Where does value come from?

Keno Juechems* and Christopher Summerfield

Department of Experimental Psychology
University of Oxford
Radcliffe Observatory Quarter
Woodstock Road, Oxford, UK

Abstract

The computational framework of reinforcement learning (RL) has allowed us to both understand biological brains and build successful artificial agents. However, in this article we highlight open challenges for RL as a model of animal behaviour in natural environments. We ask how the external reward function is designed for biological systems, and how we can account for the context sensitivity of valuation. We summarise both old and new theories proposing that animals track current and desired internal states and seek to minimise the distance to goal across multiple value dimensions. We suggest that this framework can readily account for canonical phenomena observed in the fields of psychology, behavioural ecology, and economics, and recent findings from brain imaging studies of value-guided decision-making.

*Correspondence: keno.juchems@psy.ox.ac.uk

Keywords

Reinforcement learning; value; homeostasis; reward; goal-directed decision-making; medial prefrontal cortex

The reward hypothesis and the reward paradox

What is it that animals seek to achieve by their behaviour? For evolutionary biologists, there is a simple answer to this question – that animal behaviour has evolved to maximise reproductive fitness, i.e. the propensity to produce offspring that will carry forth our genes. However, for those studying the behaviour of humans and animals in the fields of ethology, economics, psychology and neuroscience, this answer is only partly satisfying. Critically, it stops short of specifying how animals can learn which behaviours will increase or decrease their fitness. Instead, behavioural scientists propose that the environment furnishes reinforcement signals that indicate the likely costs or benefits of an action and assume that these signals can be directly sensed by the agent, allowing them to behave in ways that maximise utility over the short or long term. In machine learning research, this is sometimes called the "reward hypothesis" [1].

Over the past half century, the fields psychology and artificial intelligence (AI) have jointly developed a normative theory, known as reinforcement learning (RL), that is founded on the reward hypothesis [2]. RL conceives of each encounter with the world as involving three distinct stages: (i) an agent receives sensory observations from the environment; (ii) the agent takes actions that influence its future states; and (iii) the agent receives a scalar reward signal that is emitted by the environment, and that is processed by the agent via a dedicated input channel (**Fig. 1a**). This conceptualisation lends itself naturally to modelling the canonical laboratory paradigm for operant behaviour, whereby an animal receives a sensory stimulus (e.g. a visual signal or tone), takes an action (e.g. presses a button or licks a spout) and receives a positive or negative reinforcer (e.g. liquid, food, or money). The framework successfully accounts for a wide array of habit-based and goal-directed behaviours in humans and other animals [3], and prominently explains the neural signals that accompany the expectation and delivery of reinforcers [4]. RL has furnished the canonical computational framework for understanding value-guided decision-making in humans and other primates, and has been widely used to explain how different brain regions participate in valuation and choice [2]. Moreover, RL models are currently offering some of the most exciting advances in AI research, producing artificial agents that can behave intelligently in complex, dynamic environments such as board and video games, thereby providing a proof of concept that intelligent behaviour can emerge *de novo* under the assumptions of the reward hypothesis [5, 6].

The RL framework offers an important set of computational tools for understanding animal learning and behaviour. However, as a theory of biological intelligence, RL focusses on how rewards influence behaviour but does not specify the provenance of the external reward function against which behaviour is optimised. In machine learning, the reward function for an RL model is hand-designed by the researcher. For example, when training an autonomous vehicle, she might specify that it needs to reach its destination as quickly as possible whilst obeying the traffic laws, by assigning benefits and costs for desired and undesired behaviours. However, in natural environments, no external entity exists that can directly quantify the consequences of each action, like the points that are awarded in a video game for completing levels or shooting monsters (**Fig. 1b**). Nor is it obvious that biological systems have a dedicated channel for receipt of external rewards that is distinct from the classical senses. Rather, rewards and punishments *are* sensory observations – the taste of an apple, the warmth of an embrace – and so stimulus value must be inferred by the agent, not conferred by the world. In other words, rewards must be intrinsic, not extrinsic (**Box 1**). We call this "the reward paradox".

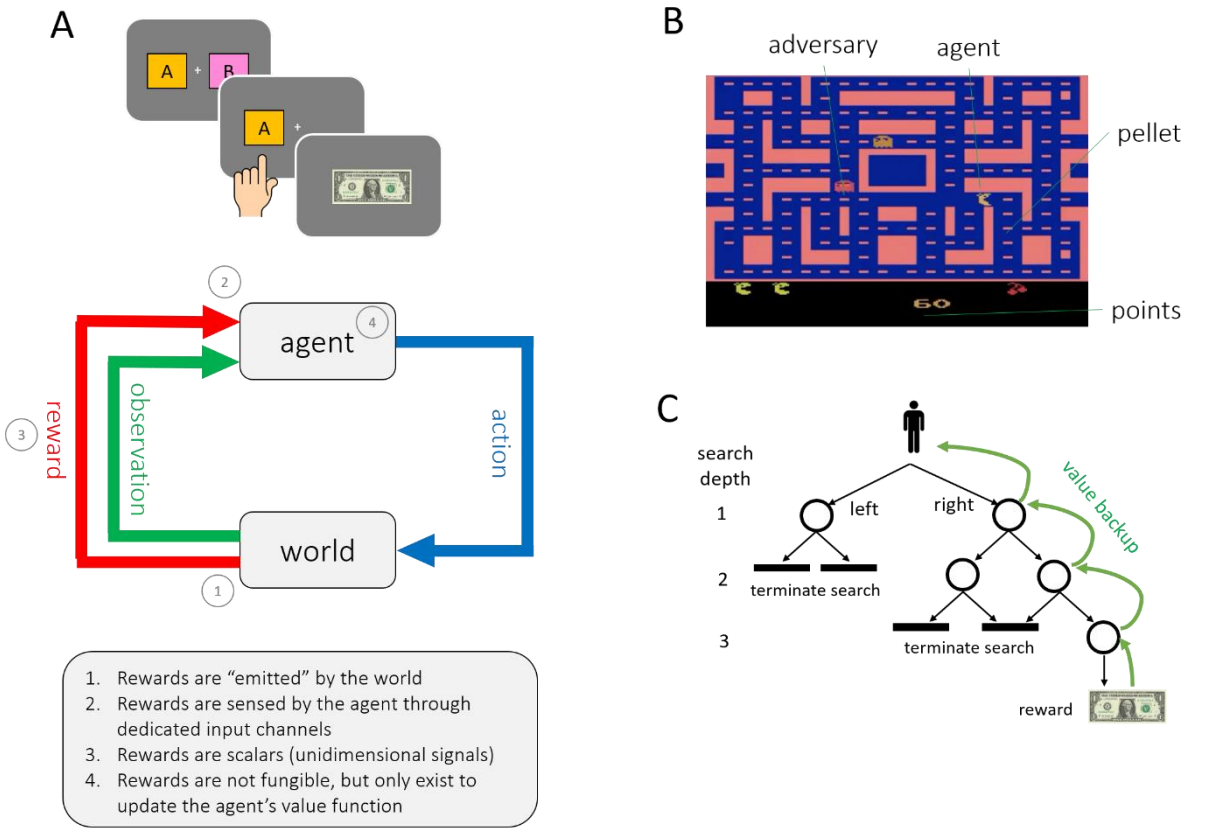

Figure 1. The RL framework A. Top: a typical "bandit" task in which participants choose between options (A and B) with uncertain reward probability, incurring financial outcomes. Below: in the RL framework, an agent receives observations (green), takes an action (blue) and receives feedback in the form of reward (red). The numbers highlight some assumptions of the basic RL framework that are less well suited to explaining motivated behaviour in biological organisms (grey box below). **B.** The RL framework applied to video games (here, Ms. Pac-Man). The agent acts in order to maximise game score, which is a scalar quantity. Unlike humans playing the game, the agent receives the score differential directly (e.g. points for eating pellets) as reward signal via a dedicated input channel. **C**. Schematic of model-based RL, where the agent searches through a tree of possible future states and actions, sometimes stopping if no reward is expected from this branch of the search tree, and ultimately choosing the trajectory through the tree that leads to the highest return.

