Action discovery and intrinsic motivation: a biologically constrained formalisation

K.N. Gurney[1], N.Lepora[1], A. Shah[1], A. Koene[2], and P. Redgrave[1]

[1] Adaptive Behaviour Research Group, Department of Psychology, University of Sheffield, UK

[2] Laboratory for Integrated Theoretical Neuroscience, RIKEN Brain Science Institute, Japan

{k.gurney, n.lepora, a.shah, p.redgrave}@sheffield.ac.uk, koene_at_brain.riken.jp

Abstract. We introduce a biologically motivated, formal framework or 'ontology' for dealing with many aspects of action discovery which we argue is an example of intrinsically-motivated behaviour (as such, this chapter is a companion to that by Redgrave et al in this Volume). We argue that action-discovery requires interplay between separate internal forward models of prediction, and inverse models mapping outcomes to actions. The process of learning actions is driven by transient changes in the animal's policy (repetition bias) which is, in turn, a result of unpredicted, phasic sensory information ('surprise'). The notion of salience-as-value is introduced and broken down into contributions from novelty (or surprise), immediate reward acquisition, or general task/goal attainment. Many other aspects of biological action-discovery emerge naturally in our framework which aims to guide future modelling efforts in this domain.

1 Introduction

As described in detail elsewhere in this volume, there are several reasons why behaviour can be described as 'intrinsically-motivated' and why intrinsically-motivated behaviour is useful. Common to many accounts is the idea that intrinsically-motivated behaviour allows us to gain competence in achieving goals in an environment by developing skills for, and knowledge of, our interaction with that environment (see for example Barto et al., 2004). In addition, intrinsically-motivated behaviour of this kind usually results in the development of internal models of the action-outcome causality or 'know-how' (Oudeyer and Kaplan, 2007). Such competences allow us to accomplish subsequent tasks and goals more effectively.

In this chapter we focus on how intrinsic motivation helps an animal determine action-outcome causality. Recently we have developed the first steps in a biologically plausible account of this process (Redgrave and Gurney, 2006; Redgrave et al., 2008). These ideas are also described in a companion chapter in this Volume **??**, and summarised in section 2. The focus of that work was on an

analysis of the physiological and anatomical evidence that implicates short latency phasic (transient) changes in the levels of the neurotransmitter dopamine in learning causality, and in particular, its role as a signal of sensory prediction error.

In the tradition of Marr and Poggio (1976) we have therefore proposed a computational rationale for phasic dopamine. Thus, in brief, phasic dopamine causes the animal to repeat any movements it may have made immediately prior to a surprising event. By repeatedly executing those movements, the animal can determine agency – did the animal's movement cause the surprising event? It can also determine exactly what movements, and under what circumstances they must be executed to provide the outcome. The acquisition of this knowledge is the discovery of a novel action – analogous in many respects to a skill or an 'option' (Barto et al., 2004) – that can be used to interact with the environment.

In Marr's analysis, the next step is to determine *how* the computation (of action discovery) is performed. In our scheme we propose that the repetition of movement execution repeatedly generates representations of those movements, their circumstances, and the outcome. These representations are them presented to associative networks in the brain responsible for building internal models of action-outcome contingency. However, in attempting to articulate this process in detail, several ideas such as 'prediction', 'prediction error', 'habituation', 'salience', 'sensory context', etc., play a prominent role. Some of them may be characterised as representations in the brain, other are putative processes manipulating these signals. In any case, their definitions are often somewhat nebulous and they all remain to be formally defined. Since much of this chapter is aimed at remedying this, it is important to understand the importance of such a project.

Many times in the cognitive sciences, debates (sometimes rather heated!) occur about the meaning, interpretation and status of concepts, terms and definitions. One pertinent example is that of 'reward'. This has a more specific interpretation in biology (being confined to appetitive stimuli such as food, liquid etc.) than it does in computational reinforcement learning theory (e.g. see Sutton and Barto, 1998), where it is semantically neutral and is defined implicitly by its symbolic occurrence in learning algorithms. An understanding of this situation lies at the heart of our recent analysis of the role of dopamine in reinforcement learning (Redgrave and Gurney, 2006). A lack of precision in our conceptualisation can lead to lack of collaboration, wasted effort and missed opportunities for advancement of the subject.

A similar set of problems occurred in Artificial Intelligence (AI) during attempts in the 1980s to capture, precisely, knowledge in a specific domain so that it could be manipulated in AI programmes like 'expert systems'. This culminated in the formulation of the notion of a domain-specific *ontology*. According to Gruber (1992) "[an ontology is] a specification of a conceptualisation... a description (like a formal specification of a program) of the concepts and relationships that can formally exist for an agent or a community of agents...". The requirement of such an ontology, as used by computer scientists, is that it is specified in a formal ontology language. However, the bioinformatics community has made

real progress with ontology-like tools with more relaxed frameworks. The Gene Ontology (Gene Ontology Consortium, 2001) is a tool for the representation and processing of information about gene products and functions. According to the Gene Ontology Consortium, "The exponential growth in the volume of accessible biological information has generated a confusion of voices surrounding the annotation of molecular information about genes and their products. The Gene Ontology (GO) project seeks to provide *a set of structured vocabularies* for specific biological domains..." [our italics]. By helping to dispel the 'confusion of voices' in its subject domain, GO has materially facilitated the progress of genetics science. We argue that cognitive science and robotics can avoid a similar cacophony by using an appropriate 'structured vocabulary' for their discourse, defined using the relevant formal methods.

In this chapter, we therefore attempt to provide a formalisation of some terms in the vocabulary of action selection, and action discovery, which will also be applicable in discussions of behaviour, and intrinsically-motivated learning in general. In defining an ontology of action discovery it is essential to understand what is being done: it is *not* the case that we are seeking to establish the 'truth' that the normal language label L 'really is' given by the formal definition D (e.g. 'action' really means D_1 rather than D_2) . Rather, we are proposing that *defining* label L to mean D is *useful* because D is useful for our purposes, and D is *plausibly* assigned the label L. Other mappings may also be plausible, but at least we should try and be clear which mapping we are using. Disagreement about such mappings should not take priority over establishing a formalism that is self-consistent, comprehensive, and useful in formulating new theories and models.

The outcome of a programme such that presented here is that we sharpen the computational hypothesis of agency being proposed here. In addition, it helps us to develop key functional architectural components for action discovery. Finally, the interpretation of alternative hypotheses in the same framework will facilitate a comparison of hypotheses. We start by describing the behavioural and neurobiological paradigm we are attempting to formalise.

2 The action-outcome paradigm

We first outline the situation we have in mind in functional terms; many of the terms such as 'action' and 'context', which we define more precisely later, occur informally here. We imagine an animal interacting with its environment and trying to discover causal relations between itself and the environment; that is, developing a sense of 'agency', in which events caused by the agent are discovered, and the causal components of behavioural output determined. This is supposed to occur through an exploration of the animal's environment in a way that is governed by the unpredictability of the outcome associated with the action(s) performed.

In much of our exposition elsewhere, we have used the term 'novelty' in an very general sense to mean this unpredictability. Here we will find it useful to

distinguish between the unpredictability of simple phasic outcomes (such as a luminance change) and the more general case that might require evaluation of complex aspects of the environment (e.g. new objects, or old ones in new situations). We will refer to the former as *surprise* and the general case as *novelty*. Thus, we conceive of surprise as a special (limited) instance of novelty, and will use 'novelty' where a general interpretation suffices, and only specialise to 'surprise' where necessary. These definitions are formalised in section 3.3. Much of our exposition will focus on surprise, which is sufficient to cause phasic dopamine release (Schultz et al., 1997).

In the action-outcome paradigm, post-action, unpredicted change may, or may not, be caused by the animal's behaviour; it might be purely coincidental and non-correlated with behaviour. The animal is therefore attempting to discover whether there is, in fact, a causal link between any aspect of its behaviour and the outcome. In any case, if the post-action environment is unexpected, there is a sensory prediction error between what is anticipated and what actually occurs.

While the outcome is surprising, the animal continues to explore possibilities for causal relations between its actions and the environment. During this exploratory phase, we suppose the animal will have its action-selection policy biased to repeat those actions which most recently caused unpredicted outcomes; we refer to this process as *repetition bias*. As well as being surprising, these outcomes may contain other novel elements requiring rich sensory representations. Repetition bias allows the reliable presentation of such representations, together with relevant contextual and motor (action-related) signals, to associative networks in the brain, thereby allowing establishment of internal models of causal relations.

In the process of 'latent learning' (Tolman, 1948) animals which have *sufficient exposure* to an environment are able to learn a model of that environment (including action-outcome contingencies) which may be recruited at a later stage for goal-directed behaviour. One view of repetition bias in our scenario, therefore, is that it is a means of ensuring rapid and plentiful exposure to the correct stimuli for learning of action-outcome contingencies for later deployment.

We now detail some neurobiological specifics which act as constraints on our thinking. We propose that the ability for a surprising outcome to effect policy change (repetition bias) is critically dependent on it being able to elicit short latency phasic (transient) changes in levels of the neurotransmitter dopamine. This, in turn, works by facilitating synaptic plasticity between cortex and the basal ganglia - a set of sub-cortical structures which are believed to be a key locus for implementing the action-policy . Details of how this synaptic plasticity causes policy change and the role of the basal ganglia in our paradigm are provided in the companion chapter. For our purposes here, it is enough to summarise the supposed sequence of neural events after an action with an unpredicted causal outcome (numbers refer to items in Figure 1).

