Arousal has been interpreted as resulting from the prediction of a need for vigour[36] 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[37].

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 appear to govern behaviour under threat. It is important to note that these particular algorithms are not simple or transparent consequences of BDT itself.

[H1] 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[38,39]. The other concerns prospective versus retrospective prediction about the future, and is associated with the distinction between **model-based** and **model-free** control[17,40-42].

By considering how behaviour under threat is controlled, we can identify several principles. Perhaps the most important in this area is the pre-programming 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 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[43]. Pre-specifying the set that is considered circumvents the more general problem of inferring which are relevant[44]. Second is plasticity: the extent to which predictors of important outcomes can be learned *de novo*? Some systems cannot learn at all, and so can only operate in a purely pre-specified manner[43]; for others, plasticity is limited[45,46]. Third is that the menu of possible actions may be restricted to different degrees, pre-specifying which is ever even considered[47]. As we describe below, various behaviours appear to be controlled by distinct algorithms that have different pre-programming characteristics, and may thus potentially represent separate controllers.

[H3] Consummatory actions

Consummatory responses — instincts, or fixed action patterns — occur in the presence of evidently significant events, such as imminent or proximal threat. They appear 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[43]. Thus, it appears to be governed by a Pavlovian controller and to be strongly pre-programmed in all of the three domains described above. However, its magnitude appears to vary according to both the prior probability of attack and opportunity costs[48]. Certain other protective actions appear to be more plastic than startling: for example, the eye blink reflex to corneal air puff[49] can become associated with predictive cues through learning.

Other threat-related consummatory responses include the suite of behaviours often labelled as fight, flight, and freeze responses[50]. 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'[51,52] or 'predatory imminence'[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[54] and inhibition of goal-oriented behaviour[55]. These responses can co-occur with feelings of surprise in humans[55]. 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[56], courtship in stickle-backs[57], egg-moving in geese[57] and potentially elementary eating actions in wild gorillas[58]. 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.

[H3] *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 model-free manner. Model-based predictions of forthcoming outcomes support specific forms of preparation; this could underly particular bodily responses such as the conditioned protective eyeblink[49] or limb withdrawal[59]. 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 appear 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[40]. Both model-based and model-free predictions could determine a unified opportunity cost of sloth[32].

Fear responses provide well-known examples of behaviours that are subject to a preparatory controller. These responses include Pavlovian actions that enable the subject to prepare for specific threats[50] and which 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 may arise from multiple separate neural controllers[60]. Precursors of threat are often learned through experience, thus requiring plasticity. This is apparent in cue-conditioned[61] and context-conditioned freezing[62]. Such learning occurs for various sensory stimuli across different modalities, although some stimulus-outcome combinations are apparently more readily learned than others[45,63] 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[33] and the 'escape from fear' paradigm, which involves the de-novo acquisition of actions that avert predicted threat[47]. 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[47].

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

[H3] 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[52]. In the latter case, the dedicated action pattern that is adopted to resolve such conflict has been termed 'anxiety-like'[8] and includes passive avoidance (that is, a complete lack of approach). In exploration or foraging paradigms, such avoidance gradually disappears over time[67]. A related response in humans is anxiety-like behavioural inhibition, which has been suggested to be partly under instrumental and possibly model-based control[37,68].

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 individual agents achieve common goals. It has been proposed to translate such approach to neuroscience, in our case by regarding algorithms as individual agents[69].

[H3] 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 appears to be Pavlovian but model-based[70]: that is, it is associated with specific outcomes, but does not consider whether the desired outcomes are actually achieved. Furthermore, there appear to be several distinct algorithms in control of these behaviours, characterised by further pre-progamming 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[14]. Instead, they resonate to a degree with theories that posit the existence of sets of distinct emotions[1,71,72]. On the other hand, the distinct algorithms highlighted here do not map onto the classical emotion categories proposed by basic emotion theory[7,73] 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 need not 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).

Since 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 unlikely 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 based just on behavioural evidence. Furthermore, organisms that occupy separate ecological niches may also employ very distinct controllers.

[H1] Neural circuits for survival

Armed with this basic architecture of control, we now turn 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/dimensional duality is also evident in the neural systems that mediate these control algorithms.