In this *Opinion*, we consider the challenges of applying the reward hypothesis to the study of biological agents in the natural world, with a focus on the neural and computational mechanisms by which humans and other animals make value-guided choices. In doing so, we highlight some ways in which the RL framework might be modified, adapted or nuanced to account for the reward paradox. We appeal to a diversity of theories concerning motivated behaviour and ask how they can explain the behaviour and brain activity observed during human decisions in cognitive, social and economic settings.

Box 1. Intrinsic Motivation. Machine learning and AI researchers know that designing a reward function by hand is costly and does not scale to real world settings. To avoid this problem, agents have been built that generate their own intrinsic motivational signals [37, 57, 58]. For these agents, the cost function is populated by additional terms that encourage the agent to pursue adaptive behaviours even in the absence of reward. Many of these additional terms draw inspiration from psychological theory, which has long noted that humans and other animals often pursue activities for their own sake (e.g. a kitten furiously chasing a piece of string, or a commuter completing the crossword on her journey to work). For example, one key intrinsic motivation may be to explore the world in a structured fashion, satisfying curiosity and avoiding boredom [59]. This allows agents to seek out experiences that may accelerate learning without directly enhancing reward, a phenomenon known as active learning. Other intrinsic motivation signals may encourage us to make actions with predictable consequences, or that exert maximal control over the environment or the behaviour of conspecifics, or those that minimise discrepant or dissonant inputs (see [57] for a review). Augmenting the reward function with these intrinsic motivational signals often leads to faster learning and more stable patterns of behaviour in artificial agents [60]. However, in most reports, intrinsic motivation signals supplement (rather than replace) rewards from the external environment, leaving the thorny question of where rewards come from only partially answered. Other approaches that avoid any external feedback by using purely unsupervised or self-supervised methods – where the agent is optimised to maximise information gain or reduce sensory surprise – hold undoubted promise but have not yet been observed to yield complex behaviours in dynamic environments that resemble the natural world [61].

Context-dependent valuation and the representation of internal state

To circumvent the reward paradox, it is sometimes assumed that the reward function has been designed by evolution. In other words, reward computation may rely exclusively on fast, mandatory, phylogenetically prespecified mechanisms that automatically convert sensory observations into hedonic signals that act as a proxy for "external" rewards. Indeed, there is no doubt that dedicated mechanisms exist for those physiological drives that lie at the bottom of our "hierarchy of needs" [7], such as hunger or fatigue. These needs are signalled by projections from the hypothalamus to neural systems that respond to rewards, including dopaminergic neurons in the ventral tegmental area [8]. On this basis, proponents of RL models have dubbed dopamine neurons the "reward retina" as if they were responsible for directly sensing reinforcement from the environment [9].

How, then, is the reward value of a stimulus is computed based on the agent's internal state as well as the context provided by the external environment? Intuitively, the value a potentially rewarding stimulus (e.g. a hamburger) will depend on the agent's ongoing bodily needs (am I hungry?). Under the RL framework, this sensitivity to internal context might be accounted for by assuming that bodily states part of the environment, i.e. that the state of "I am hungry" is external to the agent in just the same way that the presence or absence of an apple is a property of the world, not just the mind. However, this argument becomes harder to defend when one considers the range of affective and cognitive factors that can potentially modulate value. For example, a hamburger might be less valuable to someone who is nervous before an exam, or to a committed vegetarian. It seems harder to argue that private mental phenomena such as moods and or moral attitudes are properties of the external environment.

More generally, standard RL models assume that reward is a purely hedonic signal that is used solely to update the agent's value function. Reward is thus not *fungible* – that is, once acquired, it is not exchangeable for other assets, and cannot be used in a way that would alter future observations or facilitate future actions. For example, when RL agents are learning to play video games, the rewards they receive (i.e. game score differential) evaporate instantly without providing future opportunities for disbursement within the game. This is unlike natural settings, where rewards are typically

accumulated because they lead to persistent changes in state. For example, bodies act as repositories for energetic resources that are harvested by the agent during trophic behaviours, and bank accounts are used to store for income acquired in exchange for work. The agent can thus choose to exchange one stored asset for another that is needed, the defining feature of economic behaviour in the marketplace. In the laboratory, rewards may be implicitly accumulated as experimental animals or human participants become more sated or wealthy as a result of the liquid rewards or financial incentives that they receive for performing the task, but these changes in state are typically considered to be nuisance variables by the experimenter. For example, when a monkey working for liquid reward is no longer thirsty enough to work, the researcher will typically terminate the experiment. To fully account for biological behaviour, our models of value learning need to account for the internal state of the organism.

Psychologists and neuroscientists who study appetitive responses to primary reinforcers such as food and liquid have long appreciated that theories motivated behaviour need to include a representation of internal state. One prominent model, known as incentive salience theory [10], proposes that whilst an animal might "like" a stimulus on the basis of cached state values (e.g. sugar tastes good) as proposed in standard RL, the extent to which it "wants" a stimulus also depends on its physiological state (e.g. hunger), which adjusts valuation via a gain control mechanism. This ensures that stimuli that satisfy natural appetites will be consumed more readily (e.g. food when hungry and water when thirsty). Incentive salience can account for a wide range of empirical phenomena. For example, the rate of approach towards food depends on hunger levels, and adapts immediately according to internal levels of satiety irrespective of prior training [11]. Similarly, food items that have been administered during states of food deprivation are preferred over control items during a later test when the animal is sated, suggesting they have acquired greater value [12]. However, incentive salience does not describe exactly how internal state is represented or specify how the animal chooses which appetite it wishes to satisfy. More widely, a rich literature has studied the consequences of satiety on food choices, often invoking model-based RL to describe why animals take actions that elicit a reinforcer that has not been devalued over one that has. However, this literature has largely been silent about the mechanisms by which devaluation itself occurs, focussing instead on how animals learn the causal structure of the task [13].

In the remainder of this article, we explore the relevance of these considerations for the study of human decision-making. Humans routinely make decisions about complex stimuli or abstract plans of action in cognitive, social and economic settings. Our argument is that a representation of internal state may be computationally relevant even for decisions between consumer products, social opportunities, or lifestyle alternatives. We adapt models that have been previously proposed to account for choices among basic needs to these more intricate decision settings and describe how they might account for some intriguing findings concerning the neurobiology of human choice.

Goals as cognitive setpoints

Our argument begins with the observation that much human motivated behaviour is intrinsically goal-directed. Humans will often work tirelessly to complete tasks whose only apparent reward is the satisfaction of task completion itself, for example when seeking to win a race, solve a puzzle, or climb a mountain [14]. Indeed, this purposive, self-consistent property of human behaviour is so powerful that humans will seek to complete tasks even when the outcome is known to be a net loss, a phenomenon known as the sunk cost fallacy [15]. Although the RL framework embraces goal-directed behaviours, it does so under the assumption that animals explicitly learn the transition probabilities between states and use mental tree search to identify the most rewarding trajectories, which then guide "model-based" decisions (see **Fig. 1c**). Thus, for RL, goals are not the objects of computation themselves but emerge as a by-product of policies which optimise expected future reward [16]. Other elaborations of the RL framework explicitly propose additions to the cost function, such as a desire for

novelty, agency or consistency in behaviour, which draw upon the assumption from cognitive psychology that humans do more than simply maximise reward (see **Box 1**).