Cortical representations of sensory and motivational information (1) initiate representations of motor *preparation* (2). For example, in the case of reaching,

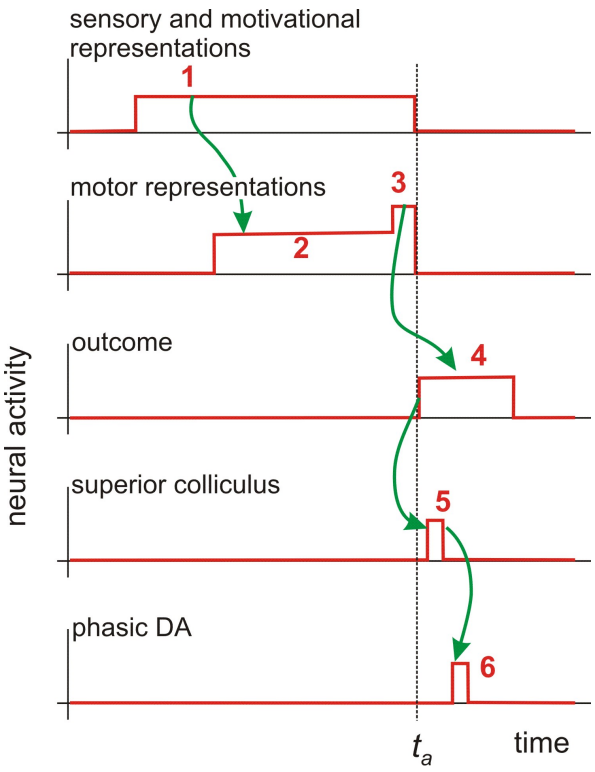

Fig. 1. Neural representations and causality in the action-outcome learning scheme. Red traces are activity levels in various neural representations. Green arrows show causal relation between representations

this preparatory activity is encoded in areas of posterior parietal cortex (PPC) (Snyder et al., 1997). Of all possible actions being initiated, usually only only one will be selected for execution by the basal ganglia.

In Figure 1, the action is supposed to occur prior to time t_a and has motor representation (3) (shown with signal level elevated with respect to that for preparatory motor representations). The basis for this selection is that highly active or *salient* representations in cortex are able to dominate competitive mechanisms in basal ganglia, thereby enabling their behavioural enactment. Salience may be influenced by task, goal or reward related information.

It is important to note that the neural signals have been shown for only one motor act. In addition, behaviour is assumed continuous, and so similar sets of signals will be in various stages of development at any time. This scheme is similar to that of Cisek (2007); Cisek and Kalaska (2010) who stress the intimate relation between partial action specification (preparation), and action selection. Thus, the environment offers a continuous stream of possible actions or *affordances* which generate an associated stream of contextual and preparatory

activity for each action. These activities are then run in a continuous competition to yield a single behavioural action. This ideas is formalised in our notion of *action request* (section 4.1). The gradual transition from perceptual to action-based representation also has resonance with the notion that perception exists *for* action, not passive evaluation of the environment (Allport et al., 1987).

The action will cause a phasic change in the environment which leads to a perceptual representation of that outcome (4). In particular, there will be a signal in superior colliculus (a mid-brain structure) that responds simply to the phasic onset of the outcome and a phasic gaze shift is initiated. The collicular signal will also elicit a phasic response in dopamine neurons via their direct innervation (Comoli et al., 2003; Dommett et al., 2005). The phasic dopamine facilitates cortico-striatal plasticity which makes the repetition of the action more likely in the current situation. Finally, after the gaze shift and after more sophisticated cortical sensory processing, more complex representations of the structural features of the outcome will occur in cortex.

As an example, consider the sensory information provided by the animal being near a light switch constituted by a long toggle or lever. If the animal doesn't know the specific action required to operate the switch, several motor preparatory representations might be elicited (pushing up or down, rotating the toggle etc). Eventually an action is selected and some outcome ensues. If it turns on the light, there will be a phasic response in the colliculus due to elevated luminance, and subsequently, a specific light will be identified as the outcome with accompanying nuances of light levels, shadows cast etc. In a more extreme scenario, the animal will not have acted intentionally with the toggle, but will have accidentally switched it on while pursuing other behaviours. More exploration is required in this case to discover causality, but the principles of action-discovery we outline here are supposed to apply quite generally.

Other relevant process at work here include habituation and sensitisation. Habituation of sensory representations refers to a decline in response when the perceptual features are no longer surprising or novel, *and* don't have any rewarding consequences (Sokolov, 1963). In the example, once the causal relation between the switch and the light onset has been discovered, we expect the colliculus response to decline (if it has no rewarding consequences). In contrast, sensitisation (elevated response) occurs if the caused event is a reward or reward predictor (Ikeda and Hikosaka, 2003; Wurtz and Albano, 1980). These phenomena will form a key part of our narrative.

3 Formalisation

3.1 The environment and its internal representation

The environment: Our basic objects are dynamically evolving environmental or 'world' states, which are experienced by the animal or agent, and corresponding internal states or neural representations, which are constituted by patterns of activity over populations of neurons. Thus, we suppose that, at any time t_s,

the external world is in some state $\boldsymbol{\gamma}(t_s) \in \Gamma_S$, where Γ_S is the set of all possible world states, and t_s is the time at which the state occurs. The use of bold symbols to represent states implies they are some kind of *vector* [3]. The evolution of the world in time through this state space then defines a vector-valued function[4], its trajectory or *world path* $\boldsymbol{\gamma}$, through Γ_S

$$\boldsymbol{\gamma} : \mathbb{R} \to \Gamma_S \ , \qquad \boldsymbol{\gamma} \in \Gamma \ ,$$

where the space of possible world paths is denoted by Γ. While there is no technical limit on the time domain, pragmatically, we may choose to limit time around some relevant epoch in the animal's history with consequent limiting of Γ.

The use of dynamic trajectories in *function* space as the grounding for our ideas, rather than instantaneously defined *states*, allows a more flexible and realistic interpretation of perception, its relation to action, and the use of prediction. It encompasses, as a special case, the use of states defined at single times, even if these are defined over a continuous time domain (as, for example, in the work of Oudeyer and Kaplan (2007)). The approach is inspired by the use of functional methods in physics for studying dynamics and, while it lacks the superficial simplicity of discrete, state-based views, we contend that because animal behaviour is at least as complex as that of inanimate systems, the functional approach will ultimately facilitate a simpler analysis when we confront the complexities of behaviour head on.

Representations of the environment: Corresponding to states in the world, we suppose there are internal states of the agent, or brain states in some set N_S. We will refer to the components of the vectors in N_S as the *neural features* of the state (which might, for example, be the activity of a population of neurons).

Then, in line with our continuous, dynamic approach we define the space of time-dependent *neural representations* N_Γ which are vector-valued functions, $\boldsymbol{y}$, of time

$$\boldsymbol{y} : \mathbb{R} \to N_S \ , \qquad \boldsymbol{y} \in N_\Gamma \ .$$

Representations are supposed to arise in the brain via sensory processes perceiving the environment as a stimulus. Although the exact trajectory of $\boldsymbol{y}$ may also depend on the history of the agent and its internal states, we will refer to a transformation from world trajectories to their neural correlates as a *sensory transformation* $\mathcal{S}$; it is a mapping from the space of world paths Γ to representations N_Γ

$$\mathcal{S} : \Gamma \to N_\Gamma \qquad \text{with} \qquad \boldsymbol{y} = \mathcal{S}[\boldsymbol{\gamma}] \tag{1}$$

[3] We use the term 'vector', but in the sense adopted in computer science to mean a 1D-array or n-tuple; there is no implication that these n-tuples form a true vector space. Indeed, if we use only positive valued components (a natural choice to indicate presence of a feature) the space does not have an additive inverse.

[4] We use the normal convention that y denotes a function and $y(t)$ its value at time t.

(where the notation $\mathcal{T}[\cdot]$ denotes a mapping from one function space to another). At this stage, $\boldsymbol{y}$ includes all internal state information about the agent; we do not distinguish between 'sensory' and 'motor' representations, although this will prove useful later. Note that we may be interested in a variety of different sensory transformations with different levels of complexity, making use of more or less limited representation spaces. However we will avoid a proliferation of such spaces as far as possible and refer to relations in (1) in a generic sense.

It is also natural to seek a mapping between neural representations and the world, a feature which will be particularly useful in dealing with prediction. In general, we assume that the world is richer than our mental representations of it. For example, visual perception is acknowledged to be an ill-posed problem (Poggio and Koch, 1985) in which multiple world states give rise to the same visual percept, and more broadly, perceptual neural representations generalise across stimuli (small nuances of the world are often lost unless we have a special reason to encode them). There is, in general therefore, an equivalence class of world paths $\tilde{\boldsymbol{\gamma}}$, which all have the same neural representation $\boldsymbol{y}$, and we define the inverse mapping from representations to the set of world path equivalence classes $\tilde{\Gamma}$

$$\mathcal{S}^{-1} : N_\Gamma \to \tilde{\Gamma} \qquad \text{with} \qquad \mathcal{S}^{-1}[\boldsymbol{y}] = \tilde{\boldsymbol{\gamma}} \; . \tag{2}$$

Contexts and outcomes: In the action-outcome situation we consider, much hinges around comparing representations either side of some critical time (t_a in Figure 1). It is therefore convenient to define subsegments of $\boldsymbol{\gamma}$ with respect to the current time t_s. From the animal's point of view, that part of $\boldsymbol{\gamma}$ in the past constitutes its *context* $\boldsymbol{\gamma}^-$. Thus, if the time domain is $T^- = (-\infty, t_s]$

$$\boldsymbol{\gamma}^- : T^- \to \Gamma_S \qquad \text{with} \qquad \boldsymbol{\gamma}^-(t) = \boldsymbol{\gamma}(t) \; , \qquad t \in T^- \; .$$

In a similar way, we define the *outcome* as the future trajectory; so, with time domain $T^+ = [t_s, \infty)$

$$\boldsymbol{\gamma}^+ : T^+ \to \Gamma_S \qquad \text{with} \qquad \boldsymbol{\gamma}^+(t) = \boldsymbol{\gamma}(t) \; , \qquad t \in T^+ \; .$$

The context has a corresponding neural representation $\boldsymbol{y}^-$, where

$$\boldsymbol{y}^- = \mathcal{S}[\boldsymbol{\gamma}^-] \; ,$$

and the functions $\boldsymbol{\gamma}^-$, and $\boldsymbol{y}^-$ reside in spaces Γ^- and N_Γ^- respectively. These spaces are simply suitable temporal restrictions of Γ and N_Γ and will be useful in introducing notions of prediction. The outcome also has representation $\boldsymbol{y}^+ = \mathcal{S}[\boldsymbol{\gamma}^+]$ with associated spaces N_Γ^+, Γ^+.