[H3] 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/flight and for different kinds of freezing behaviour may be anatomically closely related in subdivisions of the periaqueductal gray[74,75] 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[25].

In favour of macroscopically separated controllers, circumscribed brain lesions can have a profound and specific impact on emotional behaviour. For one example, amygdala lesions

impair what is termed cue-conditioned freezing[76], but appear leave intact some innate anxiety-like behaviour in rodents[77]. The latter are reduced by hippocampal lesions[77,78] which do not impact cue-conditioned freezing[76]. 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), ma require y partly separate and independent neural systems[60].

In addition, different Pavlovian actions appear 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[79,80], although the loci that relate to each type of action vary according to the familiarity of the context[81]. 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.

[H3] Distributed neural controllers

Despite the evidence outlined above, we believe that it is likely inaccurate to conceive of 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 to 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 elicit an appropriate response to predictors (as in Pavlovian fear conditioning) can be abstractly described by a single decision-making algorithm. However, it appears that considerable array of brain regions is involved[10]. This could include computation of evidence for threat in the amygdala, and 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[82]), and the additional involvement of sensory cortices for predictors with particular sensory properties[83,84].

[H3] 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[85-88], and shared representation of diverse pleasures[26]; electrophysiological recordings show encoding of global utility in the orbitofrontal cortex[89], and of reward prediction errors across various stimuli in phasic dopaminergic responses[90]. 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[64,91-95]. Furthermore, tonic dopaminergic responses appear to reflect average reward[30,32]. 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.

[H3] Arbitration between controllers

The critical remaining implementational question concerns the neural basis of arbitration and interaction amongst 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 over-exuberant activity in the serotonergic raphe that is caused by repeated negative outcomes and drives helplessness is apparently suppressed via the medial pre-frontal cortex in those subjects that are able to exert control[96]. Thus one controller which helps mediate passivity and behavioural inhibition (the raphe) is suppressed by another (the medial pre-frontal cortex). Indeed, neuromodulators and neuropeptides[97] could provide a convenient way to communicate dimensional quantities such as utility or arousal globally, in keeping with widespread dopamine[98], serotonin[99], and norepinephrine[100] projections. Circulating hormones, for example in the case of stress hormones[101], could spread even broader influences over even longer timescales.

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

[H1] 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,[4,107] and so our hope is just to show how they might fit into the current picture. Importantly, although they are regarded by some as being critical 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[1,12,108]. This raises two central questions: what are feelings, and what, if anything, is 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[18]. 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 that our subjective sense and its verbalisation hews substantially to the dimensions of valence and arousal[2,20].

Various data suggest that experienced (even incidental) feelings influence future decisions, as well as immediate actions[109]. 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[36,110,111]. Secondly, although decisions are shaped by currently experienced feelings, they are also influenced by the feelings anticipated to occur after relevant outcomes[112].

Hence, feelings experienced in the past may provide sparse and efficient signals for future deliberation of decision outcomes, and thus simplify model-based search[113] and/or memory look-up[114]. Such anticipated feelings may be rather abstract or may induce actual feelings[115]. Since feelings are only incompletely able to represent the full workings of the various controllers, their influence may appear suboptimal or irrational (just as we argued for model-free controllers). Finally, 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 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[116,117].

[H1] 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: that emotional phenomena are the output of a system for response optimization[14], just like any other behaviour. We therefore analysed the goals of behaviour in biological environments, dissected emotions into associated actions and feelings, and characterised 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, and 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 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[7,73]), or propose 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[18,118]). Many computational, psychological and neural questions remain, and we hope to have furnished a useful framework for answering them.

Box 1: Levels of theoretical analysis

In computational neuroscience, it is common to distinguish different levels of analysis that go back to Marr[21].

Computational level

At the computational level[21], theoretical analysis focuses on formalising the problem that the nervous system has to solve, and finding an appropriate, often optimal or normative solution. One optimal solution to any decision-making problem is given by Bayesian decision theory (BDT[119]). According to this theory, agents should create and maintain a so-called 'belief state' which summarises the whole history of their past observations. To do so, they must employ what is known as a generative model of 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 [119-121]. 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 their subjects.