In considering these phenomena, we turn to one of the earliest theories of motivated behaviour – one which assumes that animals seek to keep internal drives in equilibrium [17, 18]. This theory builds on early models of homeostatic function [19] and found its first prominent expression in the 1940s with Hull's Drive Reduction Theory [20]. These homeostatic models argue that motivated behaviours seek to restore imbalance among different physiological needs, such as warmth, satiety, and hydration (**Fig. 2a**). A central idea is that neural circuits encoded desired states or "setpoints" and drives result from the disparity between current and desired state. Thus, rewards are multidimensional quantities signalling the drive reduction that occurs when an imbalance among the organism's multiple internal needs is redressed.

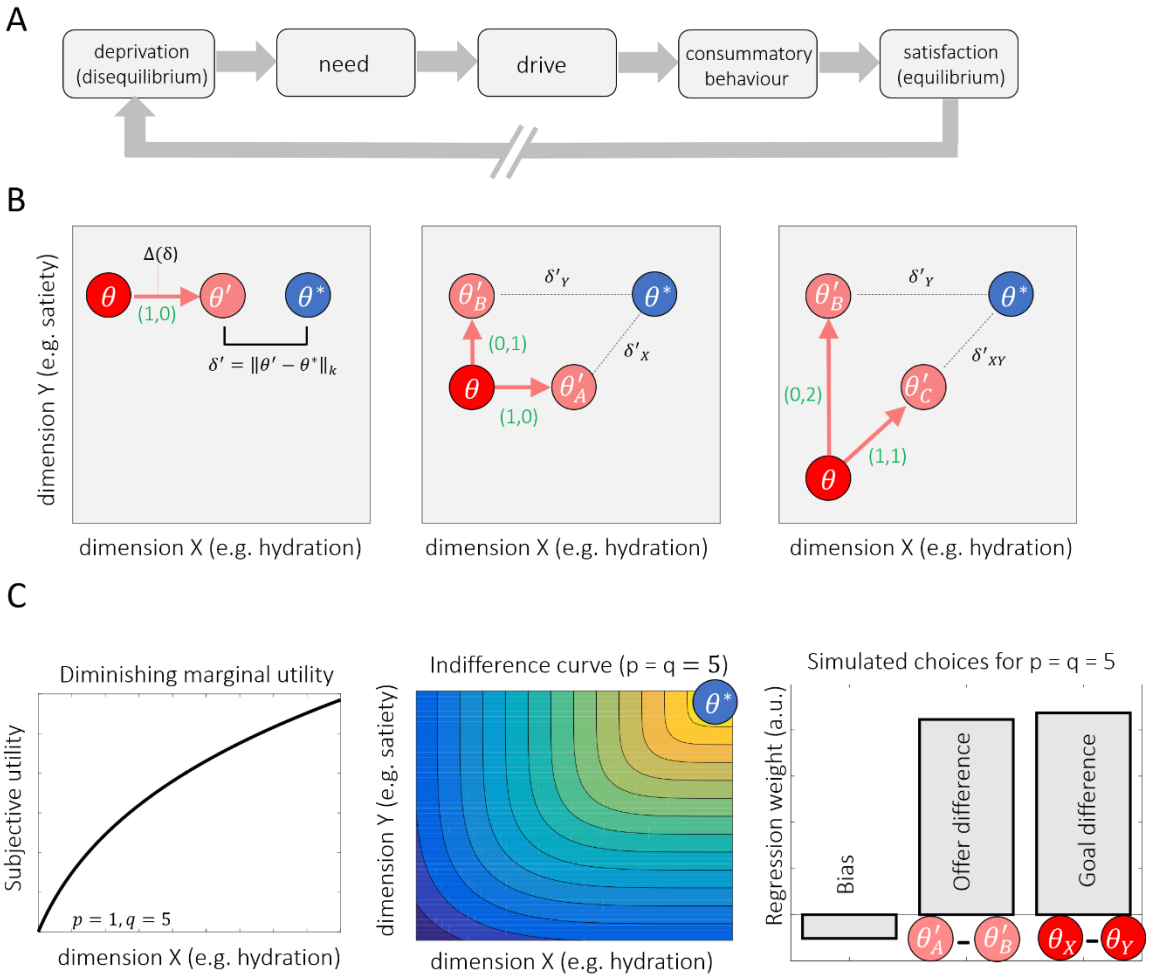

Figure 2. Theories of Motivation. **A.** Schematic of Drive Reduction Theory. Behaviours are initiated to reduce drives, which in turn are provoked by needs that arise following homeostatic imbalance. **B.** Illustration of homeostatically regulated reinforcement learning (HRRL) [18], here in two example dimensions (labelled satiety and hydration). The agent's current state is denoted θ, candidate next states θ' and the desired state (setpoint) θ^*. Offers in the (x,y) dimension are shown in green. In the leftmost panel, the agent is sated but thirsty, as in a laboratory experiment where a monkey is working for liquid. An action reduces the distance δ between the current and desired state $\|\theta - \theta^*\|_k$ (where k indicates the Minkowski distance; we use this here for simplicity, but note that a more general, two-parameter version depicted in panel **C** is needed to account for non-linearities in valuation, see **Box 2** for details). Notice that the rate of reduction in distance is proportional to what would (in the RL framework) be construed as an external reward signal. In the middle panel, an agent who is more thirsty

than hungry is offered an equivalent quantity of food (0,1) or water (1,0). Water reduces distance to goal (Euclidean, $k = 2$) more than food and is thus preferred. In the rightmost panel, the agent is both very hungry and very thirsty. An offer comprising a bundle of both food and water (1,1) is preferred over a large quantity of food (0,2). **C.** Predictions of HRRL with two parameters, p and q, governing non-linearities. The model can reproduce the ubiquitous finding of diminishing marginal utility (left panel). As p and q increase, distance estimates depend more strongly on just one dimension, indicating that maximum weight is given to the currently most needed asset (middle panel). Intermediate values p and q can reproduce one of the central findings of Juechems et al., 2019: that choices are both driven by offers and needs when assets need to be kept in equilibrium (see also **Fig. 4**).

Here, we make a connection between homeostatic models of reward, which have emphasised balance among basic drives and interoceptive processes, and recent work on value-guided decision-making in humans and other primates [21-23]. We argue that motivated behaviour in cognitive, economic or social settings can similarly be understood as seeking to restore a balance among competing "setpoints" that correspond to higher-order goals. Unlike currently popular theories of value-guided choice, which are largely inspired by RL models [2] or expected utility theory from economics [24], our view emphasises the multidimensional nature of value signals, the dependence of motivated behaviour on the internal state of the agent, and the desire to explicitly realise goal states (constraint satisfaction) rather than maximise reward. In what follows, we show how extant theories of motivation can explain some otherwise puzzling behavioural and neural findings in the literature on value-guided decision-making and economic choice.