The physics of the world dictate how contexts are transformed into outcomes and we define a *world transform* $\mathcal{W}$

$$\mathcal{W} : \Gamma^- \to \Gamma^+ \qquad \text{where} \qquad \boldsymbol{\gamma}^+ = \mathcal{W}[\boldsymbol{\gamma}^-] \; . \tag{3}$$

Reference to this kind of transform will serve to highlight the contrast between the way contexts get transformed into outcomes in the world, with the way

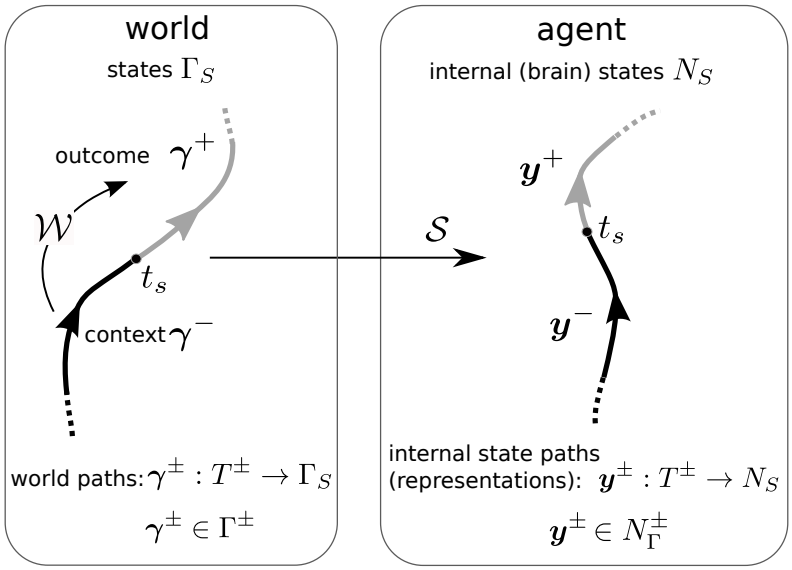

Fig. 2. Graphic depiction of the formalisation thus far. The world (or environment) is described by spaces of vector-valued states (Γ_S, N_S, respectively). However, the emphasis is on vector-valued functions of time which define 'trajectories' or 'paths' through these state spaces. It is convenient to consider the trajectory referenced to some time marker t_s so that prior/subsequent events constitute a *context/outcome* respectively. This will usually apply to some phasic event at t_s but this is not a requirement. Internal representations are derived by *sensory transforms* like $\mathcal{S}$ occurring within the agent.

their representational counterparts get transformed in the agent. The formalism described so far is summarised in Figure 2. Notice that, while we will normally regard the context/outcome boundary as some pivotal, operantly defined time $t_s = t_a$ (see Figure 1), this is not necessary, and the formalism holds for arbitrary partition of time.

3.2 Behaviour and action

Behaviour: *Behaviour* is defined in an analogous way to the environment as an evolving trajectory of *externally observable* states that the agent or animal can take. Formally, we suppose that at each time t_s, the physical pose of the animal is described by a vector $\beta(t_s)$ with $\beta(t_s) \in B_S$. Behaviour is defined as the trajectory given by the vector-valued function

$$\boldsymbol{\beta} : \mathbb{R} \to B_S \ .$$

Since the animal can observe itself, B_S is a subspace of the world states Γ_S; that is, the animal's body is part of its environment[5]. This approach enables us to subsume the effects of action into the existing framework, since behaviour is then just a part of the world's trajectory; that is, $\boldsymbol{\beta}$ is a suitable restriction of $\boldsymbol{\gamma}$ so that its range or codomain is B_S.

Action: Actions are conceived of as discrete blocks of behaviour defined over finite time intervals. In particular, we are interested in intervals just prior to operantly relevant times t_a. Thus, if $T_{\text{act}} = \{t : t_a - \Delta t < t \leq t_a\}$, for any finite Δt we define the action

$$\boldsymbol{\alpha} : T_{\text{act}} \rightarrow B_S \qquad \text{with} \qquad \boldsymbol{\alpha}(t) = \boldsymbol{\beta}(t) \ , \quad t \in T_{\text{act}} \ .$$

The functional definition of action is quite general, with no constraint on its temporal extent or ethological semantics. This generality is a useful starting point because it is notoriously non-trivial to segment behaviour in a meaningful way into discrete actions (Schleidt and Kien, 1997).

Representations of behaviour and action: We now turn to the neural representations of behaviour and action; the approach here mirrors that of the environment. Behaviour will elicit sensory representation of the agent's own body but, more importantly for us here, behaviour is associated with internal *motor states* in a subspace $N_M \subset N_S$, in which features in N_M are responsible for eliciting the behaviour. In a similar way, then, to the treatment of world trajectories and behaviour, we define a motor neural representation of behaviour, $\boldsymbol{b}$, that is a restriction of $\boldsymbol{y}$ to a function with codomain N_M. These ideas are shown in Figure 3

Note that the designation of internal state features as 'motor' (defined by their being in N_M) has no agenda about what constitutes a motor signal; indeed, different settings and models may require different selections of N_M. However, while neural representations are often allowed to have sensory and motor aspects, it is useful to be able to refer to the complementary space $N_{S\setminus M}$, of N_M with respect to N_S, as 'sensory' neural states [6]. The sensory states give rise to sensory representations $\boldsymbol{c}$ so that the entire internal representation $\boldsymbol{y}$ is defined by the pair $[\boldsymbol{b}, \boldsymbol{c}]$.

Actions are represented by suitable temporal restriction of $\boldsymbol{b}$. The neural representation of action $\boldsymbol{\alpha}$ will be denoted $\boldsymbol{a}$. Behaviour may be divided into contextual and outcome phases with respect to some specific time, t_a, and we will usually consider an action as occurring in the contextual period (that is, T_{act} terminates at t_a - see above). The context therefore has motor and sensory representations grounded in N_M and $N_{S\setminus M}$.

[5] Throughout this chapter, we use the expression 'A is a subspace of B' (or $A \subset B$) to mean that A is defined over a subset of the features in B. Also, note that B, is supposed to be upper case Greek β in keeping with the notation that Greek and Roman symbols refer to the external world and neural representations respectively

[6] $N_{S\setminus M}$ is defined over the features in N_S which are *not* contained N_M

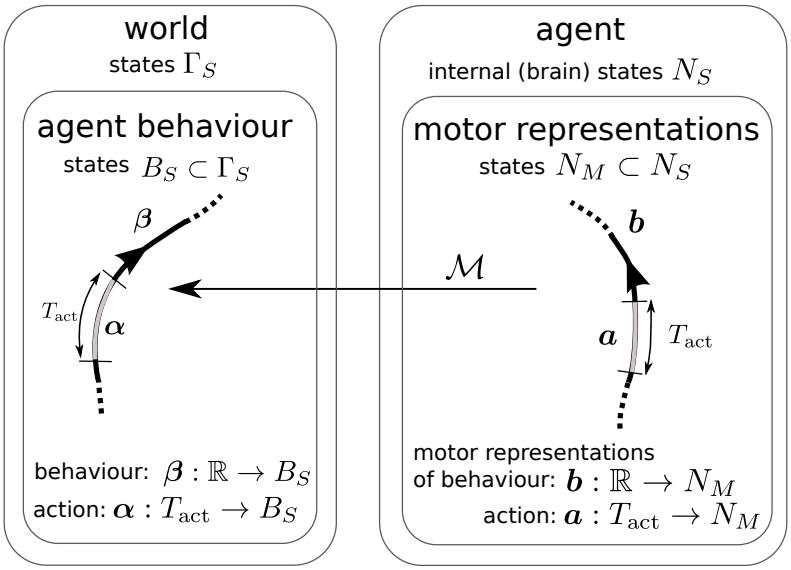

Fig. 3. Graphic depiction of the formalisation of behaviour and action. Behaviour is described with respect to a subspace of the world states $B_S \subset \Gamma_S$ via trajectories $\boldsymbol{\beta}$ through B_S. Corresponding motor representations $\boldsymbol{b}$ are defined through a subspace $N_M \subset N_S$ of internal states. Actions are defined as behaviour over a small time segment T_{act} with associated representations $\boldsymbol{a}$. *Motor transforms* like $\mathcal{M}$ occur within the agent.

From representation to behaviour: We define a *motor transformation* $\boldsymbol{\beta} = \mathcal{M}[\boldsymbol{b}]$ or its action equivalent $\boldsymbol{\alpha} = \mathcal{M}[\boldsymbol{a}]$ which shows how the internal state of the agent elicits behaviour.

Minimal and efficient action: The definition of action so far is quite general; behaviour is unsegmented agent activity, and actions, while time delimited, have no ethological definition as yet. However, we want to consider strongly causal models of the world and it may be helpful, on occasion, to be specific about which actions we are considering. Consider, the light switch toggle example described earlier. There are a multitude of actions which will 'do the job': operating the toggle while waving the other hand/paw in the air, or walking in a circle before pressing the toggle – all are effective actions in the sense defined above. However, learning that such actions are necessary for a particular outcome amounts to developing 'superstitious' behaviour (Timberlake and Lucas, 1985). What is needed is the notion of a *minimal* or most *efficient* action.