Implementational level

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

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 they exactly or approximately enforce; they can also be implemented in neural tissue or in other substrates.

Pavlovian vs. 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 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[122] or rooting with objects associated with food in pigs[123]) are pitted against experimentally-determined contingencies (such as denying or delaying rewards that are approached in this way).

Model-based vs. model-free control

At least two canonical methods have been described for making predictions when whole trajectories of future states and possibly actions must be considered. Model-based reasoning[120] involves building a precise set of beliefs about the structure of the environment and the outcomes it affords and searching the model prospectively at the time of choice through a form of dynamic programming[120]. This has some attractive properties: for example, models are often relatively easy to learn and choice 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[113]. Thus simplification is essential. One simplification is a more general form of model-based control algorithm[70], 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: here, 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, namely the summed expected utility of the future outcomes[120]. Given the way 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 alter.[124] 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[125-128], influenced by particular characteristics such as the relative uncertainties of the two sorts of controller[17], or the cost versus benefit of engaging in expensive model-based calculations to overcome potentially incorrect model-free ones[129,130].

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[131]. However, such a characterization is not well supported by the evidence and interpretations that we have adduced.

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[132]. 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 (though 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, irrational actions such as fighting) in order 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 as to where such long-run battles might end up, helping address the important questions as to how individual-level emotional characteristics maximise population-level fitness[133,134] and how heterogeneity across a population might result[135].

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 amongst us are able to access much richer, covert, physiological and neural signals about our own states, including a machinery for interoceptive inference[136]. 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[113]) such that we could be less biased observers of others than of ourselves.

Glossary

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

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

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

Constructionist approach: A family of theoretical approaches that views 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: Assume A is preferred over B, and 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 BDT 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, we use the term Pavlovian to denote an algorithm or controller whose choice of actions is insensitive to the actual consequences of those actions. We do not use the term to denote design characteristics of experiments (as is sometimes the case).

Instrumental: The term instrumental in this article refers to an algorithm or controller whose choices are contingent on their past or predicted future consequences. We do not refer here to design characteristics of experiments.

Model-based: We use the term model-based to characterise algorithms that exploit a model of the structure of the environment and the outcomes it affords to make long-run

predictions about the future. Predictions need not be action-contingent, and so can support either Pavlovian or instrumental controllers.

Model-free: We use the term model-free 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 so do not encode the specific outcomes underpinning those values. Model-free predictions need not be action-contingent, and so 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, etc.) are the output of a set of cognitive appraisals, or encompass such appraisals. Theories differ widely according to the appraisals they consider part of the set.

Author biographies:

Dominik R Bach

Assistant professor for Clinical Psychiatry Research, University of Zurich

Dominik R. Bach studied medicine, psychology and maths. He obtained a PhD in experimental psychology at Berlin Institute of Technology and trained as psychiatrist. After postdocs at UCL and Berlin School of Mind and Brain, he is now at the Psychiatric Hospital, University of Zurich, Switzerland. He is an honorary fellow at the Max Planck UCL Centre for Computational Psychiatry and Wellcome Trust Centre for Neuroimaging (UCL). His research focuses on the computational neuroscience of defensive emotions, particularly in the context of mental health.

Authors's website: www.bachlab.org

Peter Dayan

Professor of Computational Neuroscience, University College London

Peter Dayan studied mathematics at Cambridge University, did his PhD in computational neuroscience at the University of Edinburgh, and did postdocs at the Salk Institute and the University of Toronto. After three years as an assistant professor at MIT, he helped found the Gatsby Computational Neuroscience Unit at UCL in 1998. His interests center on mathematical and computational models of neural processing, with a particular emphasis on representation, learning and decision making.

Peter Dayan's homepage: http://www.gatsby.ucl.ac.uk/~dayan

References

1 Darwin, C. *The expression of the emotions in man and animals*. Vol. 1965 (University of Chicago Press, 1872).

2 Russell, J. A. A Circumplex Model of Affect. *J Pers Soc Psychol* **39**, 1161-1178, doi:DOI 10.1037/h0077714 (1980).

3 Kuppens, P., Tuerlinckx, F., Russell, J. A. & Barrett, L. F. The relation between valence and arousal in subjective experience. *Psychol Bull* **139**, 917-940, doi:10.1037/a0030811 (2013).