Homeostatically regulated reinforcement learning for cognitive behaviour

We suggest that during learning, humans form new setpoints pertaining to cognitive goals. For example, we might represent current and desired states on axes pertaining to financial stability, moral worth, or physical health as well as hunger, thirst or temperature. To sketch the computational underpinnings of this process, we appeal to an existing framework for understanding motivated behaviour for basic physiological processes [18, 25, 26], known as homeostatically regulated reinforcement learning (HRRL; **Box 2**). This theory proposes that current states and goals are encoded in a multidimensional "value map". Motivated behaviour can then be seen as an attempt to minimise the maximum distance to setpoints in this value space. Repurposing this framework for cognitive settings, agents commit to policies which focus on purposively driving the current state towards setpoints on a particular goal dimension, such as caching resources, building a shelter, obtaining a mate, or enhancing professional status. In doing so, their ultimate goal is to maintain equilibrium among all goal states, achieving what might be popularly characterised as a state of "wellbeing".

In HRRL, reward emerges naturally as a consequence of the computations required for learning, rather than being furnished by the external environment. This allows the "reward paradox" highlighted above to be sidestepped. To illustrate, consider an animal whose internal state is currently described by parameters θ and who seeks to achieve a desired state θ^*. Let us assume that the agent has a biologically plausible functional form, such as a neural network. In order to calculate the gradients that will allow network weights to be optimised for future behaviour, it is necessary to compute the loss term $\Delta(\|\theta^* - \theta\|_k)$, which indicates the change in distance between a current and desired state that is afforded by any action. Now consider a typical laboratory study in which a single reward dimension is relevant, for example because the experimenter reduces prior liquid intake and offers only water in exchange for behaviour. For a thirsty animal, this loss term is now mathematically identical to what one would construe as "reward" in the RL framework – i.e. the volume of liquid administered on a trial [18]. This equivalence will break down as θ tends towards θ^*, but by this point the researcher will typically have ended the experiment because the animal is fully hydrated. Thus, although in this experiment value is artificially constrained to be unidimensional, we note that from the agent's perspective,

multiple value dimensions may continue to be relevant. For example, a human participant might be tempted to neglect the task in favour of checking their social media account, in order to enhance social capital. However, this individual would most likely be excluded from the analysis.

Homeostatically regulated reinforcement learning readily incorporates other accounts of reference-dependent evaluation by adjusting how the agent computes θ^*. For instance, it may do so by using a running average over past options and anchor its behaviour to this reference. When $k > 1$, simply assuming that other drives exist that cannot currently be satisfied immediately gives rise to the ubiquitous phenomenon of diminishing marginal utility. In other words, as the agent acquires the

Box 2: Homeostatically Regulated Reinforcement Learning

Let us assume that sensory observations, $\boldsymbol{X_t}$ are mapped onto a persistent internal state $\boldsymbol{\theta_t}$:

$$\theta_t = X_t \times U + \theta_{t-1} \times A_\theta$$

Eq.1

Where $\boldsymbol{U}$ is a matrix mapping the chosen option $\boldsymbol{X_t}$ onto $\boldsymbol{\theta_t}$. The internal state, $\boldsymbol{\theta_t}$, exhibits dynamics whereby it will change (decay) in the absence of any input, e.g. credit card debt will increase if it is not repaid, just as thirst will increase if not slaked. These dynamics are captured by the matrix $\boldsymbol{A_\theta}$, which governs the intrinsic changes over time.

Note that this implies that the correspondence between internal and external states must be learned, i.e. our preferences are not simply given to us but acquired with experience, and that we might be uncertain about whether we are hungry, tired or happy. This relates to the framework of "active inference", under which reward values are inferred rather than being directly conferred [17]. We conceive of this internal state space as a multi-dimensional map whose dimensions variously pertain to both physiological states (e.g. satiety) and cognitive goals (e.g. social status). Internal states are subject to momentum via recurrent connections so that states (or moods) may persist in the face of a change in external circumstance.

In HRRL [18], the agent represents a set of desired (or aversive) points $\boldsymbol{\theta}^*$ on this space, allowing it to estimate the distance (and contributing features) of its current position to the goal state. In doing so, the agent needs to combine across the relevant goal dimensions and weight these according to their relative contribution to this distance by computing the Minkowski distance:

$$\delta_{t+1} = \|\theta_{t+1} - \theta^*\|_k = \sqrt[p]{\sum |\theta_{t+1} - \theta^*|^q}, for\ p = q$$

Eq.2

The parameter k allows the agent to prefer bundles that only consist of one single asset ($k < 1$), be indifferent between all bundles ($k = 1$) or prefer mixed bundles ($k > 1$). The model is thus well placed to account for commonly observed indifference curves between two or more economic goods. If we further assume that there are two non-linearities, p and q, rather than just k, our model can further account for diminishing marginal utility, **Fig. 2c**. The implications of this computational step are described more fully in ref [18].

Finally, where, then, is value (or: reward) in this framework? Value arises naturally when the agent compares its current distance to its desired state to its previous distance, which summarizes that the agent has progressed on at least one of its goals since the last time-point:

$$R_{t+1} = \delta_t - \delta_{t+1}$$

Eq.3

Value is thus a summary of whether the agent is approaching or retreating from its goals. This model encapsulates situations in which multi-attribute objects need to be integrated into a single reward signal, such that $R = X \times U$ if $A_\theta = 0$ and k = 1.

currently relevant asset (such as liquid reward) its attention will progressively tend towards other relevant assets, diminishing the contribution of liquid reward to its well-being (Fig. 2c). However, allowing k to vary freely, a range of different policies can be accommodated, including those in which one goal (e.g. attainment of a desired drug in addiction) dominates all others [27].

Implications of HRRL for cognitive decisions

More generally, we argue that some of the most complex and abstract decisions that humans make might be better described by a process that optimises over states, rather than rewards. For example, consider a high school student choosing a career path. Under the (model-based) RL framework, she must consider an impossibly large number of potential futures and select whichever is going to be most rewarding. This seems to imply the devotion of disproportionate levels of computational resources to the search problem. The approach advocated here implies that she first selects a goal state (e.g. become a lawyer) than then takes actions which minimise distance to that goal. For example, she seeks to go to law school; to maximise her chances of acceptance, she first studies hard for her exams; this in turn influences decisions about whether to socialise with friends. This explanation seems to accord better with our common sense intuition of how the complex choices faced by humans are made. However, the computations involved may build upon more phylogenetically ancient mechanisms. For example, one of the most prominent theories of insect navigation proposes that in order to reach their home base, central place foragers such as honey bees (and desert ants) initially encode an egocentric snapshot of their base and subsequently, on the return journey, use a similarity-matching process to gradually reach their goal [28]. This implies they are similarly performing gradient descent over states, akin to the process proposed here.

Our view relates to an earlier proposal that animals are motivated to satisfy a "current concern" [29] and understands disorders of mental health as disrupted goal selection, pursuit and appraisal [30]. A fleshed out computational theory of this process will require the specification of how agents learn the relationship between individual states and the degree of goal realisation. One model proposes a distinction between "primary" and "learned" value, where the former resembles the traditional reward signal in the RL framework and the latter quantifies the eligibility of states for goal realisation [14]. A very appealing aspect of this framework is that it provides a natural way to understand the affective states that pervade our everyday mental landscape, including satisfaction (goal completion), frustration (goal obstruction), and disappointment (goal abandonment), which have largely eluded computational description thus far [14]. A more directly related machine learning approach directly computes the feasibility of attainment of goal states, showing how it can be used for compositional inference under the nonstationary reward functions characteristic of real world environments [31].