Consider first the elimination of spurious components of action ('hand or paw waving'). Elimination of action components amounts to defining behaviour using proper subspaces $B_i \subset B_S$. Then, let B_I be the intersection of all such B_i rich enough to define actions that can cause a given outcome.

Elimination of spurious prior actions amounts to considering the smallest T_{act} for which a biologically realisable action exists that can yield the given outcome. This minimisation has to be done with both world and animal in mind, for whereas the world is only concerned about the action $\boldsymbol{\alpha}(t_a)$ *at* time t_a, (e.g. required force on the toggle), it isn't possible for the animal to generate $\boldsymbol{\alpha}(t_a)$ without a prior action trajectory (start moving the paw and increase speed) which may be subject to a range of dynamic and kinematic constraints (Körding and Wolpert, 2006)

With these issues in mind we define a *minimal action* $\boldsymbol{\alpha}$, for a given outcome, as that defined by B_I, and the correspondingly smallest T_{act}. A key part of the process of *action discovery* then, may be the establishing of a minimal action for a given outcome [7]

3.3 Prediction

The notion of prediction is pervasive throughout much theoretical neuroscience: it is a key idea in computational reinforcement learning (Sutton and Barto, 1998) (where it occurs as estimates of future reward), and Friston (2005) has elevated prediction to the central pillar of any theoretical account of the brain. This is not an unreasonable stance for, if an animal can predict its environment and the result of its actions, it can generate integrated, goal directed behaviour in an efficient manner.

Prediction is a result of the animal having some kind of *internal model* of the world. In our framework, an internal model can be used to allow neural representations derived from *context* to influence future sensory representations of *outcome*, where these terms are used in the sense defined in section 3.1. We have in mind here that processes like habituation, sensitisation from reward, or 'priming' of sensory systems via task information, all result from signals becoming manifest via internal models of one kind or another, and that they all modify sensory representations.

Internal models, as we envision them, have several important characteristics in addition to their general role as predictors. First, they result from structural parameters within the brain architecture of networks of neurons and the strengths of the connections therein. Second, the internal model may be viewed as performing a form of data compression or abstraction on the input: 'raw' input is transformed into key abstract features, and it is these features that are used to generate predictions. For example, a red ball is represented not as a multitude of independent red 'pixels', but rather as the abstract concept of a sphere parameterised by colour and size. In addition, as a result of the data compression, the model generalises so that it makes similar predictions from a variety of related inputs. The data compression characteristic of an internal model enables us to make contact with theories of novelty and intrinsic motivation that are related to

[7] Other notions of action efficiency/optimality could have been used (for example, minimal energy expenditure) but temporal optimality and action simplicity seem most appropriate in the context of action discovery

information compression (Schmidhuber, 2009). Third, while the model is structurally encoded in a network architecture, the model only becomes 'expressed' or 'manifest' when it elicits output signals; that is, it generates neural activity that represents predicted sensory information (such as the appearance of a red ball within the visual field of the agent).

As an example to illustrate these ideas, consider a feedforward neural network with 'hidden' units, conceived of as a statistical model of some data. One interpretation of its operation is that, given a pattern of 'context' at the input layer, the network delivers a 'prediction' at the output layer. The model is encoded in the network connection strengths, but the prediction is made manifest only when the net delivers its output signals. The network may be viewed as performing a data compression on the input because the hidden layers extract only key features of the data (especially if the number of hidden units is less than that of the inputs). As a result of the data compression, the network generalises and will make similar 'predictions' from a wide variety of inputs.

Representation of predictions: The *internal* representations of predictions elicited by an internal model will, in general, be trajectories of internal states denoted $\mathbf{y}^*$. These constitute a class of internal representations and, since we want predictions to interact with sensory derived representations – typically in some process of comparison – we identify the space of $\mathbf{y}^*$ with N_Γ. While a sensory prediction $\boldsymbol{y}^*$ is *developed* during the contextual period (if it is to reliably influence representations in the outcome, $\boldsymbol{y}^+$) they are *deployed* during the outcome when they are compared with $\boldsymbol{y}^+$. We therefore consider the space of trajectories $\boldsymbol{y}^* \in N_\Gamma^+$.

Internal prediction models: An internal prediction model $\mathcal{I}_P$ is a map from representations of context to those of outcome, generating an internal prediction $\boldsymbol{y}^*$

$$\mathcal{I}_P : N_\Gamma^- \to N_\Gamma^+ \qquad \text{with} \qquad \boldsymbol{y}^* = \mathcal{I}_P[\boldsymbol{y}^-] \,. \tag{4}$$

We also refer to $\mathcal{I}_P$ as a feedforward model to distinguish it from inverse models defined later. In practice, the trajectories will be defined over a suitable subspace of N_S since the entire brain state is not usually required to generate a prediction. The paths $\boldsymbol{y}^*, \boldsymbol{y}^-$ are then suitable restrictions defined over this subspace.

Phasic events and their predictions: A phasic event is defined by a segment of world path, $\boldsymbol{\gamma}_\phi$, restricted to only a short time interval $T_\phi = [t_\phi, t_\phi + \Delta t]$; that is $\boldsymbol{\gamma}_\phi : T_\phi \to \Gamma_S$. Under a sensory transform, there will be a corresponding short-lived representation of the event, $\boldsymbol{y}_\phi = \mathcal{S}[\boldsymbol{\gamma}_\phi]$. We will denote the associated internal representations of predictions with their event suffix $\boldsymbol{y}^*_\phi$ where the intention is that $\boldsymbol{y}^*_\phi$ is active for a time of the order of Δt centred around T_ϕ.

Predictions in the world: We have emphasised prediction as something occurring internally to the agent. However, we also speak of predictions as being

grounded in the world – simply saying 'the light will come on' refers to the world, not our internal state. This aspect of prediction is addressed through an inverse mapping such as (2), in which the predicted sensory representation $\mathbf{y}^*$ can be transformed to an equivalence class of world trajectories, $\mathcal{S}^{-1}[\boldsymbol{y}^*] = \tilde{\boldsymbol{\gamma}}$.

Sensory error functions: Predictions only become useful if they are used to make comparisons with representations of reality, derived directly from sensory transforms of the environment, that is $\boldsymbol{y}^+$. This is accomplished using an error function $\mathcal{E}[\boldsymbol{y}^*, \boldsymbol{y}^+]$ to derive error signals e. Such functions may or may not be be true metrics or divergences, and several may be needed to capture all relevant aspects of the contrast between prediction and percept. Error signals may then be used to drive adjustment of the prediction models $\mathcal{I}_P$ so that they become increasingly accurate, and are able to deliver better predictions (in the sense of minimising e).

Novelty and surprise: Novelty and surprise have been discussed in computational terms in a variety of ways by other authors (see for example, Baldi and Itti, 2010; Oudeyer and Kaplan, 2007; Ranganath and Rainer, 2003). Our interpretation was noted in section 2 and we formalise it here. Thus, we take surprise to mean an error defined over a phasic outcome and its prediction, which makes use of only a single feature y_ϕ^+ (e.g. luminance change). That is

$$\text{surprise} : \mathcal{E}(y_\phi^+, y_\phi^*) \; . \tag{5}$$

The prediction model which gives rise to y_ϕ^* is denoted $\mathcal{I}_{P_\phi}$. Any error measure which is more general than this (i.e. with non-phasic features, or vector valued representations in general) will necessarily require the use of the term 'novelty'.

$$\text{novelty} : \mathcal{E}(\boldsymbol{y}^+, \boldsymbol{y}^*) \; . \tag{6}$$

Such an error will be associated with the general prediction model $\mathcal{I}_P$. The definition in (6) subsumes that in (5) as a special case and this is reflected in our use of terminology; we will use 'novelty' where a general interpretation suffices, and only specialise to 'surprise' where necessary.

The ideas of this section are illustrated in Figure 4. We now investigate a particularly important representation - that of *salience* - and how it is implicated in several important prediction processes.

4 Salience and action selection

In our previous work on action selection we introduced the notion of the *salience* of sensorimotor representations that contribute to the selection of actions (see companion chapter and (Redgrave et al., 1999)). Roughly speaking, salience represents the urgency of, or degree of demand for, an action, as encoded in the strength of signals afferent to basal ganglia.

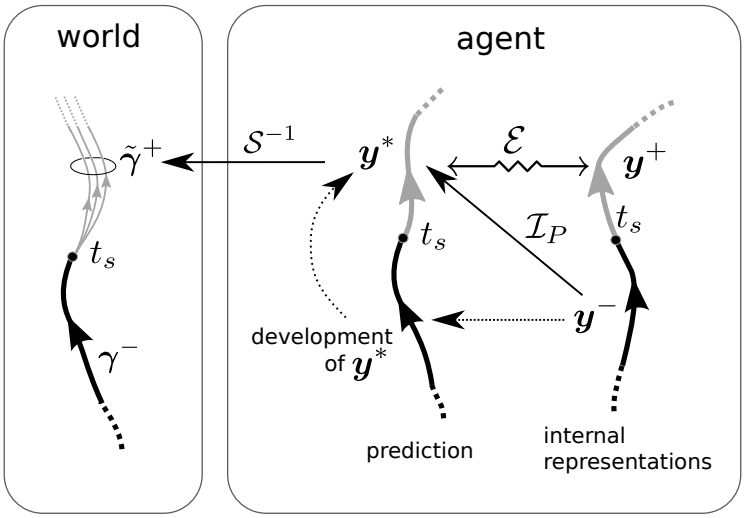

Fig. 4. Graphic depiction of the formalisation of prediction. Internal representations $\boldsymbol{y}^-$ prior to some time t_s (differentiating context from outcome) start to cause predictions to occur ('development of $\boldsymbol{y}^*$'). These continue to evolve beyond t_s and are deployed after t_s. Error signals are obtained by comparing $\boldsymbol{y}^+$ and $\boldsymbol{y}^*$ using an error function $\mathcal{E}$ (a metric or divergence between pairs of functions in the space N_Γ^+). The model $\mathcal{I}_P$ is defined as function which maps $\boldsymbol{y}^-$ to $\boldsymbol{y}^* \in N_\Gamma^+$. Internal predictions like $\boldsymbol{y}^*$ are associated (via inverse sensory maps like $\mathcal{S}^{-1}$) with non-unique world paths which collectively from equivalence classes like $\tilde{\boldsymbol{\gamma}}^+$.

