Hedonic Value: Enhancing Adaptation for Motivated Agents

Ignasi Cos[*‡] Lola Cañamero[‡] Gillian M. Hayes[†] Andrew Gillies[**]

[†] *Institute of Perception, Action and Behaviour,*
School of Informatics, University of Edinburgh,
Informatics Forum, 10 Crichton Street,
Edinburgh, EH8 9AB, Scotland, UK.
[‡]*Adaptive Systems Research Group,*
School of Computer Science, University of Hertfordshire,
Hatfield, Herts, AL10 9AB, UK.
[**] *Institute of Artificial and Neural Computation,*
School of Informatics, University of Edinburgh,
5 Forrest Hill, Edinburgh, EH1 2QL, Scotland, UK.

Abstract

Reinforcement learning (RL) in the context of artificial agents is typically used to produce behavioural responses as a function of the reward obtained by interaction with the environment. When the problem consists of learning the shortest path to a goal, it is common to use reward functions yielding a fixed value after each decision, for example a positive value if the target location has been attained and a negative one at each intermediate step. However, this fixed strategy may be overly simplistic for agents to adapt to dynamic environments, in which resources may vary from time to time. By contrast, there is significant evidence that most living beings internally modulate reward value as a function of their context to expand their range of adaptivity. Inspired by the potential of this operation, we present a review of its underlying processes and we introduce a simplified formalisation for artificial agents. The performance of this formalism is tested by monitoring the adaptation of an agent endowed with a model of motivated actor-critic, embedded with our formalisation of value and constrained by physiological stability, to environments with different resource distribution. Our main result shows that the manner in which reward is internally processed as a function of the agent's motivational state, strongly influences adaptivity of the behavioural cycles generated and the agent's physiological stability.

Keywords: Hedonic Value, Motivation, Reinforcement Learning, Actor-Critic, Grounding.

[*]Current Address: ISIR, Universite Pierre et Marie Curie, 4 Place Jussieu, 75005 Paris, France. Email: ignasi.cos@isir.upmc.fr

1 Introduction

Although it is possible to learn efficient behavioural sequences in the context of reinforcement learning (RL) by propagating value backwards to previously visited states (Sutton and Barto, 1998), this procedure by itself may fall short in the more demanding context of a motivated agent that has to adapt to and survive in a situated, dynamic environment. If an agent has several needs, a reasonable strategy to survive may consist of learning behavioural patterns that prioritise the compensation of an internal resource over another, as a function of what the environment affords and of the internal rate of consumption of each resource, in a similar fashion to most animals. However, to reach this behavioural flexibility via RL in changing environments may demand mechanisms that modulate the criterion of internal assessment to influence behaviour in an adaptive manner. An example of this is observed in the the level of pleasure associated to food consumption, which varies a great deal as a function of the level of hunger (Shizgal, 1997). While an empty stomach typically reinforces the pleasure and urge with which a meal is consumed, this pleasure progressively diminishes, sometimes even reverts, as one gradually satiates. Although the modulation of value by the physiological state is a well known phenomenon, often studied in the context of stimulus devaluation (Dittrich and Klauer, 2011; Eder and Rothermund, 2008), we still lack a unified view of its underlying mechanisms that facilitate devising procedures to modulate behaviour in RL for artificial agents (Konidaris and Barto, 2006).

Related to this, a tremendous amount of studies in neuro-physiology have devoted to investigate the underlying mechanisms of *hedonic value* (HV) (Smith et al., 2011; Grabenhorst et al., 2008). Loosely speaking, HV refers to the subjective, internally perceived value resulting from any given interaction. It may influence decision-making and adaptation, and may vary as a function of the state of the animal and of its perception of the environment. Although hedonic phenomena encompass a large number of brain areas whose implication is under current investigation (Rolls, 2004; Damasio, 2000), there is some consensus that the neural encoding of HV involves recurrent projections between ventral striatum, amygdala and prefrontal areas (Alexander et al., 1990), such as orbito-frontal, anterior cingulate or dorso-lateral pre-frontal cortices (Reynolds and O'Reilly, 2009; Hazy et al., 2007; Tanji and Hoshi, 2001), as well as the dorso-lateral striatum (Guitart-Masip et al., 2011; Rolls, 2004). We dedicate the next section to review some of this evidence, which inspired us to propose an artificial implementation of a modulatory mechanism of value, based on these principles. However, our goal is to investigate the contribution to adaptation of a mechanism of value modulation grounded on the situated nature of our agent (Cañamero, 1997; Wilson, 1991), rather than providing a descriptive model of the parts of the brain involved in this process. This mechanism would be tailored to each sort of stimulus, experience or sensory modality that biases decision-making across candidate options and its associated reinforcement learning process (McClure et al., 2003). Our proposal may be therefore viewed as an extension of the ecological relationship between the agent and its environment (Gibson, 1986; Pfeifer, 1996), to include the dynamics of each internal variable, the resources offered by the environment to replenish them and the range of policies the agent can possibly learn. We assume that value modulation varies within a specific range for each resource, expressed along a single scale that makes it possible to reconcile assessments across dissimilar options (Grabenhorst and Rolls, 2011; Gurney et al., 1998).