Many of the more complex behaviours exhibited by humans involve decision about money. Money has the peculiar property of being fluidly exchangeable for other assets (except, as the Beatles remind us, for love). Our here might seem to break down in financial settings, because it seems incongruous to propose a "setpoint" for wealth – anecdotally at least, the accumulation of wealth does not always blunt the desire to acquire new wealth. However, there is evidence that even wealth accrual may be subject to some of the constraints identified here. For example, taxi drivers on shifts of variable length will tend to work until a fixed return has been achieved, returning home even when this "wealth setpoint" falls during a period of high yield – for example, even if customers are plentiful and fare rates are high [32]. It is also observed that humans use mental accounting to allocate wealth to different potential value dimensions in advance [33]. Thus, although money can be deployed fluidly to facilitate movement in any direction in value space, humans mentally pre-allocate funds to the different directions of advancement, treating an unexpected excess in one direction with profligacy rather than reallocating it facilitate progress along other dimensions. This latter phenomenon is known as the "house money" effect.

Pure drive reduction theories fell from favour with the suggestion that animals will work for rewarding stimuli when their drives are satisfied, i.e. purely for their hedonic value [34, 35]. For example, animals will continue to eat after food has been administered intra-gastrically [34], and will consume stimuli with no nutritional value, such as saccharine water, when they are hungry [36]. Although alternative explanations for these phenomena exist [19], we caveat our argument with the acknowledgement that some behaviours are driven by "liking" alone, irrespective of any relation to the current setpoint. Similarly, it is unlikely that all human cognitive decisions are based on minimising distance to a desired goal – the mechanisms proposed here may exist alongside other processes that allow behaviours that have been pleasurable or successful in the past to be repeated in the future. Thus, we might buy a pair of shoes on impulse, sneak an extra bite of dessert even when we are full, or linger at a party even after our social duties have been fulfilled. However, our argument is that the human tendency to optimise towards desired states has been overlooked in the face of enthusiasm about rewards as the primary drivers of behaviour.

A map of internal state in the OFC

In the final sections, we consider the implications of the perspective advanced here for the neurobiology of decision-making. For valuation to depend on the internal state of the agent, these states must be explicitly represented in brain signals. However, most neural studies of value-guided choice have used a paradigm in which the animal or human participant makes a succession of unrelated, independent choices between food items or monetary gambles. This paradigm is not well suited to measuring how variation in internal state (how sated am I right now? How much wealth have I accumulated?) might be coded in neural circuits. However, recently researchers have begun to employ more complex paradigms that may shed some light on this issue.

At the neural level, the integrity of the orbitofrontal cortex (OFC) is critical for goal-driven reward behaviour [13]. Interestingly, a recent theory proposes that the OFC constitutes a "map" of both observable and latent states within a task [37, 38], similar to the framework proposed here (c.f. Eq.1 in Box 2). Moreover, OFC may code endogenous (internal) as well as exogenous (external) value, as firing rates [39] and BOLD signals [40] can be choice-predictive even in the baseline period prior to stimulus onset. An emerging theory suggests that OFC encodes affective states that pertain to ongoing positive or negative value, i.e. moods or "momentum" in internal value states [41, 42]. Together, these findings imply that OFC is a likely candidate for representing ongoing estimates of position on the value map.

One recent study directly tested whether BOLD signals in OFC tracked the internal accumulation of financial resources across an episode in a way consistent with the ongoing internal state representation proposed here [43]. Participants performed an economic task that involved gambling for financial incentives over successive, independent trials. For a rational agent performing this task, there is no imperative to track the accumulated level of payoff incurred by choices – to maximise return, one simply needs to select, on each trial, the gamble with the highest expected value. However, when humans performed this gambling task over multiple discrete episodes, each signalled by a prominent contextual cue that was structurally irrelevant to the task, brain activity in the medial OFC tracked the level of accumulated resources (or "wealth") over the context (**Fig. 3**). This "value accumulation" signal occurred even though the accumulated rewards were never overtly signalled to participants, nor were they instructed or incentivised to calculate their ongoing wealth.

Another brain imaging study revealed that a nearby region of medial prefrontal cortex tracks internal states during wealth accumulation [44]. Participants were asked to choose among four alternatives (bandits) that either paid or charged an uncertain sum of money. Critically, each selected bandit was fungible – it continued to incur financial costs or benefits across a short block of trials, as if the agent

had purchased that asset and was continually profiting (or otherwise) from its possession ("rule in" condition). In a different condition, the bandits began in the agent's possession and decisions had to be made about which (if any) bandits to "rule out". This allowed us to assess how neural encoding of value depended on whether an asset was being bought or sold. Critically, BOLD signals in the ventromedial prefrontal cortex (vmPFC) encoded the cumulative expected value of assets across the block (**Fig. 3**), and only coded for momentary rewards when decisions were made with future consequences for reward (e.g. when a bandit was selected in the "rule in" condition, or not selected in the "rule out" condition). Together, these studies suggest that the reward system tracks level of accumulated resources over time, for example so that future decision policies can adapt according to the current internal state, in a way that is compatible with the theory advanced here.

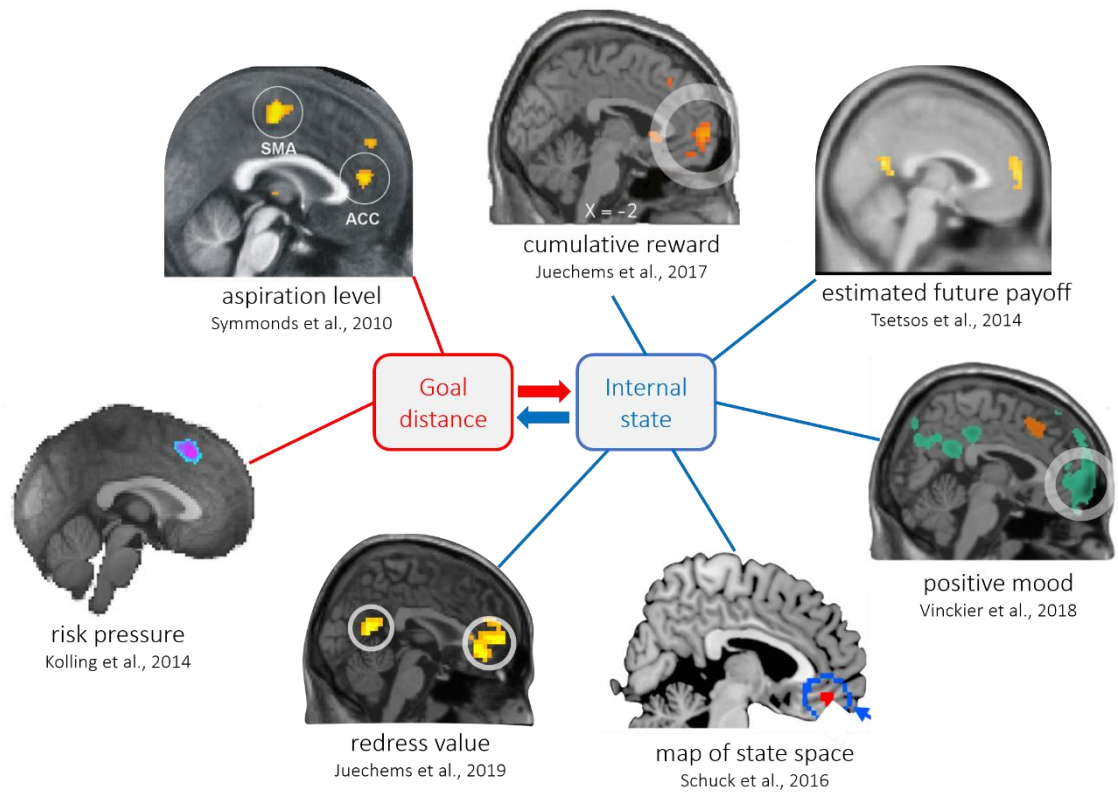

Figure 3. Neural coding of internal state and goal distance. Cortical areas on the medial surface correlating with diverse aspects of internal state and goal distance. Reprinted with permission from refs [37, 42-44, 46, 47, 55].