For example, a strong local luminance change will, in many cases, elicit an orienting movement to the locus of that change. The stimulus gives strong activity in the corresponding areas of the brain that respond to these stimuli (such as superior colliculus and frontal eye fields). The locus of the activity in the neural tissue is often topographic with respect to spatial location of the stimulus, suggesting spatially defined 'features' in the neural representation of salience. This localised activity elicits the orienting movement (saccade or head movement) with an 'urgency' contingent on its overall 'strength' or the salience level.

4.1 Formalising salience

Salience is referenced to a collection of sensorimotor features, suggesting a vector, but at the same time, has an overall (scalar) 'strength'. The vector aspect chimes with our definition of neural representations since the supporting space (N_S and its subspaces) contain vectors. Thus, we may write $\boldsymbol{y} = (y_1, y_2, ...)$, where individual component functions y_f or features have been articulated. The function y_f may signal the presence of a particular kind of object or *world-feature*, which is, in turn, described by several components of the world trajectory $\boldsymbol{\gamma}$. To simplify

notation, we will use the index f to refer to the object or world-feature signalled by y_f. We will also use the shorter term 'feature' to refer to the 'world-feature' where the context is clear.

Salience maps: We now suppose there is a particular class of internal representations $\boldsymbol{s} \in N_\Gamma$ which is deployed in action selection policy formation. We have in mind here that $\boldsymbol{s}$ are signals afferent to basal ganglia and will call this vector the *salience map* (This reference is inspired by the common occurrence of topography of $\boldsymbol{s}$ in the brain). The scope of $\boldsymbol{s}$ is supposed to be large enough to encompass selection of all actions of interest.

Action requests: Consider now a single action $\boldsymbol{\alpha}$ occurring prior to time t_a which defines a context/outcome partition. Let $N_\alpha \subset N_\Gamma^-$ be a subspace of representations containing just sufficient features of the salience map to elicit $\boldsymbol{\alpha}$. Let $\boldsymbol{s}^{(a)}$ be the restriction of $\boldsymbol{s}$ to N_α, then we refer to $\boldsymbol{s}^{(a)}$ as the *action request* for $\boldsymbol{\alpha}$.

Two kinds of action discovery: In terms of the preceding framework, a key part of the problem of action discovery is to learn how to partition the salience map into sets N_α defining action requests for minimal actions. There are two interesting special cases of action discovery to consider here. First, the action itself (e.g. lever pressing) may be well-learned, and so any motor components of the request are well established, but the sensory contextual components are not (the lever is in a novel situation with novel outcome). The action discovery may then have the behavioural appearance of *purposeful investigative behaviour* (repeated skillful lever pressing). In a second case, both the sensory context *and* the action have to be established. Here, there will be *wide-ranging, exploratory behaviour* (discovering how to interact with levers efficiently) and will involve learning a minimal or efficient action (section 3.1)).

Salience as strength of action request : The scalar strength or *salience*, S_α of the action request is supposed to be captured by some measure of its overall level of activity. If we interpret the component features as representative of neural activity, then they are all positive and we can write

$$S_\alpha = \sum_f s_f^{(a)} \,, \tag{7}$$

where the sum is over features in the space N_α.

4.2 Salience and value

We now turn to an interpretation of the salience map that allows us to make closer contact with intrinsically-motivated learning and concepts from machine

learning. If an action is urgently requested (and assuming a well-trained animal) it is reasonable to suppose that the action is of high 'value' to the animal. We use the term 'value' in a loose sense here but, for goal directed behaviours, we suggest that the salience of an action request is, or at least is influenced by, the estimated 'action value' as used in machine learning. In action discovery, while the feature sets (defined via $N_\alpha \subset N_\Gamma$) are being determined, the value estimates are being refined. We are now naturally led to the question: what is it about an action that endows it with value?

While our arguments are valid for any sensory modality and class of actions, we draw examples from orienting movements in response to visual stimuli. Thus, consider the topographic maps of salience for visual orienting in colliculus or cortical areas devoted to gaze control such as frontal eye fields. We now suggest there are three broad reasons why a 'hotspot' of salience may have developed in the salience maps for orienting (with ensuing re-direction of gaze). First, features in the visual field are unpredicted and so, objects at that location in space are 'interesting' (surprising or novel). This results in action discovery via intrinsically-motivated behaviour. Second, the animal may have associated the object with primary reward (food, drink, etc) and can direct behaviour immediately to that reward. Third, the object may have been identified as task-relevant and, while not known to deliver reward immediately, is believed to be useful in guiding behaviour toward reward as a distal goal.

There is a fourth possibility which, perforce, implies a lack of a correspondence between salience and value. Thus, in the above situations, it is assumed that behaviour is goal directed, in the sense that it is *not* habitual (Balleine and Dickinson, 1998; Gurney et al., 2009b). Habitual actions are not contingent on the value of the current goal and are evoked instead by sensory context alone. For habits, therefore, we suppose that action selection is not conducted with reference to internal models of the world and prediction. However, since these concepts are our main concern, we limit ourselves henceforth to goal directed behaviour and now go on to explore the three components of salience discussed above.

4.2.1 Novelty and habituation: The first component of value defined above refers to unpredicted sensory features. We make links with the notion of habituation (a decline in response to a feature when that feature is no longer novel and has no rewarding consequences (Sokolov, 1963)), and identify the need to consider two kinds of prediction error: failure to predict correctly the occurrence of a feature, and failure to predict correctly its absence.

To simplify the arguments, we restrict ourselves at first to a single feature f, and consider evaluation over intervals during which the representations may be considered approximately constant. In this case the feature outcome is represented by a scalar y^+ and the prediction by another scalar y^*. One possible measure of the discrepancy, or error, s_P resulting from the prediction is

$$s_P = |y^+ - y^*| \,. \tag{8}$$

The notation s_P is supposed to suggest that this is a salience signal derived from a prediction. Poor prediction results in a larger value for s_P and there is then surprise or novelty, according to whether the feature f is (respectively) a phasic event, or more continuously available. As the prediction becomes more accurate, s_P becomes smaller, and the detection of the feature habituates. Habituation is therefore viewed as a consequence of development of the internal model responsible for y^*.

The right hand side of (8) is symmetric with respect to $\text{sign}(y^+ - y^*)$. There is however, a profound contrast in interpretation under sign change. For, if $y^+ > y^*$, this signifies novelty through a failure to adequately predict a feature's *presence*. However, if $y^+ < y^*$, this signifies novelty because of failure to predict an *absence* of the feature. Mechanistically, if we assume one of y^+, y^* supplies inhibitory and the other excitatory signals to a neural novelty detector, then it is not clear how the neuron can be excited by a net negative activity (if, say $y^+ < y^*$).

It is therefore more natural to split the salience due to novelty across two kinds of detectors: one which signals presence of unpredicted features, $s_{P\wedge}$, and one which signals absence of predicted features, $s_{P\vee}$.

$$\begin{aligned} s_{P\wedge} &= [y^+ - y^*]_{\geq 0} \\ s_{P\vee} &= [y^* - y^+]_{\geq 0} \ , \end{aligned} \tag{9}$$

where $[x]_{\geq 0}$ is a halfwave rectification function ($[x]_{\geq 0} = x$ if $x \geq 0$, and is 0 otherwise)[8]. Their separation makes sense, not just from a mechanistic point of view, but because different actions may be required according to whether novelty is determined via the absence or presence of a feature.

The existence of 'absence detectors' like $s_{P\vee}$ is more than just a theoretical possibility. In our account of prediction errors in section 5 we will introduce the idea that the lateral habenula may encode such signals (Matsumoto and Hikosaka, 2007). In the meantime we will focus on those like $s_{P\wedge}$, which we suppose are the norm. Our exposition has focussed on a single features, and if that feature refers to a phasic outcome, the error measure denotes surprise. In all other cases we are dealing with a salience map from novelty which will, in general, be a vector-valued function of time $\boldsymbol{s}_P$, obtained by considering all features and their time evolution. This may not simultaneously contain 'feature present' and 'feature absent' forms, but we can write, in general

$$\boldsymbol{s}_P = \boldsymbol{s}_{P\wedge} + \boldsymbol{s}_{P\vee} \ . \tag{10}$$

We now move on to consider other contributions to salience features which are intimately linked to goal directed behaviour. Essentially, our hypothesis is that behaviour directed at obtaining reward, or achieving goal-directed outcomes, occurs because the requests for the pertinent actions are generated by 'inverse models' from goals/rewards to action-requests. Specifically, action requests are

[8] This is equivalent to $xH(x)$ where $H(x)$ is the Heaviside function, but the notation in the text is more expedient here.

produced via a process of *sensitisation* of representations of salience features via top-down signals derived from the internal models.