To test our formulation of hedonic value, we expanded a motivated architecture (Cos et al., 2010), which initially focused on the dynamics of perception only, with an actor-critic algorithm to learn decision-making strategies (Sutton and Barto, 1981). Therein, our neuro-inspired notion of subjective assessment is implemented as a value function, and is tested by learning behavioural responses to different stimuli and physiological states in a manner compliant with the hypothesis of phasic dopamine as an error signal (Khamassi et al., 2005; McClure et al., 2003; Schultz et al., 2000; Houk

et al., 1995). The adaptivity of the agent has been assessed in terms of its physiological stability (Ashby, 1965), as a function of the response to changes in the availability of resources of the environment. The results show the influence of the subjective interpretation of reward during adaptation, and importantly, the dependence of the behavioural patterns and of the agent's physiological stability on the agent's subjective view of the environment.

2 Background and Related Research

Value-based decisions are based on differences of expected value (Wallis, 2012; Kennerley et al., 2011; Wallis and Miller, 2003), and their related learning algorithms are sensitive to the matching between expected reward and outcome value (McClure et al., 2003; Houk et al., 1995). However, differences in value between options are not solely dependent on actions and stimuli, they include as well a component of subjective perception, constrained by the specifics of an individual and its relationship with the environment (Pfeifer, 1996; Gibson, 1986). Here we describe this additional dimension to enhance adaptivity.

In general, most attention in RL has been devoted to structure the problem in a tractable manner, typically by devising an efficient hierarchy of motor primitives; either imposed by prior design constraints (Matarić and Brooks, 1990) or by self-adaptation on the basis of sensorimotor interaction (Toussaint, 2003). The result is a dramatic decrease of the learning interval as a result of a reduced dimensionality of the RL state-space and a more parsimonious behaviour. In a complementary fashion, some of the architectures best suited to reproduce aspects of animal behaviour have incorporated the dynamics of interaction between the animal and the statistics of the environment to the operation of their adaptation mechanisms (Konidaris and Hayes, 2005; Matarić and Brooks, 1990). As an extension to these, we propose to endow motivated agents with an additional element of behaviour control, namely a mechanism of *internal assessment*, constrained by the agent's condition of situatedness (Cos et al., 2010; Velásquez, 1998; Cañamero, 1997; Wilson, 1991). In the same manner that both perception and value (or reward) based decision-making are processes influenced by internal physiological processes, our proposal of subjective assessment should be founded on the specifics of valuation of physiological effect as a result of behaviour executions (Dickinson and Balleine, 2001), organised in cycles of sensorimotor interaction (McFarland and Sibly, 1975; Seth, 2000). In a continuous fashion, drives may express an urge for action (Hull, 1943) and incorporate internal information to give rise to the agent's motivations, which may exert a direct influence on the saliency of the agent's behaviours, as part of the internal-external dynamics seeking physiological stability (Ashby, 1965). Motivation, defined in several contexts and disciplines (McDougall, 1913; Freud, 1940; Tinbergen, 1951; Lorenz, 1966), often with different nuances, always conveyed the role of "a substance, capable of energising behaviour, held back in a container and subsequently released in action" (Hinde, 1971, 1960), hence relating physiology to behaviour.

Intrinsically, the notion of value includes an objective and a subjective component. Objective is the part independent of the physiological state, subjective is the part dependent on it (Grabenhorst and Rolls, 2011; Conover et al., 1994; Conover and Shizgal, 1994). Revisiting the adaptation processes implemented by some of the aforementioned robotic architectures, most of them could be classified as based on objective value, as the notion of motivation does not explicitly incorporate motivation into their modulation of value. An exception is however the Hullian drive based architecture of Konidaris and Barto (2006), which includes both the expression of intended purpose of any motivated architecture, combined with the learning of priorities by reinforcement learning. As a novelty, this architecture includes a procedure of internal modulation that biased its internal motivations as a function of the environment statistics, e.g., scarce resources result an overexpressed related motivation, showing that this may lead to a better adaptation. Likewise, Coninx et al. (2008)

extended Konidaris' architecture with a model of the basal ganglia (Girard et al., 2008) that arbitrates between several actions, showing that two different policies arise when there are different environments. Although these architectures do not explicitly address the notion of hedonic value by using fixed reward formulae, they strongly suggest that mechanisms of internal modulation do influence the overall behaviour of the robot. Along a different line of research, Damoulas et al. (2005) proposed a RL context wherein genetic algorithms were used to evolve an interpretation of physiological effect in the form of Q-values (Sutton and Barto, 1998), showing that only a small fraction of agents yielded physiologically stable behaviour to be transferred to the next generation. In conclusion, Damoulas et al. (2005) showed that adaptivity depends on the manner reward is assessed over generations. In a complementary fashion to these studies, we propose to investigate the role of a subjective interpretation of reward, modulated by the environment and by the internal dynamics of the agent, which influences the manner in which reward is used by an actor-critic to adapt behaviour to the environment. Finally, seeking the maximisation of reward and the avoidance of penalty, other architectures have included principles of contextual grounding and RL to learn behavioural policies adapted to a certain environment (Butz et al., 2010).

A brief summary of the neuroscience of decision-making and hedonic value. Despite the progress about the neural organization of the brain and of the different processes encompassing decision-making, our knowledge about the interplay of the neural structures implicated in hedonic value remains incomplete (Wallis, 2012; Kennerley et al., 2011). Recent evidence has gradually revealed a complex and subtle organization of the different factors and sub-roles implicated in decision-making: expected outcome, energetic cost, hedonic value, time, risk or confidence in decision, associated with specific brain areas. Numerous experiments have been performed to characterize the operation of different brain areas during the learning of stimulus-response (SR) and behaviour-reward relationships (Balleine and O'Doherty, 2010), showing that the brain areas mainly responsible for the encoding of the function and the limbic function are the pre-frontal cortex (Balleine and O'Doherty, 2010) and the ventral striatum (Guitart-Masip et al., 2011). Next we list a summary of some of the main relevant aspects of neural encoding related to hedonic value.