4 Scherer, K. R., Schorr, A. & Johnstone, T. *Appraisal processes in emotion: Theory, methods, research*. (Oxford University Press, 2001).

5 Oatley, K. & Johnson-Laird, P. N. Cognitive approaches to emotions. *Trends in cognitive sciences* **18**, 134-140, doi:10.1016/j.tics.2013.12.004 (2014).

6 Ekman, P. & Oster, H. Facial Expressions of Emotion. *Annual Review of Psychology* **30**, 527-554, doi:Doi 10.1146/Annurev.Ps.30.020179.002523 (1979).

7 Izard, C. E. Basic emotions, relations among emotions, and emotion-cognition relations. *Psychological review* **99**, 561-565 (1992).

8 Calhoon, G. G. & Tye, K. M. Resolving the neural circuits of anxiety. *Nat Neurosci* **18**, 1394-1404, doi:10.1038/nn.4101 (2015).

9 Burgdorf, J. & Panksepp, J. The neurobiology of positive emotions. *Neuroscience and biobehavioral reviews* **30**, 173-187 (2006).

10 Herry, C. & Johansen, J. P. Encoding of fear learning and memory in distributed neuronal circuits. *Nat Neurosci* **17**, 1644-1654, doi:10.1038/nn.3869 (2014).

11 Stephan, K. E. *et al.* Charting the landscape of priority problems in psychiatry, part 1: classification and diagnosis. *Lancet Psychiatry* **3**, 77-83, doi:10.1016/S2215-0366(15)00361-2 (2016).

12 LeDoux, J. E. Coming to terms with fear. *Proc.Natl.Acad.Sci.U.S.A* **111**, 2871-2878, doi:10.1073/pnas.1400335111 (2014).

13 Blanchard, D. C., Griebel, G., Pobbe, R. & Blanchard, R. J. Risk assessment as an evolved threat detection and analysis process. *Neuroscience and biobehavioral reviews* **35**, 991-998, doi:Doi 10.1016/J.Neubiorev.2010.10.016 (2011).

14 Scherer, K. R. in *Appraisal processes in emotion* (eds K. R. Scherer, A. Schorr, & T. Johnstone) 92-120 (Oxford University Press, 2001).

15 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, doi:10.1523/JNEUROSCI.4316-06.2006 (2006).

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

17 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).

18 Lindquist, K. A. & Barrett, L. F. A functional architecture of the human brain: emerging insights from the science of emotion. *Trends in cognitive sciences* **16**, 533-540, doi:10.1016/j.tics.2012.09.005 (2012).

19 Jack, R. E., Garrod, O. G. & Schyns, P. G. Dynamic facial expressions of emotion transmit an evolving hierarchy of signals over time. *Current biology : CB* **24**, 187-192, doi:10.1016/j.cub.2013.11.064 (2014).