Budget rules: aspiration and avoidance points

The natural world is structured in such a way that some states are critical for survival or have substantial impact on long-run future outcomes. For example, the student introduced above might work hard to pass her exams in the knowledge that it will open up interesting career opportunities. These states are often attained when accumulated resources reach, or fall below, a critical threshold. Behavioural ecologists have argued that animals' risky foraging behaviour adapts to satisfy a "budget rule" that seeks to maintain energetic resources at aspirational levels that safely offset future scarcity. For example, birds make risky foraging choices at dusk in order to accrue sufficient energy to survive a cold night [45]. This view is neatly accommodated within the framework proposed here, in that the aspiration level reflects the setpoint against which current resource levels are compared, and the driver of behaviour is the disparity between current state and goal.

Recent work with brain imaging [46, 47] suggests that this goal distance signal might be computed in a different medial prefrontal cortex region, the dorsal anterior cingulate cortex (dACC). Symmonds [47] asked participants to make a series of gambles which only paid out if a critical threshold or aspiration point was reached. BOLD signals in the vmPFC scaled with the expected future return over a series of gambles, similar to [44], whereas dACC coded for the aspiration point itself (**Fig. 3**). Using a very similar design, Kolling [46] asked human participants to accept or reject monetary gambles over a short block, with cumulative earnings increased by a multiplier if they reached a fixed aspiration level. BOLD signals in the dACC coded for number of steps to goal, as well as the discrepancy between accumulated return and the aspiration level, scaled by the distance to the end of the block. This signal, which the authors call "risk pressure", seems to be signalling the agent's current position with respect to a goal, rather than merely the arrival of a punctate reward. In another study involving planning in a virtual subway environment, dACC BOLD signals coded for the distance to goal in number of context switches [48]. Indeed, the dACC is one brain region where single-cell responses gradually build up in association with a series of actions that precede reward delivery [49], and where lesions provoke failures of task persistence [50], as if this region participated more generally computing the disparity signal $\theta^* - \theta$ in the service of goal-directed behaviour.

The multidimensional nature of value

In the standard RL framework, each state-action pair is associated with a scalar value, often called a Q-value [1]. In other words, most RL models assume that value is a unidimensional quantity (although "multi-objective" RL models relax this assumption [51]). This seems incongruous when considering the challenges that animals face in natural environments, which involve jointly satisfying many different constraints (for example, maintaining levels of satiety, hydration, warmth and social capital). Rewards take many different forms – from food items to social and sexual signals to secondary reinforcers such as money. A popular view in the hybrid field of neuroeconomics is that neural signals allow potentially incommensurable rewards to be encoded in a single neural "common currency" [23]. For example, single cells in OFC code for the "offer value" of different liquid rewards irrespective of their taste [52]. The existence of a common neural code for different assets allows economists to understand rational decision-making in consumer settings, which might involve choosing between a meal in a restaurant or tickets to the theatre. However, HRRL suggests a different perspective, whereby assets are explicitly encoded along different dimensions (needs) in the value map, and decisions are made according to whichever need is currently greatest.

Without further assumptions, HRRL predicts that animals will treat different assets as if they are imperfect substitutions for one another. Multidimensional prospects (or "bundles") should be preferred if they contain a mixture of different assets, because this allows multiple drives to be satisfied simultaneously (**Fig. 2b**). For example, it is better to be neither too hungry nor too thirsty than to be starving or dehydrated. Several longstanding empirical results support this view. First of all, in early conditioning experiments it was observed that a cue that is associated either simultaneously or successively with multiple assets – such as food and water - has greater value that an equivalent cue associated with only a single asset (or drive) [53]. Finally, a preference for compromise decisions is also apparent from studies of so-called "decoy" effects in multi-attribute, multi-alternative choices. Faced with a choice between two iso-preferred items that have complementary costs and benefits (e.g. an expensive but high quality consumer product A, or a cheaper but lower quality product B, the introduction of a "decoy" that is higher in price/quality than A will bias preferences towards A, whereas an item that is lower in price/quality than B will bias choices towards B, as if participants had an intrinsic preference for the compromise solution [54]. This preference is predicted by a need to keep goals in equilibrium, by assuming that agents aspire to high quality and low price simultaneously, rendering the agent's indifference curve over options convex to the origin (**Fig. 2b**).

In natural environments we are faced with choices between offers of different quantity and quality, in variable states of need. In such scenarios, our proposal predicts that both goal difference (the relative level of two internal resources A and B) and offer difference (the relative quality of offer of A and B) will predict choices, **Fig. 2c**. This prediction was tested in a recent study in which participants made decisions that directly trade off multiple assets. Juechems [55] asked participants to perform a task that involved managing a virtual zoo containing lions and elephants. On each trial they populated the zoo by choosing among offers of variable numbers of each animal, and the financial success of their zoo was proportional to the minimum number of either lions or elephants present on each trial (which was not directly observed), **Fig. 4a**. Each choice was classified according to whether it "redressed" the imbalance in cumulative lions and elephants and decisions to "redress" were predicted by both goal difference and offer difference, **Fig. 4b**. Critically, BOLD signals in the vmPFC signalled the level of redress incurred by a choice – the extent to which a choice brought the numbers of animals into equilibrium, but independent of the total animal numbers and thus of any overall financial return. Interestingly, the dACC coded for the level of imbalance among assets, i.e. the pressure to redress this imbalance, reminiscent of the "risk pressure" signal described above [46], **Fig. 4c**. Whilst the conclusions that one can draw from this study may be limited by the fact that (by experimental necessity) lions and elephants were ultimately converted to a single currency (money), we think it provides preliminary evidence for the framework advanced here.