- The main brain area specialised in the encoding of hedonic value in an independent manner of behaviour is the Orbito-Frontal Cortex (OFC). The OFC is an area of integration of multimodal sensory and limbic information (Wallis, 2012; Cardinal et al., 2002), receiving major afferents from sensory cortex, hypothalamus, dorsal and ventral striatum and amygdala (Rolls, 2005, 2004). Consistently with the operation of hedonic assessment, dependent on the internal physiology, neurons in the OFC activate with pleasant or positively hedonic stimuli and are sensitive to stimulus devaluation after satiation (Grabenhorst and Rolls, 2011). Furthermore, OFC neurons can rapidly reverse their responses to a visual stimulus, depending on whether its previous association was rewarding or punitive (Grabenhorst and Rolls, 2011). In addition to OFC, neighbouring areas such as the Anterior Cingulate Cortex (ACC) and the Lateral Pre-Frontal Cortex (LPFC) have been shown to encode aspects of value related to the energy cost of the candidate options of different goal-directed actions (Kennerley et al., 2011; Hazy et al., 2007).
- *Absolute vs. Relative Reward.* The encoding of hedonic value cannot solely be attributed to the OFC, as several brain areas exert different functions during the perception-action loop that necessitate of this notion. Most important may be the bonding between action and value attributed to ACC (Grabenhorst and Rolls, 2011; Quilodran et al., 2008). While the OFC encodes the absolute value associated to stimuli via projections from the amygdala, the ventral striatum (nucleus accumbens), and projections to ACC modulate activity and value as a function of external stimuli, incorporating a component of cost into its encoding of value. For example, value representations in ACC are sensitive to the presence of fat in food (Shizgal, 1997; Conover

et al., 1994). Furthemore, the nature of value encoding is constrained by the need of implementing decisions, hence requiring the comparison across often dissimilar options. Therefore, although areas unrelated to the execution of behaviour such as the OFC encode value in absolute terms, value in areas also encoding motor responses, such as PPC and PMd, are relative to make comparisons possible (Chib et al., 2009).

- Although the involvement of the nigrostriatal circuitry in the arbitration of behaviour has been long established (Redgrave et al., 1999; Houk et al., 1995), its relation to hedonic value is still under investigation. A recent study has shown, probably as a means to learn behavioural policies in an actor-critic-like manner (Li and Daw, 2012; McClure et al., 2003; Houk et al., 1995), a specialization of the medial part of the substantia nigra pars compacta (SNc) and ventral tegmental area (VTA) for the encoding of hedonic value (Guitart-Masip et al., 2011), and of the nucleus accumbens (ventral striatum) for prediction errors (Rushworth et al., 2009; Hare et al., 2008; Rolls and Grabenhorst, 2008). These areas modulate value by projecting to the ventro-medial cortex (vmOFC), fine tuning value predictions of future actions (Peters and Buechel, 2010; Samejima et al., 2005). Different aspects of value are also encoded by different brain regions; while those aspects of value related to the *goal* may be attributed to the mOFC, mPFC and amygdala, those other related to the decision are attributed to the central OFC (cOFC) (Kable and Glimcher, 2007; Schaeffer and Rotte, 2007). In this context, the contribution of the basal ganglia is the correction of prediction errors of value and the control of the intensity with which the behaviour is elicited (Jin and Costa, 2010; Turner and Desmurget, 2010).

In summary, a full description of HV or the sense of valency, as this notion is also referred to in AI (Ackley and Littman, 1991), would require a descriptive model of the recurrent activity across OFC, ACC and the ventral striatum. However, rather than a description of the neural activity across these brain areas, we deemed more useful for its application to artificial agents to build a model that captures the principle of operation of HV. In this light, the main principle of our model consists of making the modulation of value dependent on loop of interaction between the agent's homeostasis and the sensorimotor cycle. In other words, value is not only dependent on the agent's internal state, but also on the perception the agent may build of its environment. For example, the agent will not value a consummatory action in the same manner if its internal level of energy is close to satiation than to depletion, and will also vary if the necessary resource is easily attainable or scarce. This set of principles is captured by a reward formula embedded into an actor-critic RL algorithm, which aims at maintaining the agent's internal physiology within the boundaries that permit the agent's operation (Cañamero, 1997) — see sections 3.1 and 3.2. We have tested the performance of this formulation by monitoring the resulting patterns of decision-making of an artificial agent endowed with this formulation of HV, tested in several simulated environments. Overall, the results show that a subjective component to the assessment of value increases learning speed and enhances the stability of the behavioural patterns.

3 Theoretical Model

The model consists of three main parts: a module of *artificial physiology* (top right fig. 1), which is an abstraction of internal physiological processes; a *perception module* (bottom right figure 1), which provides grounded knowledge about the behaviours afforded by each object nearby the agent; a *motivated actor-critic* module (centre fig. 1), which learns behavioural patterns adapted to the environment; and a module to calculate *hedonic value* (see Value Function, centre-top fig. 1). Each module is described in the next sections, however, we have considered appropriate to introduce a

preliminary description of the agent's *behavioural structure* and of the manner in which it interacts with the environment.