20 Nowlis, H. H. & Nowlis, V. The description and analysis of mood. *Annals of the New York Academy of Sciences* **65**, 345-355 (1956).
21 Marr, D. C. & Poggio, T. From Understanding Computation to Understanding Neural Circuitry. *Neurosciences Research Program Bulletin* **15**, 470-491 (1977).
22 Kording, K. P. & Wolpert, D. M. Bayesian decision theory in sensorimotor control. *Trends in cognitive sciences* **10**, 319-326 (2006).
23 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, doi:10.1038/nrn3289 (2012).
24 Pouget, A., Beck, J. M., Ma, W. J. & Latham, P. E. Probabilistic brains: knowns and unknowns. *Nat Neurosci* **16**, 1170-1178, doi:10.1038/nn.3495 (2013).
25 Graebner, A. K., Iyer, M. & Carter, M. E. Understanding how discrete populations of hypothalamic neurons orchestrate complicated behavioral states. *Frontiers in systems neuroscience* **9**, 111, doi:10.3389/fnsys.2015.00111 (2015).
26 Berridge, K. C. & Kringelbach, M. L. Pleasure systems in the brain. *Neuron* **86**, 646-664, doi:10.1016/j.neuron.2015.02.018 (2015).
27 Samuelson, P. A Note on the Pure Theory of Consumers' Behaviour. *Economica* **5**, 61-71 (1938).
28 Blondel, V. D. & Tsitsiklis, J. N. A survey of computational complexity results in systems and control. *Automatica* **36**, 1249-1274 (2000).
29 Hinton, G. E. & Nowlan, S. J. How learning can guide evolution. *Complex systems* **1**, 495-502 (1987).
30 Niv, Y., Daw, N. D., Joel, D. & Dayan, P. Tonic dopamine: opportunity costs and the control of response vigor. *Psychopharmacology* **191**, 507-520, doi:10.1007/s00213-006-0502-4 (2007).
31 Choi, J. E., Vaswani, P. A. & Shadmehr, R. Vigor of movements and the cost of time in decision making. *J Neurosci* **34**, 1212-1223, doi:10.1523/JNEUROSCI.2798-13.2014 (2014).
32 Dayan, P. Instrumental vigour in punishment and reward. *Eur J Neurosci* **35**, 1152-1168, doi:Doi 10.1111/J.1460-9568.2012.08026.X (2012).
33 Mowrer, O. H. 2-Factor Learning Theory - Summary and Comment. *Psychological review* **58**, 350-354, doi:DOI 10.1037/h0058956 (1951).
34 Maia, T. V. Two-factor theory, the actor-critic model, and conditioned avoidance. *Learn Behav* **38**, 50-67, doi:10.3758/Lb.38.1.50 (2010).
35 Lloyd, K. & Dayan, P. Safety out of control: dopamine and defence. *Behav Brain Funct* **12**, 15, doi:10.1186/s12993-016-0099-7 (2016).
36 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, doi:10.3390/bs3030501 (2013).
37 Bach, D. R. Anxiety-Like Behavioural Inhibition Is Normative under Environmental Threat-Reward Correlations. *PLoS computational biology* **11**, e1004646, doi:10.1371/journal.pcbi.1004646 (2015).
38 Mackintosh, N. J. *Conditioning and associative learning*. (Oxford University 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 H. Pashler & R. Gallistel) 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. Actions and habits: Variations in associative representations during instrumental learning. *Information processing in animals: Memory mechanisms*, 143-165 (1981).
43 Yeomans, J. S., Li, L., Scott, B. W. & Frankland, P. W. Tactile, acoustic and vestibular systems sum to elicit the startle reflex. *Neuroscience and biobehavioral reviews* **26**, 1-11 (2002).
44 Wilson, R. C. & Niv, Y. Inferring relevance in a changing world. *Frontiers in human neuroscience* **5**, 189, doi:10.3389/fnhum.2011.00189 (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, doi:10.1037/0097-7403.33.4.451 (2007).
48 Bach, D. R. A cost minimisation and Bayesian inference model predicts startle reflex modulation across species. *J Theor Biol* **370**, 53-60, doi:10.1016/j.jtbi.2015.01.031 (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. *Annual Review of Psychology* **39**, 43-68 (1988).
51 McNaughton, N. & Corr, P. J. A two-dimensional neuropsychology of defense: fear/anxiety and defensive distance. *Neuroscience and biobehavioral reviews* **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 University Press, 2000).
53 Fanselow, M. S. & Lester, L. S. A functional behavioristic approach to aversively motivated behavior: Predatory imminence as a determinant of the topography of defensive behavior. (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. *Behaviour. Supplement*, III-159 (1967).
57 Tinbergen, N. *The study of instinct*. (Oxford University Press, 1951).
58 Byrne, R. W. & Byrne, J. M. E. Complex leaf-gathering skills of mountain gorillas (Gorilla g. beringei): variability and standardization. *American Journal of Primatology* **31**, 241-261 (1993).