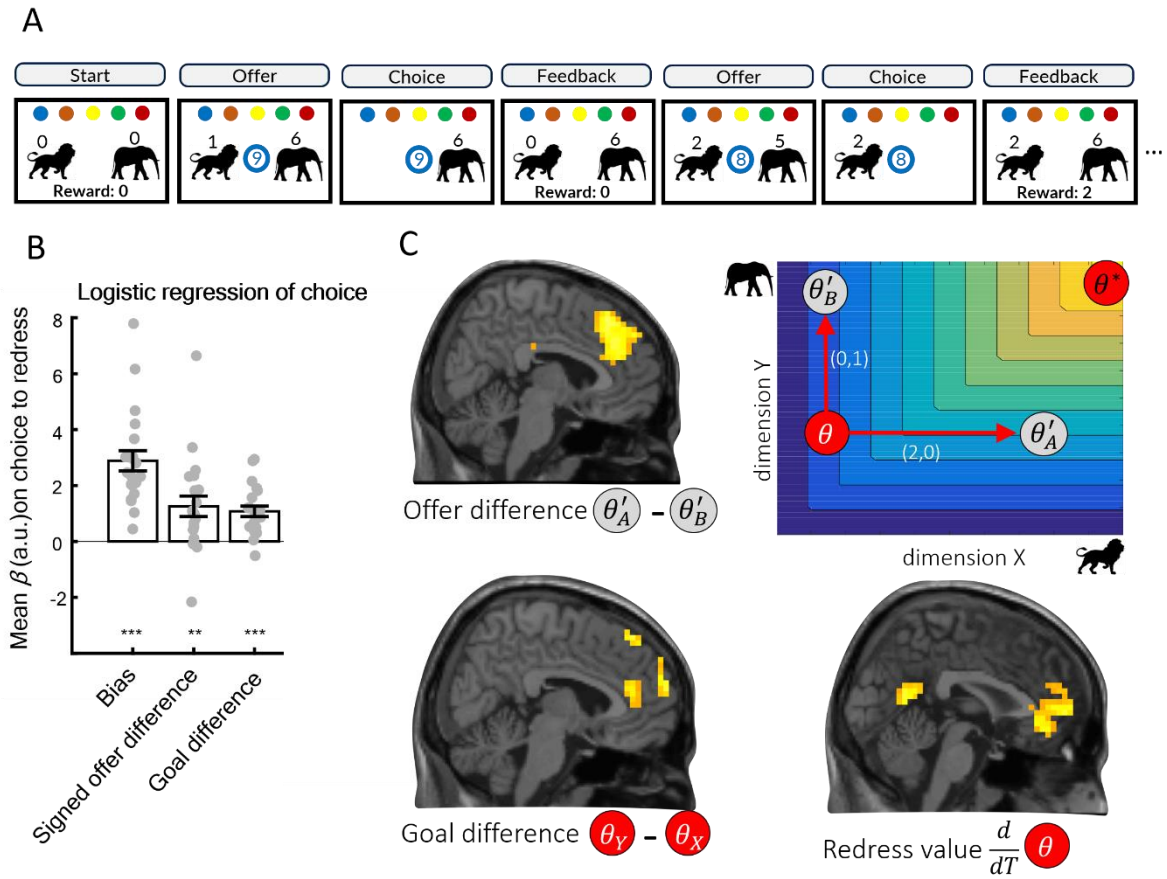

Figure 4. Neural signals for computing goal equilibrium. A. Experimental task used in Juechems et al., 2019. Participants managed virtual zoos in which they needed to accumulate as many lions and elephants as possible. Starting out from an empty zoo, participants were offered varying amounts of lions and elephants to acquire. Importantly, reward was calculated as the lower tally of animals currently owned within the zoo (e.g. when participants owned 2 lions and 6 elephants, they earned 2 units of reward). This enforced a rule by which the two tallies should be held approximately in equilibrium across time. Participants completed five zoos per block with their respective length indicated as a countdown in the centre of the screen. **B.** Logistic regression predicting a

"redress" choice – i.e. whether participants chose to increase the lower tally of animals currently in the zoo (as given by the feedback screen). Participants placed approximately equal weight on the offer and goal (tally) differences, as in the simulation in **Fig. 2c**. **C.** Indifference map over the two assets (lions and elephants) and three key neural correlates. Regions in the dACC and vmPFC encoded the offer difference, goal difference, and the redress (the decrease in distance to goal between trials). Reprinted with permission from ref [55].

Where do goals come from?

Our focus here has been on "where value comes from". Drawing upon homeostatic theories of motivation, we argue that the value of actions taken in cognitive, social and economic settings depends at least in part on the extent to which they drive an agent towards – or away from – a setpoint. However, this begs another, potentially more thorny question: how are the setpoints set in the first place? In those use cases to which HRRL has been applied thus far, pertaining to physiological needs, this is relatively uncontroversial, as it is safe to assume that setpoints for satiation and hydration are innate. What, then, of our desire to climb a mountain or become a lawyer?

In the RL framework, goals are chosen because of their potential reward value. But if reward is a reflection of the distance to a goal, then in our hands this argument becomes circular. One possibility is that goal-setting involves a hierarchical means-end reasoning process, by which proximal needs beget simple goals, the achievement of which requires more complex goals, and so on in a hierarchy of abstraction. In fact, the earliest theories of executive function and the prefrontal cortex proposed just such a mechanism for goal-directed behaviour [56]. Another factor that should not be overlooked is the role of social influence in goal setting. We often seek to achieve goals because they were suggested by our friends and associates, or by commercial advertising. However, we have to admit that we do not know how goals are computed. We think it is likely that the process by which we set ourselves goals is at least partly arbitrary, and otherwise intricately linked with the most intricate and ephemeral aspects of human cognition that pertain to our identity and sense of self.

Concluding Remarks

Our contribution here is twofold. Firstly, we highlight some challenges for RL as a computational theory of motivated behaviour in biological agents. The RL framework as developed within machine learning assumes that rewards originate in the environment, that they are sensed by dedicated input channels, that they are not fungible, and (typically) unidimensional. Secondly, we suggest that an alternative framework [14, 18, 20, 29] that treats reward as the by-product of computing distance to largely self-defined (*intrinsic*) goals, may have rich explanatory power for understanding human decisions. We show how this framework readily accounts for commonly observed effects in human and animal behaviour, such as reward-maximization, reference-dependence, diminishing marginal utility, and convex indifference between goods. Empirical tests of the theory's wider predictions will require tasks (or field observations) that involve accumulation of multiple assets. Such tasks are rare in the literature concerning value-guided choices, and where multiple value dimensions inadvertently come into play, they are often even discarded as nuisance variables. We hope that the ideas advanced here will suggest ways in which to probe goal-directed behaviour using richer tasks which more closely emulate the fact that human (as well as animal) behaviour is driven by multiple, competing goals.

Acknowledgements

This work was funded by the European Research Council (grant REP-725937 to C.S.) and also received funding from the European Union's Horizon 2020 Framework Programme for Research and Innovation

under Specific Grant Agreement 785907 (Human Brain Project SGA2, to C.S.). We are grateful to a large number of colleagues for interesting discussions on this topic and constructive feedback about a preprint version of the MS (and especially Tom Ringstrom); to Mehdi Keramati and two anonymous reviewers for helpful and constructive critique; and to Boris Gutkin, who thoughtfully provided an informal review of the revised submission.