Figure 1

1. The agent is endowed with a behaviour repertoire of consummatory behaviours, that the model has to learn to sequence to attain physiological stability. By consummatory we refer to the behaviours of the agent's repertoire, which may compensate one or more of the agent's physiological deficits, and are executed by interacting with nearby objects affording their execution (see the end of section 3.1 for further detail). For example, eating is only possible when an edible object is nearby. Therefore, we have first endowed our agent with a wandering behaviour to facilitate exploration and therefore object encountering. As soon as one or more objects is perceived, a decision is made about which consummatory behaviour to execute, which is followed by a *behavioural interaction* with the relevant object.
2. The unit of behaviour assessment is the *behaviour compensatory cycle*, which is a sequence of consummatory behaviours leading the agent's state from a random initial physiological state to a state of satisfaction (see section 3.1). The RL problem we introduce in section 3.2 has been structured accordingly as a learning exercise in which the agent is repeatedly initialised to random physiological states, to be compensated for by learning appropriate behaviour compensatory cycles. These cycles also serve the purpose of providing a measure of average performance for the learning algorithm.

3.1 Artificial Physiology

Our model of artificial physiology consists of a set of homeostatic, survival-related variables, a set of drives that depend on the internal variables and a repertoire of behaviours. Behaviours are selected by an arbitration mechanism based on an actor-critic algorithm (see section 3.2).

Homeostatic variables. These are abstractions representing the dynamics of the agent's internal resources. We have used three basic homeostatic variables: *nutrition*, *stamina* and *restlessness*. Nutrition is an abstraction for a number of elements controlling food intake (e.g., glucose level in blood). As time goes by, its level decreases, as described by equation 1, if food/rest is not periodically consumed. Stamina represents the level of energy of the agent, which decreases over time if the agent does not regularly rest. Restlessness is an abstraction of the level of internal agitation, which increases over time. Each of these variables has an optimal set point h_{op}, and a range of values (their viability zone) for the agent to remain viable in its environment (Cañamero, 1997; Ashby, 1965). Note that restlessness is not a survival variable. If the values of any other variable exceed their respective physiological boundaries, the agent can no longer operate. Furthermore, each homeostatic variable exhibits a status of normality (when its value approaches the set point), deficit or excess.

Formally speaking, each controlled *homeostatic variable* h_i varies due to internal bodily dynamics and to the interactions of the agent within its environment as described by equation 1.

$$\dot{h}_i = -h_i/\tau_i + \sum_k \alpha_{ki} \sum_{j_k} \delta(t - t_{j_k}). \tag{1}$$

τ_i is the exponential decay rate of variable h_i. α_{ki} is the amount of compensation that results from the execution of behaviour b_k over the variable h_i. t_{jk} lists the instants at which behaviour b_k is successfully executed, mathematically signalled by a $\delta(t)$ Kronecker function. Equation 2 shows the analytical solution to equation 1.

$$h_i(t) = h_0 e^{-t/\tau_i} + \sum_k \alpha_{ki} \sum_{j_k} u(t - t_{j_k}) e^{-(t-t_{j_k})/\tau_i}. \tag{2}$$

$u(t)$ is the Heaviside-step function. t_{j_k} is the j^{th} physiological compensation related to the execution of behaviour b_k. Again, α_{ki} is the amount of compensatory effect that results from the execution of behaviour b_k over the variable h_i.

Drives. Each homeostatic variable may express its status of deficit or excess via a set of *drives* (Hull, 1943). Drives are abstractions denoting urges to action based on bodily needs related to self-sufficiency and survival. They monitor the divergence of each homeostatic variable with respect to its set point as defined by equation 3,

$$d_m(t) = \sum_l a_{lm}(h_{op_l} - h_l(t)), \tag{3}$$

and initiate the process of correction. d_m is drive m and the h_l the value of the l^{th} related homeostatic variable and h_{op_l} its the optimal (or set) value. a_{lm} are the coefficients relating the variable l to the drive m. In a preliminary fashion, we have restricted its general expression to the case where each drive depends linearly on a single homeostatic variable. Thus, when a homeostatic variable diverges from its optimal point, the drive expresses an urge for compensation. In the general case, the value of several homeostatic variables may influence each drive. In the study presented here, we have used three drives: *hunger* (which controls nutrition), *fatigue* (controlling stamina), and *curiosity* (controlling restlessness). At each time step, the drives are assigned an intensity (activation level) proportional to the deviation of the controlled variable from its set point, which define the agent's internal state (see equation 3).

Figure 2

Behaviour repertoire. The agent has been endowed with a default exploratory *behaviour* that makes it wander in its environment. Furthermore, there is a set of three consummatory behaviours, which may be engaged by the decision-making module whenever in the vicinity of a consummatory object (see next section). All behaviours are coarse-grained sub-systems composed of simpler actions that implement different competencies. The execution of a behaviour modifies (increases or decreases) the levels of particular internal variables (see equation 1), therefore affecting the satisfaction of drives. To facilitate later analyses of internal dynamics while respecting our principles of design, we have constrained the complexity of our homeostatic system, assigning a single behaviour to compensate each drive — "eat" (approach edible objects) satisfies hunger, "rest" in a shelter satisfies fatigue, and "interact" with any object of the environment satisfies curiosity. For the behaviour execution to exert a *compensatory effect* (increment or decrement of the internal homeostatic variables), it must occur in a context offering that affordance. For example, to compensate the variable nutrition, the execution of the behaviour eating has to be executed with an edible object nearby.