59 Zhang, S., Mano, H., Ganesh, G., Robbins, T. & Seymour, B. Dissociable Learning Processes Underlie Human Pain Conditioning. *Current biology : CB* **26**, 52-58, doi:10.1016/j.cub.2015.10.066 (2016).
60 Gross, C. T. & Canteras, N. S. The many paths to fear. *Nat Rev Neurosci* **13**, 651-658, doi:10.1038/nrn3301 (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. *Annals of the New York Academy of Sciences* (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. *Psychological review* **108**, 483-522 (2001).
64 Dolan, R. J. & Dayan, P. Goals and habits in the brain. *Neuron* **80**, 312-325, doi:10.1016/j.neuron.2013.09.007 (2013).
65 Korn, C. W. & Bach, D. R. Maintaining homeostasis by decision-making. *PLoS computational biology* **11**, e1004301, doi:10.1371/journal.pcbi.1004301 (2015).
66 Rangel, A. Regulation of dietary choice by the decision-making circuitry. *Nat Neurosci* **16**, 1717-1724, doi:10.1038/nn.3561 (2013).
67 Fonio, E., Benjamini, Y. & Golani, I. Freedom of movement and the stability of its unfolding in free exploration of mice. *Proc.Natl.Acad.Sci.U.S.A* **106**, 21335-21340, doi:10.1073/pnas.0812513106 (2009).
68 Bach, D. R. The cognitive architecture of anxiety-like behavioral inhibition. *J Exp Psychol Hum Percept Perform* **43**, 18-29, doi:10.1037/xhp0000282 (2017).
69 Alonso, R., Brocas, I. & Carrillo, J. D. Resource Allocation in the Brain. *The Review of Economic Studies* **81**, 501-534, doi:10.1093/restud/rdt043 (2014).
70 Dayan, P. & Berridge, K. C. Model-based and model-free Pavlovian reward learning: revaluation, revision, and revelation. *Cogn Affect Behav Neurosci* **14**, 473-492, doi:10.3758/s13415-014-0277-8 (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, doi:10.2466/pms.1964.18.1.119 (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? *Psychological review* **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, doi:10.1016/j.bbr.2007.10.018 (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 peraqueductal 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. & Bannerman, 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. *Current Biology* **24**, 541-547, doi:Doi 10.1016/J.Cub.2014.01.046 (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, doi:20026734 (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, doi:10.1038/nn2061 (2008).
82 Sharpe, M. J. & Killcross, S. The prelimbic cortex uses higher-order cues to modulate both the acquisition and expression of conditioned fear. *Frontiers in systems neuroscience* **8**, 235, doi:10.3389/fnsys.2014.00235 (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, doi:10.1038/nn.3749 (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, doi:10.1093/cercor/bhv001 (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 corticostriatolimbic circuits. *Physiol Behav* **86**, 717-730, doi:10.1016/j.physbeh.2005.08.061 (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, doi:10.1523/JNEUROSCI.2711-11.2011 (2011).
95 Liljeholm, M. & O'Doherty, J. P. Contributions of the striatum to learning, motivation, and performance: an associative account. *Trends in cognitive sciences* **16**, 467-475, doi:10.1016/j.tics.2012.07.007 (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, doi:10.1523/JNEUROSCI.3630-06.2006 (2006).
97 Dayan, P. Twenty-five lessons from computational neuromodulation. *Neuron* **76**, 240-256, doi:10.1016/j.neuron.2012.09.027 (2012).
98 Fallon, J. H. Topographic organization of ascending dopaminergic projections. *Annals of the New York Academy of Sciences* **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, doi:10.1007/s00213-010-2089-z (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, doi:10.1523/JNEUROSCI.5754-12.2013 (2013).
103 Tovote, P., Fadok, J. P. & Luthi, A. Neuronal circuits for fear and anxiety. *Nat Rev Neurosci* **16**, 317-331, doi:10.1038/nrn3945 (2015).
104 Likhtik, E., Stujenske, J. M., M, A. T., Harris, A. Z. & Gordon, J. A. Prefrontal entrainment of amygdala activity signals safety in learned fear and innate anxiety. *Nat Neurosci* **17**, 106-113, doi:10.1038/nn.3582 (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, doi:10.1016/j.neuron.2009.12.002 (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, doi:10.1016/j.neuron.2011.07.027 (2011).
107 Roseman, I. J. & Smith, C. A. in *Appraisal processes in emotion* (eds K. R. Scherer, A. Schorr, & T. Johnstone) Ch. 1, 3-19 (Oxford University 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, doi:10.1126/science.1076358 (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, doi:10.1098/rspb.2010.0303 (2010).
111 Eldar, E., Rutledge, R. B., Dolan, R. J. & Niv, Y. Mood as Representation of Momentum. *Trends in cognitive sciences* **20**, 15-24, doi:10.1016/j.tics.2015.07.010 (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* (2015).
113 Huys, Q. J. *et al.* Bonsai trees in your head: how the pavlovian system sculpts goal-directed choices by pruning decision trees. *PLoS computational biology* **8**, e1002410, doi:10.1371/journal.pcbi.1002410 (2012).
114 Redish, A. D. Vicarious trial and error. *Nat Rev Neurosci* **17**, 147-159, doi:10.1038/nrn.2015.30 (2016).
115 Loewenstein, G. & Lerner, J. S. Handbook of affective sciences (Eds. RJ Davidson, KR Scherer, HH Goldsmith), Oxford 2003: Oxford University Press.
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**, 773-783 (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*, doi:10.1093/scan/nsw154 (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, MA, 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, doi:DOI 10.1037/h0040090 (1961).
124 Dickinson, A. & Balleine, B. Motivational Control of Goal-Directed Action. *Anim Learn Behav* **22**, 1-18, doi:Doi 10.3758/Bf03199951 (1994).
125 Boureau, Y. L., Sokol-Hessner, P. & Daw, N. D. Deciding How To Decide: Self-Control and Meta-Decision Making. *Trends in cognitive sciences* **19**, 700-710, doi:10.1016/j.tics.2015.08.013 (2015).
126 Dayan, P. How to set the switches on this thing. *Current opinion in neurobiology* **22**, 1068-1074, doi:10.1016/j.conb.2012.05.011 (2012).
127 Alexander, W. H. & Brown, J. W. Computational models of performance monitoring and cognitive control. *Topics in cognitive science* **2**, 658-677, doi:10.1111/j.1756-8765.2010.01085.x (2010).
128 Botvinick, M. & Weinstein, A. Model-based hierarchical reinforcement learning and human action control. *Philosophical transactions of the Royal Society of London. Series B, Biological sciences* **369**, doi:10.1098/rstb.2013.0480 (2014).
129 Keramati, M., Dezfouli, A. & Piray, P. Speed/accuracy trade-off between the habitual and the goal-directed processes. *PLoS computational biology* **7**, e1002055, doi:10.1371/journal.pcbi.1002055 (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, doi:10.3389/fpsyg.2013.00092 (2013).