4.2.2 Reward and salience: We start by considering behaviour (such as orienting, reach, and approach) which may be directly driven by reward-related stimuli comprising high level visual features or objects. Let $x_R(t)$ be the motivational level for seeking reward of type R (e.g. food). Then we suppose there is a learned, *inverse* or 'top-down' internal model[9] $\mathcal{I}_{\mathrm{R}}$ which causes x_R to sensitise representations of features $\boldsymbol{y}_R \in N_\Gamma$, related to objects that carry reward R (Ikeda and Hikosaka, 2003; Wurtz and Albano, 1980). Another term often used here is that the features have been *conditioned* through learning about the contingencies associated with the reward. This process is indeed one of 'sensitisation' of pre-existing representations derived through perception, rather than initiation of new ones (otherwise the agent would be hallucinating the presence of the associated features and objects). This is captured by putting $\mathcal{I}_{\mathrm{R}}(\boldsymbol{y}_R, x_R) = \boldsymbol{y}_R f(x_R)$, where $f(x_R)$ is a positive scalar, and supposing that the sensitised feature representation $\hat{\boldsymbol{y}}_R$ is given by

$$\begin{aligned} \hat{\boldsymbol{y}}_R &= \boldsymbol{y}_R + \mathcal{I}_{\mathrm{R}}(\boldsymbol{y}_R, x_R) \\ &= \boldsymbol{y}_R(1 + f(x_R)) \ . \end{aligned} \tag{11}$$

If the sensitisation is transmitted to salience for action, it therefore contributes an amount

$$\boldsymbol{s}_R = \mathcal{I}_{\mathrm{R}}(\boldsymbol{y}_R, x_R) = \boldsymbol{y}_R f(x_R) \tag{12}$$

to that salience.

Often this process may take place in two stages. First, spatially invariant representations of entire objects are sensitised; this kind of learning may take place at any time in the agent's lifetime. Then, second, these representations are transformed to topographic representations of salience that request actions such as visual orienting, reach, and approach (Connor et al., 2004; Cope et al., 2009; Thompson et al., 2005). These transforms might be an integral part of the sensorimotor system and are acquired in early development of the animal.

4.2.3 Goal or outcome-based salience: We now turn to models that go to the heart of our exposition: those dealing with the relation between learned actions and outcomes. In this case we want top-down information about desired outcomes or goals to become manifest at the level of motor action selection. Each such outcome may be a stage on the way to some final goal (acquiring food may require going outside, making a car journey, going shopping etc) but we will refer to each sub-task as a goal in its own right.

The argument proceeds in an analogous way to that for reward, but our starting point here are representations of goal $\boldsymbol{y}_G \in N_\Gamma$. These may include features about objects, whole scenes, spatial relationships etc. In any case, we

[9] Some researchers call this a *competence* model **??**

suppose there are inverse models $\mathcal{I}_G$ providing contributions $\boldsymbol{s}_G$ to action salience via mappings of the form

$$\mathcal{I}_G[\boldsymbol{y}_G] = \boldsymbol{s}_G \ . \tag{13}$$

Once again, (just as for $\mathcal{I}_R$), $\mathcal{I}_G$ may be composed of multiple stages; from 'high level' representations of outcome situations to objects to salience maps for motoric acts such as reach, gaze, etc. However, we also admit the possibility that some instances of $\mathcal{I}_R$ map one high level representation to another in generating sequence behaviour, in which one task outcome generates the next desired outcome. These high level representations, occurring in limbic and associative structures, may also be subject to selection under basal ganglia control (Yin and Knowlton, 2006).

Notice that the action-outcome model $\mathcal{I}_G$ is a model of how the action for the outcome is invoked; that is, how a representation of the desired outcome in $\boldsymbol{y}_G$, is transformed into an action request s_G. Thus, $\mathcal{I}_G$ is a model of *deployment*, not *prediction*, which is the role of forward models such as $\mathcal{I}_P$ described in section 3.3. These two models may be intimately related. The prediction (e.g. my computer will wake from sleep) derived from a context (sat at the computer and pressing the space bar) under $\mathcal{I}_P$ may, itself, correspond to a desired outcome. Thus, under a related model $\mathcal{I}_G$, this outcome (computer waking from sleep) must elicit an action request (press space bar) which was part of the original context for $\mathcal{I}_P$. It is in this sense that one model is the inverse of the other. The main point to note is that two distinct models are required for action-outcome learning.

4.3 Combining salience contributions

We now synthesise the results of the previous subsections on the definition of salience contributions to define the overall salience $\boldsymbol{s}$ for goal directed action. We assume that salience, within a particular salience map or brain structure, is additive with respect to its various contributions components $\boldsymbol{s}_P, \boldsymbol{s}_R, \boldsymbol{s}_G$ defined via (10), (12), (13) respectively.

$$\begin{aligned} \boldsymbol{s} &= \mathcal{E} \text{ using } \mathcal{I}_P + \mathcal{I}_{\mathrm{R}}(\boldsymbol{y}_R, x_R) + \mathcal{I}_G[\boldsymbol{y}_G] \\ \boldsymbol{s} &= \underset{\text{novelty and surprise}}{\boldsymbol{s}_P} + \underset{\text{immediate reward}}{\boldsymbol{s}_R} + \underset{\text{goal/outcome}}{\boldsymbol{s}_G} , \end{aligned} \tag{14}$$

where the first equation indicates the internal models used and the second, the notation we use for salience.

5 Sensory prediction errors

Errors initiated by surprise at phasic events are at the centre of our exegesis of the role played by phasic dopamine. Here, we build on the work in section 4.2.1 to develop a more detailed account of such errors. The implications for habituating to, and prediction of, these phasic events is described, but the full implications for learning internal models is deferred until section 6.1.

Sensory prediction errors and their implementation in the brain: It is often useful to allow error signals for learning to take positive and negative values so that they can force increments and decrements in model parameter respectively. In our current context, the target is what actually transpires in the form of the outcome $\boldsymbol{y}^+$, and the prediction is $\boldsymbol{y}^*$. That is we define a *sensory prediction error* Δ by

$$\Delta = \boldsymbol{y}^+ - \boldsymbol{y}^* \ . \tag{15}$$

We now specialise to the case based on surprise, where the outcome is encoded by a single feature y^+_ϕ of a phasic event and the scalar prediction y^*_ϕ is a result of a prediction model $\mathcal{I}_{P_\phi}$ (see section 3.3) . Then (15) becomes

$$\Delta_\phi = y^+_\phi - y^*_\phi \ . \tag{16}$$

A brain structure which has been implicated in such processing is the superior colliculus (SC) (Wurtz and Albano, 1980). However, the SC cannot encode Δ_ϕ itself since it delivers only non-negative salience signals $s_{P\wedge}$

$$s_{P\wedge} = [y^+_\phi - y^*_\phi]_{\geq 0} \ , \tag{17}$$

where we have used the notation (9) to indicate this is a detection of surprise through detection of unpredicted features. In order to compute Δ_ϕ, and make use of existing resources such as SC, we require a complementary signal

$$s_{P\vee} = [y^*_\phi - y^+_\phi]_{\geq 0} \ , \tag{18}$$

for then

$$\Delta_\phi = s_{P\wedge} - s_{P\vee} \ . \tag{19}$$

Our hypothesis (discussed in detail in the companion chapter **??**) is that phasic dopamine encodes Δ_ϕ in equations (16) and (19). Thus, positive phasic changes in dopamine indicate a failure to predict the occurrence of the feature signalled by y^+_ϕ, whereas negative changes ('dips') in dopamine signal the absence of the feature when it was predicted. Physiologically, it will require the combination of an excitatory signal $s_{P\wedge}$, and an inhibitory one $s_{P\vee}$. We have argued that SC generates $s_{P\wedge}$, and there is now evidence that SC directly excites midbrain dopamine neurons (Comoli et al., 2003; Dommett et al., 2005). Recent work by Matsumoto and Hikosaka (2007) has shown that the lateral habenula is able to signal negative prediction errors in the phasic dopamine signal, and that it acts on dopamine neurons in an inhibitory way. It is therefore a candidate for encoding $s_{P\vee}$.

Reward sensitisation of sensory prediction: In spite of our interpretation of phasic dopamine in terms of a sensory prediction error, there is evidence that the the strength of the phasic dopamine signal can, under certain circumstances, be modulated by the precise reward value supplied by a stimulus (for review see Schultz, 2010). In particular, Fiorillo et al. (2003); Tobler et al. (2005) have

shown that, with well trained animals, graded reward probabilities associated with unpredictable phasic events produced phasic dopamine responses which reflected the expected amount of reward. This is often cited as strong evidence that phasic dopamine is signalling reward-prediction error. However, we think the situation is rather subtle and can be incorporated into our scheme as follows.

Recall from section 4.2.2 that salience may be augmented by sensitisation through reward. Superior colliculus may be sensitised in this way, thereby overcoming complete habituation to a predicted stimulus in the presence of reward association with that stimulus (Ikeda and Hikosaka, 2003) . Therefore, using (12)

$$s_{P\wedge} = [y_\phi^+(1 + f(x_R)) - y_\phi^*]_{\geq 0} \, . \tag{20}$$

This suggests that the sensory error in (16) should be modified to

$$\Delta_\phi = y_\phi^+(1 + f(x_R)) - y_\phi^* \, . \tag{21}$$

However, for this to hold requires that the habenula signal is modified as well. That is

$$s_{P\vee} = [y_\phi^* - y_\phi^+(1 + f(x_R))]_{\geq 0} \, . \tag{22}$$

The lateral habenula does indeed show reward modulated activity in which absence/presence of reward facilitates/inhibits activity (Matsumoto and Hikosaka, 2007). We therefore take (21) as a more complete definition of sensory error which is able to account for the reward modulation of phasic dopamine.

Notice that Δ_ϕ is still a *sensory* prediction error - there is no mention of a difference between observed *reward* as such, and its prediction. It is true that y_ϕ^+ has been sensitised or 'tagged' with reward (if the term $f(x_R)$ is non-zero), and so is, itself, a predictor of reward - but the basic signals here are a sensory feature representation, y_ϕ^+, and its prediction y_ϕ^*. Moreover, the development of sensitisation via $f(x_R)$ requires learning (under a model $\mathcal{I}_\mathrm{R}$) which may require massive exposure to the reward stimulus and (simultaneously) the feature y_ϕ^+. This is a hallmark of many laboratory experiments but not so in many natural situations in which we may therefore expect $f(x_R) \approx 0$ (Redgrave et al., 2008). In these cases Δ_ϕ is most certainly a sensory (only) prediction error and so we argue that the emphasis on the sensory encoding/error-production is more general. However, the fact that reward should find a role in action discovery is not surprising; it is reasonable to suppose that actions delivering reward should be learned more quickly and/or reliably.