3.2 Motivated Actor-Critic

This section describes the actor-critic model and the manner in which motivation and reinforcement are defined and integrated therein. The actor-critic is an on-policy RL algorithm that can learn the optimal policy to maximise cumulative reward by interaction with the environment (Sutton and Barto, 1981). Its operation has been shown to be in remarkable agreeement with experimental evidence in the learning of SR associations (Li and Daw, 2012; Schultz et al., 1993), as well as in the learning of behavioural policies (McClure et al., 2003; Houk et al., 1995). Concomitantly, we propose a motivated version of this algorithm to learn behavioural policies that minimise the agent's internal deficits in environments with different availability and distribution of resources. In consequence, we defined the vector state $\underline{s}_k$ at the learning episode k as shown by equation 4; comprising the agent's drives $\underline{d}_k$ and the potential courses of action as *perceptual affordances* $\underline{a}_k$ (note that vectors are underlined) — see equation 4.

$$\underline{s}_k = \{\underline{d}_k, \underline{a}_k\} \tag{4}$$

Although the behaviour of the robot operates in a continuous fashion, learning occurs only after each behaviour execution (k indexes the executions of behaviour). The vector $\underline{a}_k$ is the result of a transformation of sensory input into behavioural saliency, which our agent performs in an automatic fashion. The resulting values, between 0 and 1, indicate the behaviours afforded to the agent. The reader may refer to Cos et al. (2010) for a more detailed explanation of how these values may be learned via sensorimotor interaction.

The *executive function* of the model is managed by the *actor*, which evaluates the policy function for each behaviour and orders the execution of the most salient behaviour. In other words, the policy expresses the saliency of each behaviour for the current state. This function has been implemented as a set of three different functions, each implemented as a feed-forward neural network with six units in the input layer, fifteen in the hidden layer and one in the output layer (6-15-1). The output layer is linear (cf. centre fig. 3).

In a complementary fashion, the *critic* has been implemented by a feed-forward neural network with node distribution 6-15-1 and linear output layer (cf. figure 3), to calculate the expected cumulative reward $V^*(\underline{s}_k)$ — see equation 5 — from the current state s_k until the goal is attained (the state s_k is within the optimal zone of the agent's physiological space, see 4).

$$V^*(\underline{s}_k) = E\{\sum_{l=0}^{k} \gamma^{k-l} r_l | \underline{s} = \underline{s}_k\} \tag{5}$$

r_l is the predicted reward resulting from the transition between state s_k and state s_{k+1}, and E the calculation of expected value. We next summarise the operation of the motivated actor-critic in three processes:

Figure 3

1. The actor makes decisions among the three consummatory behaviours: eat, shelter and interact. First, the policy values for each behaviour are calculated for the current state s_{k-1}. Since we have used an ϵ-greedy policy, the behaviour exhibiting the highest policy value is executed 80% of the time —see equation 6. The remaining 20% of the times, a behaviour to be executed is selected at random.

$$b(\underline{s}_k) = \mathit{argmax}_i\{\Pi_i(\underline{s}_k)\} \tag{6}$$

   Π is the policy function, s_k is the state, and $b(\underline{s}_k)$ is the preferred behaviour at state $\underline{s}_k$.

2. The convergence of both actor and critic during the learning process is mutually dependent, as a function of the learning of the Temporal Difference (TD) algorithm (Sutton and Barto, 1981), see equation 9, adapted to our motivated agent.

3. Each learning episode is internally triggered by the hormone S (see equation 7), whenever the derivative of any homeostatic variable surpasses the threshold X^*, and lasts until the level of the hormone difuses (it becomes smaller than S^*).

$$\dot{S(t)} = -S/\tau_s + \sum_i \chi_i \sum_n \delta(t - t_n^i), t_n^i = t||\dot{h}_i| > X^* \tag{7}$$

   δ is the Kronecker delta, χ_i is the maximal intensity at activation, τ_s its rate of decay, and t_n^i the time at which the homeostatic variable h_i experiences a sudden compensatory variation. As such, the hormonal response peaks whenever the derivative of any homeostatic variable $\dot{h}_i$ suddenly surpasses a certain threshold X^*. Since the rate of change of the homeostatic variables is several orders of magnitude larger than that of the hormonal response ($\tau_s <<< \tau_i$),

the value of the hormone will rapidly decrease to zero after the behavioural effect has been exerted.

At each learning episode, the reward function (see next) yields a quantification of the reward (r_k) associated to a variation of the agent's physiological effect. The reward obtained r_k is then used by the critic to calculate the TD prediction error (see equation 8).

$$\delta_k = r_k + \gamma \tilde{V}(\underline{s}_k) - \tilde{V}(\underline{s}_{k-1}) \tag{8}$$

δ_k is the prediction error, r_k is the reward obtained, $\tilde{V}(\underline{s}_k)$ the estimate of the state-value function and γ is the discount factor that regulates the influence of past states, which we chose equal to 0.9. δ_k is then used to update the neural representation of the state-value function $V(\underline{s}_{k-1})$ at the state from which the decision was made $\underline{s}_{k-1}$. The update is performed by an implementation of the TD algorithm, which back propagates the error to the middle and input layer weights of the neural network (Rumelhart et al., 1986). Likewise, δ_k is also used to update the neural estimate of the preference for the behaviour just executed (see equation 9). The behaviour selection and learning operation are repeated in cycles until convergence (see experimental section).

$$p_i(\underline{s}_k) = p_i(\underline{s}_k) + \beta \delta_k; \tag{9}$$

β is the learning step and p_i is the i^{th} behaviour preference.