131 Kahneman, D. *Thinking, fast and slow*. (Farrar, Straus and Giroux, 2011).
132 Camerer, C. *Behavioral game theory: Experiments in strategic interaction*. (Princeton University Press, 2003).
133 Devaine, M., Hollard, G. & Daunizeau, J. Theory of mind: did evolution fool us? *PloS one* **9**, e87619, doi:10.1371/journal.pone.0087619 (2014).
134 Johnson, D. D. & Fowler, J. H. The evolution of overconfidence. *Nature* **477**, 317-320, doi:10.1038/nature10384 (2011).
135 Giske, J. *et al.* Effects of the emotion system on adaptive behavior. *Am Nat* **182**, 689-703, doi:10.1086/673533 (2013).
136 Seth, A. K. Interoceptive inference, emotion, and the embodied self. *Trends in cognitive sciences* **17**, 565-573, doi:10.1016/j.tics.2013.09.007 (2013).

Competing interest

The authors declare no conflict of interest.

Acknowledgements

We thank Giuseppe Castegnetti, Lasana Harris, Christoph Korn, Mike Mendl, Hiro Nakahara, Liz Paul, our reviewers, and many others, for inspiring discussions during the writing of this article. This work was supported by the University of Zurich (DRB) and the Gatsby Charitable Foundation (PD), and a grant from the UK National Centre for the Replacement Refinement and Reduction of Animals in Research (K/00008X/1, to Mike Mendl, Elizabeth Paul and PD). The Wellcome Trust Centre for Neuroimaging is supported by core funding from the Wellcome Trust (091593/Z/10/Z).