References

1. Sutton, R. and Barto, A. (1998) Reinforcement Learning, MIT press.
2. Dayan, P. and Daw, N.D. (2008) Decision theory, reinforcement learning, and the brain. Cogn Affect Behav Neurosci 8 (4), 429-53.
3. Dolan, R.J. and Dayan, P. (2013) Goals and habits in the brain. Neuron 80 (2), 312-25.
4. Schultz, W. et al. (1997) A neural substrate of prediction and reward. Science 275 (5306), 1593-9.
5. Mnih, V. et al. (2015) Human-level control through deep reinforcement learning. Nature 518 (7540), 529-33.
6. Silver, D. et al. (2017) Mastering the game of Go without human knowledge. Nature 550 (7676), 354-359.
7. Maslow, A.H. (1943) A theory of human motivation. Psychological Review 50, 370–96.
8. Fadel, J. and Deutch, A.Y. (2002) Anatomical substrates of orexin-dopamine interactions: lateral hypothalamic projections to the ventral tegmental area. Neuroscience 111 (2), 379-87.
9. Schultz, W. (2015) Neuronal Reward and Decision Signals: From Theories to Data. Physiol Rev 95 (3), 853-951.
10. Berridge, K.C. (2012) From prediction error to incentive salience: mesolimbic computation of reward motivation. Eur J Neurosci 35 (7), 1124-43.
11. Mollenauer, S.O. (1971) Shifts in deprivation level: different effects depending on the amount of preshift training. Learning and Motivation 2, 58-66.
12. Pompilio, L. and Kacelnik, A. (2005) State-dependent learning and suboptimal choice: When starlings prefer long over short delays to food. Animal Behaviour 70, 571–578.
13. Burke, K.A. et al. (2008) The role of the orbitofrontal cortex in the pursuit of happiness and more specific rewards. Nature 454 (7202), 340-4.
14. O'Reilly, R.C. et al. (2014) Goal-Driven Cognition in the Brain: A Computational Framework. arXiv arXiv:1404.7591.
15. Knox, R.E. and Inkster, J.A. (1968) Postdecision dissonance at post time. Journal of Personality and Social Psychology 8, 319–323.
16. Daw, N.D. et al. (2005) Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. Nat Neurosci 8 (12), 1704-11.
17. Pezzulo, G. et al. (2015) Active Inference, homeostatic regulation and adaptive behavioural control. Prog Neurobiol 134, 17-35.
18. Keramati, M. and Gutkin, B. (2014) Homeostatic reinforcement learning for integrating reward collection and physiological stability. Elife 3.
19. Berridge, K.C. (2004) Motivation concepts in behavioral neuroscience. Physiol Behav 81 (2), 179-209.
20. Hull, C.L. (1943) Principles of Behavior: An Introduction to Behavior Theory, Appleton-Century-Croft.
21. Rushworth, M.F. et al. (2011) Frontal cortex and reward-guided learning and decision-making. Neuron 70 (6), 1054-69.
22. Rangel, A. et al. (2008) A framework for studying the neurobiology of value-based decision making. Nat Rev Neurosci 9 (7), 545-56.
23. Kable, J.W. and Glimcher, P.W. (2009) The neurobiology of decision: consensus and controversy. Neuron 63 (6), 733-45.
24. Glimcher, P.W. (2004) Decision, Uncertainty and the Brain: : The Science of Neuroeconomics, MIT Press.
25. Keramati, M. and Gutkin, B., A Reinforcement Learning Theory for Homeostatic Regulation, Advances in Neural Information Processing Systems, 2011.
26. Hulme, O. et al. (2019) Neurocomputational Theories of Homeostatic Control. https://psyarxiv.com/s2q46/.
27. Keramati, M. et al. (2017) Cocaine addiction as a homeostatic reinforcement learning disorder. Psychol Rev 124 (2), 130-153.

584 28. Cartwright, B.A. and Collett, T.S. (1983) Landmark learning in bees. Journal of comparative
585 physiology 151, 521-543.
586 29. Klinger, E. (1975) Consequences of Commitment to and Disengagement from Incentives. Psychol
587 Rev 82, 1-25.
588 30. Keramati, M. et al. (2017) Misdeed of the need: towards computational accounts of transition to
589 addiction. Curr Opin Neurobiol 46, 142-153.
590 31. Ringstrom, T. and Schrater, P. (2019) Constraint Satisfaction Propagation: Non-stationary Policy
591 Synthesis for Temporal Logic Planning. arXiv:1901.10405.
592 32. Camerer, C. et al. (1997) Labor supply of New York City cabdrivers: one day at a time. Quarterly
593 Journal of Economics, 407-441.
594 33. Thaler, R. (2015) Misbehaving: The Making of Behavioural Economics, W. W. Norton & Company.
595 34. Miller, N.E. and Kessen, M.L. (1952) Reward effects of food via stomach fistula compared with
596 those of food via mouth. Journal of Comparative and Physiological Psychology 45, 555–564.
597 35. Olds, J. and Milner, P. (1954) Positive reinforcement produced by electrical stimulation of septal
598 area and other regions of rat brain. Journal of Comparative and Physiological Psychology 47, 419-
599 427.
600 36. Young, P.T. (1966) Hedonic organization and regulation of behavior. Psychol Rev 73 (1), 59-86.
601 37. Schuck, N.W. et al. (2016) Human Orbitofrontal Cortex Represents a Cognitive Map of State
602 Space. Neuron 91 (6), 1402-1412.
603 38. Wilson, R.C. et al. (2014) Orbitofrontal cortex as a cognitive map of task space. Neuron 81 (2),
604 267-79.
605 39. Padoa-Schioppa, C. (2013) Neuronal origins of choice variability in economic decisions. Neuron
606 80 (5), 1322-36.
607 40. Abitbol, R. et al. (2015) Neural mechanisms underlying contextual dependency of subjective
608 values: converging evidence from monkeys and humans. J Neurosci 35 (5), 2308-20.
609 41. Eldar, E. et al. (2016) Mood as Representation of Momentum. Trends Cogn Sci 20 (1), 15-24.
610 42. Vinckier, F. et al. (2018) Neuro-computational account of how mood fluctuations arise and affect
611 decision making. Nat Commun 9 (1), 1708.
612 43. Juechems, K. et al. (2017) Ventromedial Prefrontal Cortex Encodes a Latent Estimate of
613 Cumulative Reward. Neuron 93 (3), 705-714 e4.
614 44. Tsetsos, K. et al. (2014) Neural mechanisms of economic commitment in the human medial
615 prefrontal cortex. Elife 3.
616 45. Caraco, T. (1981) Energy budgets, risk and foraging preferences in dark-eyed juncos (Junco
617 hyemalis). Behav Ecol Sociobiol. 8, 213–217.
618 46. Kolling, N. et al. (2014) Multiple neural mechanisms of decision making and their competition
619 under changing risk pressure. Neuron 81 (5), 1190-1202.
620 47. Symmonds, M. et al. (2010) A behavioral and neural evaluation of prospective decision-making
621 under risk. J Neurosci 30 (43), 14380-9.
622 48. Balaguer, J. et al. (2016) Neural Mechanisms of Hierarchical Planning in a Virtual Subway
623 Network. Neuron 90 (4), 893-903.
624 49. Shidara, M. and Richmond, B.J. (2002) Anterior cingulate: single neuronal signals related to
625 degree of reward expectancy. Science 296 (5573), 1709-11.
626 50. Kennerley, S.W. et al. (2006) Optimal decision making and the anterior cingulate cortex. Nat
627 Neurosci 9 (7), 940-7.
628 51. Roijers, D.M. et al. (2013) A Survey of Multi-Objective Sequential Decision-Making. Journal of
629 Artificial Intelligence Research 48, 67-113.
630 52. Padoa-Schioppa, C. and Assad, J.A. (2006) Neurons in the orbitofrontal cortex encode economic
631 value. Nature 441 (7090), 223-6.
632 53. Wike, E.L. and Barrientos, G. (1958) Secondary reinforcement and multiple drive reduction.
633 Journal of Comparative and Physiological Psychology 51, 640-643.

54. Simonson, I. (1989) Choice Based on Reasons: The Case of Attraction and Compromise Effects. Journal of Consumer Research 16, 158-174.
55. Juechems, K. et al. (2019) A Network for Computing Value Equilibrium in the Human Medial Prefrontal Cortex. Neuron 101 (5), 977-987 e3.
56. Miller, G.A. et al. (1960) Plans and the structure of behavior, Holt, Rhinehart, & Winston.
57. Oudeyer, P. and Kaplan, F. (2007) What is Intrinsic Motivation? A Typology of Computational Approaches. Front Neurorobotics 1:6.
58. Ryan, R.M. and Deci, E.L. (2000) Intrinsic and Extrinsic Motivations: Classic Definitions andNew Directions. Contemporary Educational Psychology 25, 54–67.
59. Gottlieb, J. and Oudeyer, P.Y. (2018) Towards a neuroscience of active sampling and curiosity. Nat Rev Neurosci 19 (12), 758-770.
60. Guo, X. et al. (2016) Deep Learning for Reward Designto Improve Monte Carlo Tree Search in ATARI Games. arXiv.
61. Friston, K. (2010) The free-energy principle: a unified brain theory? Nat Rev Neurosci 11 (2), 127-38.