4. Based on the evidence reviewed in section 2, we propose a mechanism to calculate reward as a function of the dynamics of the agent's physiology and of sensorimotor interaction. Equation 10 describes this implementation of HV to monitor the agent's overall behaviour and its related physiological consequences.

$$r_k = \frac{1}{1+\lambda_S} \frac{(\underline{d}_{k-1} - \underline{d}_k) \cdot \hat{r}}{|\underline{d}_k|} \tag{10}$$

r_k is a scalar reward value resulting from the physiological shift of the internal state from $\underline{d}_{k-1}$ to $\underline{d}_k$, $\hat{r}$ is the radial unitary vector in the agent's physiological space, pointing from the origin radially towards the agent's physiological state d_k (see figure 2). By contrast, λ_S is the parameter of the poisson process we used to model the encounters with objects in our simulated scenario, and varies accordingly to the availability of resources of the environment (see the experimental section and discussion for further information). Briefly, objects are encountered at random, in average each λ_S time units. Hence, this parameter introduces the influence of a discount factor dependent on the distribution of resources of the environment. In general, our reward function yields a positive reward for any physiological compensation. However, its value increases if close to the optimal comfort zone (the inverse of the modulus of $|\underline{d}_k|$ is large).

Although the variety of behavioural patterns will significantly depend on the type of environment, we hypothesise that as a result of this formulation of value, the behaviour compensatory cycles will, at least, include the following regimes: First, it will encourage those behaviours leading away from the lethal boundary the fastest, as this is as well most rewarding. Second, whenever one or more of the homeostatic variables are sated, it should encourage behaviours to compensate any remaining deficits. Third, it should promote consumption of resources rarely encountered. Importantly, this function of reward is harmonious with the notion of physiological stability introduced by Ashby (1965), as this is, ultimately, a necessary requirement for the agent to survive.

4 Experiments

The goal of the experiments is to test the influence of the formulation of HV we just proposed (see equation 10) by assessing the agent's behaviour, to adapt to changes of their environment and of their internal homeostatic dynamics, to seek physiological stability. Although the agent's behaviour repertoire consists of a set of elementary behaviours executed individually (see section 3.1), our analysis focuses on the manner in which these are sequenced to compose *cycles of behavioural compensation*, and on other metrics that capture the agent's adaptation. We briefly describe these two behavioural levels next. Firstly, the *behaviour interaction*, which may be considered as the unit of interaction with the environment, initiated after each interaction with an object. It may be described as the following sequence of events.

- First, in the absence of any object, the agent executes its default exploratory behaviour, wandering at random.
- Upon detection of a novel object, the agent's policies are evaluated by the actor for the current state $\underline{s}_k$, and a decision is made about the behaviour to execute next (see section 3.2 for further detail).
- If the execution of the behaviour selected is successful, this may result in an improvement of the level of one or more homeostatic variables, and on a subsequent hormonal release (see equation 7).
- Following each hormonal release, reward is calculated as a function of the physiological effect resulting from interaction with the environment (see eq. 10).
- The agent's internal homeostatic variables decay over time as described by equation 1. In addition to this, the reinforcement process is initiated with the calculation of the TD error (δ_k) — see equation 8. Both the critic and the actor update their predictions of cumulative reward and behavioural preference for the previous state $\underline{s}_{k-1}$.

Secondly, we call *behavioural compensatory cycles* to the sequences of behaviour executions, starting at a random state of the agent's physiological space and ending when the optimal physiological zone has been attained (see figure 4A) — see section 3.2. These cycles result from the actor-critic's training process, consisting of repeatedly resetting the physiological state ($\underline{s}_k$) to a random set of values within the agent's physiological boundaries, which the agent has to lead to the optimal zone.

Behavioural Metrics

Our analysis of the influence of HV as modelled here, has focused on the reward value to different contingencies, parametrized by the following parameters: the *rate of decay* of the agent's homeostatic variables (τ_i decay constants); the *distribution of resources* of the environment (described by its distribution of affordances: ideal, abundant or scarce), and their availability, characterised by the λ_S parameter. The *metrics* of learning and behavioural performance are described next.

Figure 4

1. *Time until stability*. This is the time-interval required for the error in the prediction of reward δ_k (see equation 8) to reach an asymptotic value, smaller than a certain threshold ϵ (fixed to 0.3 in our simulations).
2. Additionally, we have also assessed the average performance of the resulting behavioural patterns along two complementary axes: the dynamics of its internal physiology and the behavioural cycles that result from the learning process. It is important to notice that to obtain reliable values of assessment, we have alternated phases of learning and assessment at each

simulation. Specifically, we stopped the training of the actor-critic to assess the current performance at regular intervals of 2,000 decisions (see end of introduction, section 4). Each assessment interval lasted over 200 decisions. Next we introduce the two complementary metrics we defined to this end.

First, the metric of **physiological stability**, which captures the effect of the learned behavioural strategies on the agent's internal physiology. It is calculated as the average of the agent's internal drives over the assessment interval, see equation 11:

$$Physiological\ Stability = E\{\frac{1}{N}\sum_{m=0}^{N-1} d_m\}, \tag{11}$$

and is inspired by the viability indicators of Avila-García and Cañamero (2002). d_m is the value of drive *m* averaged over the duration of a time interval determined *ad hoc* for evaluation purposes (typically 200 behaviour interactions), N is the total number of drives of the agent. Thus, when all homeostatic variables are optimally satisfied, the value of this metric should be close to zero.