6 A framework for intrinsically-motivated learning

Here we bring together the threads we have developed to give an account of action discovery as intrinsically-motivated learning driving the development of internal models of prediction and action-outcome contingencies.

Learning to 'listen' to action requests: The first process we consider is one in which basal ganglia learns to encode the new action. In (7) salience was defined via an action request – a representation $\boldsymbol{s}^{(a)}$, with components $s_f^{(a)}$, for an action $\boldsymbol{\alpha}$. In the brain, this request is first 'filtered' by neurons in striatum[10] before taking part in a selection process in the basal ganglia for behavioural expression. The neuronal response in striatum $\hat{S}_\alpha(t)$ is given by

$$\hat{S}_\alpha = \sum_f w_f s_f^{(a)} \,, \tag{23}$$

where w_f are the synaptic weights on striatal neurons receiving cortical inputs $s_f^{(a)}$. The notion of filtering here is inspired by the the formal equivalence of (23) and the convolution sums defining finite impulse response (FIR) filters. Here, the cortico-striatal weights w_f play the role of filter coefficients acting on signals $s_f^{(a)}(t)$. This view highlights the fact that a key part of biological action discovery is the adjustment or 'tuning' of the striatal filter (i.e. the weights) so that basal ganglia can 'listen' effectively to the action request.

In vector terms, if $\boldsymbol{w}$ is the weight vector comprising components w_f, the right hand side of (23) is the inner product $\boldsymbol{w} \cdot \boldsymbol{s}^{(a)}$ which takes its maximal value when both vectors are in the same 'direction'. Striatal tuning in action discovery is therefore one of 'weight vector rotation'.

Dopamine is known to facilitate cortico-striatal plasticity using learning rules that are otherwise broadly Hebb-like (Reynolds and Wickens, 2002). Now suppose that the most recently selected action $\boldsymbol{\alpha}$ (associated with $\hat{S}_\alpha$) causes a sensory prediction error Δ_ϕ due to unexpected phasic change in the environment. The resultant phasic dopamine is able to reinforce the match between $\mathbf{w}$ and $\mathbf{s}^{(a)}$ as long as the pattern of neural activity in $\boldsymbol{s}^{(a)}$ does not decline substantially before delivery of the dopamine signal [11].

Repetition bias: The increased match between $\boldsymbol{w}$ and $\boldsymbol{s}^{(a)}$ promoted by phasic dopamine will causes an increase in $\hat{S}_\alpha$, as long as the relevant context contributing to the action request is maintained. This will, in turn, be reflected in a change in the animal's policy as an increase in the probability $\pi(\boldsymbol{\alpha})$ of selecting $\boldsymbol{\alpha}$. Eventually any contextual salience that may have sustained $\boldsymbol{s}^{(a)}$ will decline as habituation occurs to the contextual features in the action request. The striatal response $\hat{S}_\alpha$ becomes small again, and so too therefore, does $\pi(\boldsymbol{\alpha})$. We refer to the temporary increase in $\pi(\boldsymbol{\alpha})$ as *repetition bias*.

6.1 Learning internal models

Action-outcome discovery would appear to involve establishing two internal models: a (forward) prediction model $\mathcal{I}_P$, and an (inverse) deployment model $\mathcal{I}_G$.

[10] The striatum is the main input nucleus of the basal ganglia

[11] Additionally, this may be supported by some kind of *eligibility trace* associated with $s^{(a)}$ which dopamine acts upon (Gurney et al., 2009a)

Further, we have postulated a division of prediction so that $\mathcal{I}_P$ is a combination of a model $\mathcal{I}_{P_\phi}$, based on surprise, and another dealing with the remaining structural complexity of outcome. How does the animal learn these models? As described below, repetition bias is a key driver for these processes.

Action-outcome pairings: During action discovery, the increase in the probability $\pi(\boldsymbol{\alpha})$ for action $\boldsymbol{\alpha}$ causes representations of context $\boldsymbol{y}^-$ (containing $\boldsymbol{a}$) and its consequent outcome $\boldsymbol{y}^+$, to be present with an increased probability. The sustained presence of such representations will induce plasticity in the relevant brain areas responsible for learning the associations implied in $\mathcal{I}_P$ and $\mathcal{I}_G$.

The dynamics of repetition bias are governed by learning of models like $\mathcal{I}_{P_\phi}$. However, learning in the other models may not necessarily follow an identical time course, which poses the possibility of incomplete learning of all required models. This problem may be overcome during further behaviour by the animal in ways which rely on the relationship between model representations. Suppose, for example, that $\mathcal{I}_G$ has been inadequately learned, resulting in performance errors for the action involved in this model. This will cause unexpected phasic outcomes which will, in turn, incur sensory prediction errors. These will drive a further round of repetition bias resulting in action discovery refinement.

Models without action contingency: We now consider the learning of models $\mathcal{I}_P$ which do not contain phasic-related components. These are non-operant models which might, for example, concern the representation of environmental elements in relation to each other. Novelty (not simply surprise) may then arise if new elements are found where they are unexpected ('why is there a football shirt in my office!?') or there is an absence of expected elements ('where has my computer gone!?'). The novelty in these situations will induce high salience for investigative behaviours directed at the novel elements (or the space they previously occupied). These actions will be maintained as long as the novelty remains unexplained. Two possibilities can now occur. First, there is no unexpected causal outcome associated with the investigative behaviour (picking up the football shirt does not, for example, turn the room lights on). In this case, novelty is reduced by developing explanatory accounts of the situation which might predict how the situation arose. Second, there *is* an unexpected outcome upon investigation, in which case we are now back in the action-outcome scenario described above. That is, a new prediction model presents itself for learning which *does* have a phasic component of the form $\mathcal{I}_{P_\phi}$. Note that, this new model is separate from the non-operant, non-phasic model which induced the investigative behaviour in the first place.

The need for prediction and inverse models: One question which arises here is why the animal needs an inverse model $\mathcal{I}_G$ for deploying the learned action-outcome element, *and* a prediction model? First, acquisition of the submodel of prediction $\mathcal{I}_{P_\phi}$ instigates repetition bias; phasic sensory prediction error

is a trigger for learning the other models and, just as important, a lack of error (through prediction) terminates learning. Second, suppose the full prediction model $\mathcal{I}_P$ was not completely developed. This would give rise to a constant stream of salience due to (general) novelty, and the agent would be continually drawn to stimuli that it had seen countless times. This would make it difficult for the animal to engage in purposeful behaviours driven, not by novelty, but by salience derived from $\mathcal{I}_G$ and $\mathcal{I}_R$. Prediction therefore prevents 'attentional deficit' behaviour which would constitute a continuous stream of exploration without pursuit of specific goals.

7 Discussion

7.1 Summary of main ideas

We have introduced a formal framework for considering the relationships between animals, their environment, their internal neural representations of that environment, and their behaviour. In particular, we define relations between a context, an action in that context, and a causally related outcome. This framework is a functional one, considering entire path histories or trajectories of state variables

Our notion of prediction is based on internal neural models of causal relations which become manifest ('make predictions') by expressing signals at their output. The resulting neural representations interact with sensory processes to drive habituation and sensitisation.

Salience maps were defined via the components of sensorimotor representations defining action policy, and action requests comprise a subset of features in a salience map. Salience (as a scalar value) is defined as the sum of feature values in an action request. We made the hypothesis that (for non-habitual actions) salience may be indicative of the value of an action and that salience comprises three components: feature novelty, sensitisation due to reward, and task-driven (outcome) priming. Sensitisation is driven by inverse models of reward conditioning and action-outcome contingency like $\mathcal{I}_R$ and $\mathcal{I}_G$ in (12) and (13) respectively.

An analysis of novelty with a simple example led to the idea that habituation comprises inhibition of the sensory representation by the prediction. The example also led naturally to the existence of two opponent novelty detectors (for feature presence and absence). A sensory prediction error was defined as the difference between the opponent novelty detectors (see equation (19)). For the special case of surprise detection, we putatively identified the feature-presence/absence detectors as superior colliculus and lateral habenula respectively. Phasic dopamine was then identified as the sensory prediction error corresponding to surprise. In addition, the ability of reward to sensitise the colliculus and habenula led to the idea that phasic dopamine can be modulated by reward (as observed experimentally), but remains best described as a sensory prediction error.

Finally we have shown how learning of the internal models may proceed under temporary changes in policy – repetition bias – and that this learning could be robust via continued interaction with the environment.

7.2 Implications for intrinsic motivation

We started by noting that action-outcome skill acquisition is a hallmark of many formulations of what constitutes intrinsically-motivated learning. In this sense our framework lies at the heart of the field. However, in the survey of Oudeyer and Kaplan (2007), they admit the possibility that some intrinsically-motivated learning deals with 'passive' observation of the environment and learning intra-world contingencies rather than agent-environment ones. This process was discussed in section 6.1 ('Models without action contingency '). We therefore agree that the learning of novelty-driven prediction models, without associated inverse models, is a suitable candidate for the tag of 'intrinsically-motivated learning'. The common ground here is that the learning is promoted by novelty (taken to subsume surprise as special case). More specifically we might argue that it is a successful *reduction* in novelty, through information compression in prediction-model construction, which is the key characteristic. This idea has been explored more fully by Schmidhuber (Schmidhuber, 2009).

If novelty-induced behaviour is a hallmark of intrinsically-motivated learning, our framework suggests a quantitative formal definition. Thus, we *could* define the level of intrinsic motivation according to the relative contribution of novelty salience, $\boldsymbol{s}_P^{(a)}$, in the action request, $\boldsymbol{s}^{(a)}$, for the current action; use the dissection of $\boldsymbol{s}^{(a)}$ given in (14) and compute $||\boldsymbol{s}_P^{(a)}||/||\boldsymbol{s}^{(a)}||$. This definition is given in the spirit of ontology construction (see the Introduction) – it is a *plausible* formal definition of 'intrinsic motivation' without laying claim to be a 'truth'.