Secondly, **behavioural effectiveness**, which results from an analysis of the agent's behavioural cycles (see equation 12). The use of this metric is inspired in the cycles proposed by McFarland and Spier (1997) in the two-resource problem, and is intended to extract a quantitative account of the agent's behavioural performance. To perform a systematic test, during the probing intervals only, we used four meaningful initial physiological states and recorded the agent's behaviour compensatory cycles: the first is a highly deficient physiological state close to the lethal boundary (hunger 0.9, tiredness 0.8 and restlessness 0.7). In the three remaining initial states, two out of three drives were strongly deficient: (0.5, 0.8, 0.9), (0.9, 0.5, 0.8), (0.9, 0.8, 0.5) — see figure 4B, and we quantified the behavioural efficiency of each cycle, defined as the average ratio between the amount of physiological compensatory effect, fixed to Δr_i for all behaviours during the probing, and obtained as a result of the agent's behaviour interactions over the number of behaviour interactions M.

$$Effectiveness\ =\ \frac{1}{M}\sum_{i=0}^{M-1} \Delta r_i \tag{12}$$

Δr_i is the amount of physiological effect following the successful execution of the i^{th} behaviour and M is the number of behaviours of the cycle (see figure 4). The rationale behind this metric is that failed behaviour executions yield no physiological effect, and would therefore reduce its value. By contrast, if each decision within a cycle yields a physiological compensation, this would yield a maximal effectiveness value (close to Δr). If plotting this specific compensatory cycle on the agent's physiological space, it would resemble a straight line from the initial state until the optimal zone (see figure 2). Importantly, the metric of effectiveness has been designed to yield approximately the same value irrespective of the cycle's initial state, since only the proportion of compensatory executions of the cycle exerts the most significant influence on the final value. For example, a cycle of five compensatory and two failed behaviour executions should yield a similar effectiveness value than a sequence of ten compensatory and four failed executions ($\frac{5+2}{7} \approx \frac{10+4}{14}$). These two values will only differ in so far the longer sequence will be slightly influenced by the negative value of the failed executions, although these values will be very small if compared with $Deltar$.

Figure 5

Figure 6

Experimental setup: Modulating Reward

As previously described, our working hypothesis is that the agent's adaptivity will improve by assigning value as a function of the current physiological state and of the statistics of sensorimotor interaction. To test this hypothesis, we performed a series of simulated runs and compare the results obtained when varying two factors: the decay of the agent's internal physiological variables (τ decay constant), and the distribution of resources of the environment. Consistent with a previous study (Cos et al., 2010), we used three different scenarios, each of which characterised by a certain distribution of resources.

1. We first used a baseline *ideal* environment in which all objects afford all behaviours to be executed by the agent. Furthermore, λ_S was set to 0.9, meaning that objects were encountered, in average, every other simulation steps each. Hence, reward value was divided by two.
2. Second, an *abundant* scenario in which the object affordances were constrained, but were still relatively frequent. Objects smaller than 0.4 afford eating (50%), objects larger than that afford shelter (50%) and all objects afford interaction — see figure 5A. λ_S was set to 0.3, hence encounters were every three to four steps each (reward value was scaled by a factor of three).
3. Third, a *scarce* scenario to study the contribution of the resources from the environment and the agent's internal physiology to the computation of reward, the *scarce* scenario. In this scenario, each object affords a single behaviour as a function of its size; objects smaller than 0.4 afford graping (50%), objects whose size is between 0.4 and 0.7 afford shelter (25%) and objects larger than 0.7 afford interaction (25%) — see figure 5B. λ_S was set to 0.1, meaning that encounters were, in average, every ten simulation steps each (reward value was scaled by a factor of 10).

We performed twenty simulation runs per condition, and recorded the time-course of the agent's internal physiology and the encompassing behavioural cycles throughout this time. Figure 7

Results

The graphs in figure 6A and 6B show the length of the average behaviour compensatory cycle during the time-course of the learning phase for the abundant and scarce environments, when compared with the ideal case. As expected, the cycle length exhibits a gradual shortening in all cases, as the knowledge about the environment improves and the behavioural policy becomes increasingly effective. Figure 8

- A visual comparison of the time-course for the scarce and abundant cases (see figure 6A vs. 6B), shows that it takes a similar time to reach a *stationary regime* in both environments (80 decisions). Note that the *stationary regime* starts once the agent has attained a stable behaviour, which yields the short compensatory cycles and a highly efficient behaviour in terms of physiological stability (see figure 6A-D). The final average length of the compensatory cycle obtained for each scenario were: 11 for the ideal scenario, 18 for the abundant and 26 for the scarce. The shortest possible cycle length for the given τ decay constant (10ms) is obtained in the ideal scenario: 11 behaviour executions.
- Similarly, figures 6C and 6D show that the *physiological stability* (see eq. 11) exhibits a significant decline over time until reaching a stationary value. Final values close to 0.05 and 0.035 have been reached for the abundant and scarce scenarios, respectively. The similarity between the time-course of the cycle length and the physiological stability suggests a strong correlation between both metrics. In other words, it should be expected that an efficient behavioural policy facilitated a quick satisfaction of internal needs.

Figure 9

Policy Characterisation. As a means to characterise *the behavioural policies*, we made a comparative analysis between the learnt policies and a set of theoretical but meaningful policies: a purely *motivated policy*, which would satisfy the drive exhibiting the largest urge first (Winner-Take-All of the drives — WTA-D), and a purely *incentive-driven policy*, i.e., a policy that executes the behaviour afforded by the closest object (Winner-Take-All of the affordances — WTA-A). We calculated these metrics during probing intervals of two hundred trials every two thousand decisions, each during simulations in all three scenarios. This is the basis of a comparative analysis between them at the beginning and end of the simulated run and after the behavioural policy has attained a stationary regime. The results are shown in figure 7 for the ideal and scarce cases and in figure 8 in the abundant case (fig. 8A-C).