While this formal definition of intrinsic motivation is plausible, it may not satisfy other interpretations. Intrinsically-motivated behaviour has also been described as "doing something for its inherent satisfactions rather than for some separable consequence" (Ryan and Deci, 2000). We take 'separable consequences' here to mean external primary rewards. The action discovery scenario, driven as it is (in general) by biologically neutral events caused by the agent is, therefore intrinsically-motivated from this perspective too. It might be claimed that the unexpected outcomes observed during action discovery constitute such external rewards but, as we have argued elsewhere (Redgrave and Gurney, 2006), this does not adhere to conventional notions of biologically defined primary reward. Further, while it is difficult to determine whether action discovery is being conducted subjectively 'for its own sake', we know from experience that this is, indeed, often the case with accompanying feelings of 'curiosity' and 'satisfaction'. Finally, we argue that, irrespective of any experience of 'curiosity', the brain mechanisms invoked during internal model building of causality, contingent on unexpected outcomes, will be substantially the same as those deployed when 'curiosity' might be established as a causative factor. We therefore contend that a mechanistic account of action discovery can shed light on intrinsically-motivated learning in general.

7.3 Prospectus

The ideas presented here do not specify any particular detailed model of intrinsically-motivated learning. Rather, they provide a framework for computational modelling work. Our thinking here was informed by neurobiological relevance and so we would hope that this will facilitate the future construction of biologically plausible models. The processes and mappings demonstrated here, when concatenated together, may also help specify future ground-plans for functional brain architectures. The specific formalism of action-discovery ontology presented here may be incomplete, and some features remain less well-defined than others. More radically, others may disagree altogether with our approach and demand an alternative formalism, but we would welcome this if it engages with the effort of building an ontology of action selection and discovery, and more widely, of intrinsically-motivated learning.

Notwithstanding the general scope of this work, we have dealt with an analysis of sensory prediction errors and phasic dopamine at a level which makes contact with specific neural functions – see equations 20, 21,22. Establishing the accuracy of these putative functions offers an immediate programme of work for modelling these the colliculus, habenula and mid-brain dopamine circuits.

Acknowledgements

Written while the authors were in receipt of research funding from The Wellcome Trust, BBSRC and EPSRC. IM-CLeVeR is supported by the European Commission under the FP7 Cognitive Systems, Interaction, and Robotics Initiative, grant no. 231722.

Bibliography

Allport, A., Sanders, H., and Heuer, A. (1987). Selection for action: some behavioural and neurophysiological considerations of attention and action. In *Perspectives on perception and action*. Lawrence Erlbaum Associates Inc., Hillsdale, NJ.

Baldi, P. and Itti, L. (2010). Of bits and wows: A bayesian theory of surprise with applications to attention. *Neural Networks*, 23(5):649–666.

Balleine, B. W. and Dickinson, A. (1998). Goal-directed instrumental action: contingency and incentive learning and their cortical substrates. *Neuropharmacology*, 37(4-5):407–419.

Barto, A., Singh, S., and Chentanez, N. (2004). Intrinsically motivated reinforcement learning. In *18th Annual Conference on Neural Information Processing Systems (NIPS)*. Vancouver.

Cisek, P. (2007). Cortical mechanisms of action selection: the affordance competition hypothesis. *Philos.Trans.R.Soc.Lond B Biol Sci*, 362(1485):1585–1599.

Cisek, P. and Kalaska, J. (2010). Neural mechanisms for interacting with a world full of action choices. *Annual review of neuroscience*, 33:269298.

Comoli, E., Coizet, V., Boyes, J., Bolam, J., Canteras, N., Quirk, R., Overton, P., and Redgrave, P. (2003). A direct projection from superior colliculus to substantia nigra for detecting salient visual events. *Nat Neurosci*, 6(9):974–980.

Connor, C. E., Egeth, H. E., and Yantis, S. (2004). Visual attention: Bottom-Up versus Top-Down. *Current Biology*, 14(19):R850–R852.

Cope, A., Chambers, J., and Gurney, K. (2009). Object-based biasing for attentional control of gaze: a comparison of biologically plausible mechanisms. *BMC Neuroscience*, 10(Suppl 1):P19.

Dommett, E., Coizet, V., Blaha, C., Martindale, J., Lefebvre, V., Walton, N., Mayhew, J., Overton, P., and Redgrave, P. (2005). How visual stimuli activate dopaminergic neurons at short latency. *Science*, 307(5714):1476–1479.

Fiorillo, C. D., Tobler, P. N., and Schultz, W. (2003). Discrete coding of reward probability and uncertainty by dopamine neurons. *Science*, 299(5614):1898.

Friston, K. (2005). A theory of cortical responses. *Philosophical Transactions of the Royal Society B: Biological Sciences*, 360(1456):815–836.

Gene Ontology Consortium (2001). Creating the gene ontology resource: design and implementation. *Genome Research*, 11(8):1425–1433.

Gruber, T. (1992). A translation approach to portable ontology specification. http://www-ksl.stanford.edu/kst/what-is-an-ontology.html.

Gurney, K., Humphries, M., and Redgrave, P. (2009a). Cortico-striatal plasticity for action-outcome learning using spike timing dependent eligibility. *BMC Neuroscience*, 10(Suppl 1):P135.

Gurney, K., Hussain, A., Chambers, J., and Abdullah, R. (2009b). Controlled and automatic processing in animals and machines with application to autonomous vehicle control. In *Controlled and automatic processing in animals and machines with application to autonomous vehicle control*, volume 5768 LNCS of *Lecture Notes in Computer Science*, pages 198–207. Springer.

Ikeda, T. and Hikosaka, O. (2003). Reward-dependent gain and bias of visual responses in primate superior colliculus. *Neuron*, 39(4):693–700.

Körding, K. P. and Wolpert, D. M. (2006). Bayesian decision theory in sensorimotor control. *Trends in Cognitive Sciences*, 10(7):319–326.

Marr, D. and Poggio, T. (1976). From understanding computation to understanding neural circuitry. Technical report, MIT AI Laboratory.

Matsumoto, M. and Hikosaka, O. (2007). Lateral habenula as a source of negative reward signals in dopamine neurons. *nature*, 447(7148):1111–1115.

Oudeyer, P. and Kaplan, F. (2007). What is intrinsic motivation? a typology of computational approaches. *Frontiers in Neurorobotics*, 1:6. PMID: 18958277.

Poggio, T. and Koch, C. (1985). Ill-posed problems in early vision: from computational theory to analogue networks. *Proceedings of the Royal society of London. Series B. Biological sciences*, 226(1244):303.

Ranganath, C. and Rainer, G. (2003). Neural mechanisms for detecting and remembering novel events. *Nat Rev Neurosci*, 4(3):193–202.

Redgrave, P. and Gurney, K. (2006). The short-latency dopamine signal: a role in discovering novel actions? *Nat. Rev. Neurosci.*, 7(12).

Redgrave, P., Gurney, K., and Reynolds, J. (2008). What is reinforced by phasic dopamine signals? *Brain Res Rev.*, 58(2):322–339.

Redgrave, P., Prescott, T., and Gurney, K. (1999). The basal ganglia: a vertebrate solution to the selection problem? *Neuroscience*, 89:1009–1023.

Reynolds, J. N. J. and Wickens, J. R. (2002). Dopamine-dependent plasticity of corticostriatal synapses. *Neural Networks: The Official Journal of the International Neural Network Society*, 15(4-6):507–521.

Ryan, R. M. and Deci, E. L. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions* 1. *Contemporary educational psychology*, 25(1):5467.

Schleidt, M. and Kien, J. (1997). Segmentation in behavior and what it can tell us about brain function. *Human Nature*, 8(1):77–111.

Schmidhuber, J. (2009). Driven by compression progress: A simple principle explains essential aspects of subjective beauty, novelty, surprise, interestingness, attention, curiosity, creativity, art, science, music, jokes. In *Anticipatory Behavior in Adaptive Learning Systems*, pages 48–76.

Schultz, W. (2010). Dopamine signals for reward value and risk: basic and recent data. *Behavioral and Brain Functions*, 6(1):24.

Schultz, W., Dayan, P., and Montague, P. (1997). A neural substrate of prediction and reward. *Science*, 275:1593–1599.

Snyder, L. H., Batista, A. P., and Andersen, R. A. (1997). Coding of intention in the posterior parietal cortex. *Nature*, 386(6621):167–170.

Sokolov, E. N. (1963). Higher nervous functions: The orienting reflex. *Annual Review of Physiology*, 25(1):545–580.

Sutton, R. and Barto, A. (1998). *Reinforcement Learning : An Introduction*. MIT Press, Cambridge, MA.

Thompson, K. G., Bichot, N. P., and Sato, T. R. (2005). Frontal eye field activity before visual search errors reveals the integration of Bottom-Up and Top-Down salience. *J Neurophysiol*, 93(1):337–351.

Timberlake, W. and Lucas, G. A. (1985). The basis of superstitious behavior: chance contingency, stimulus substitution, or appetitive behavior? *Journal of the Experimental Analysis of Behavior*, 44(3):279.

Tobler, P., Fiorillo, C., and Schultz, W. (2005). Adaptive coding of reward value by dopamine neurons. *Science*, 307(5715):1642.

Tolman, E. (1948). Cognitive maps in rats and men. *Psychological review*, 55(4):189.

Wurtz, R. H. and Albano, J. E. (1980). Visual-motor function of the primate superior colliculus. *Annual Review of Neuroscience*, 3(1):189226.

Yin, H. H. and Knowlton, B. J. (2006). The role of the basal ganglia in habit formation. *Nature Reviews. Neuroscience*, 7(6):464–476.