1. Figure 7A-C shows the results of the aforementioned metrics in the *ideal* scenario. Figure 7A-B shows the time-course of the agent's drives during a few cycles, at the beginning of a simulation and when the stationary regime has been attained, respectively. The magnitudes shown, from top to bottom are: the value for each drive, the actor's policy value for each behaviour, the drive exhibiting the highest urge (WTA-D), and the preferred behaviour according to the actor's policy (WTA-B). Initially, the effect of the learning process may be visually assessed by comparing the value of the drives and policies between both time intervals. To facilitate a visual comparison, the colour of each behavioural policy matches the colour of the drive it is associated with (red for the hunger/eat, green for tiredness/rest and blue for restlessness/interact). Remarkably, when the stationary regime has been attained, at each cycle, the drives decrease efficiently towards their minimum values (see top fig. 7B). Accordingly, the actor's policies prioritises those behaviours which could compensate the most urgent deficit the quickest. This is confirmed by the remarkable similarity between the two WTA graphs, showing that the most urgent drive (WTA-D) and the behaviour whose policy value is the highest (WTA-B) match one another. In other words, the policy prioritises to execute the behaviour that would satisfy the most urgent drive first.

   Also, to illustrate the similarities between policies, we have plotted the percentage or similarity between the purely motivated policy, and the actor's policy — see figure (7C). Although the percentage of agreement starts as low as 20%, it reaches a remarkable 72%. The remaining 28% may be explained by the $\epsilon = 0.2$ greedy policy, which obliges the actor to select 20% of the time at random for exploratory purposes.

2. Figure 7D-F shows the same comparative analysis, performed for the data obtained in the *abundant* scenario. Again, a visual inspection of the behavioural cycles shows that after reaching the stationary regime, the drives decrease efficiently towards small values at each cycle due to an efficient policy function (see fig. 7E). As for the previous case, these results suggest that the actor's final policy prioritises behaviours to compensate the drive exhibiting the highest urge first. In other words, the policy is consistent with the principle that when resources are abundant, the expression of the internal motivations should dominate the behavioural patterns, as this facilitates the gain of physiological stability and the gathering of cumulative reward. This tendency is confirmed by the similarity between the theoretical motivated policy (WTA-D) and the actor's policy (WTA-B), shown in figure 7DF. As for the previous case, we also show the percentage of agreement between the hypothetical motivated policy and the actor's behaviour, averaged over twenty simulation runs (see fig. 7F). Consistently with our hypothesis, the final stationary value is around 65%. Although smaller than the 72% obtained in the ideal scenario, this demonstrates that the policy is predominantly motivation-driven, rather than stimulus-driven. The remaining 35% of mismatch may be distributed between the 20% excluded by the greedy policy ($\epsilon = 0.2$) and a 15% due to decisions of behaviours not match-

ing the affordance offered by the object nearby or because its physiological effect would on a homeostatic variable already sated.

3. Figure 8A-C shows the results in the case of the *scarce* scenario. To characterise scarcity, each object in this environment afforded a single behaviour to our agent (see fig. 5B). Under these circumstances, a purely motivated policy would be inefficient, as the agent may frequently encounter situations in which the behaviour required to compensate the most urgent need may not be afforded by the nearby object, yielding long behavioural cycles and relatively low cumulative reward. By contrast, from an ecological perspective, we believe that it may be reasonable to follow an incentive-driven policy to profit from any available resource. Consistently with this, figure 8AB shows the sequence of afforded behaviours encountered and the preferred policy values, during a few initial behavioural cycles and during the stationary regime. As a test of similarity, we have shown the affordance perceived with the highest value (WTA-A) at all times, and the preferred behaviour according to the actor's policy (WTA-B). Despite the initial mismatch, both metrics exhibit a significant similarity once the stationary regime has been attained. Hence, the learned policy is mostly driven by external stimuli (stimulus-driven) rather than by the agent's internal drives (motivation-driven) — see figure 8C.

Behavioural Cycles. Figure 9 shows a few behavioural cycles recorded after the policy was stationary in the abundant and scarce environments. Each cycle starts whenever the agent's physiological state is initialised at one of three highly deficient states (see figure 4B) and ends once the drives are satisfied. As previously described, the policies obtained in both environments favour those specific behaviours leading to the optimal zone the fastest. However, as a result of the specifics in resource availability and subjective reward assignment in each case, the resulting policies are strongly biased towards prioritising internal drives (motivated policy) in the abundant scenario and towards externally driven responses in the scarce one. In particular, encounters in a scarce scenario are less frequent, implying that the λ_S characterising the average frequency of object encounter within this scenario is smaller (see equation 10), which results in a magnification of value in that context. The distribution of resources also exerts a significant influence on the effectivity with which each behaviour execution compensates the agent's physiological deficits, and consequently, on the resulting cycles of behavioural compensation. As shown at the beginning of section 4, cycles tend to be significantly longer in the scarce than in the abundant scenario. Hence, although the final level of physiological stability may be comparable, the velocity with which these stationary values exhibits a significant difference. This intuition is reinforced by the quantification of the metric of behavioural effectiveness, which is 34% larger in the abundant scenario ($0.88\Delta r_i$) than in the scarce one ($0.54\Delta r_i$) — see end of section 4.

5 Discussion

Previous studies have investigated several principles related to the notion of value to devise increasingly flexible strategies of adaptation for mobile robotics: behavioural cycles to interact with the environment (McFarland and Spier, 1997; Ahlgren and Halberg, 1990), new algorithms to incorporate novelty (Huang and Weng, 2002), hierarchical RL architectures for skill learning (Konidaris et al., 2010; Baldassarre, 2002), or the use of algorithms to select learning goals automously using a value system in a RL context (Merrick, 2010). The operation of each of these approaches is to some extent based on exploiting the interaction with the environment as a guide to structure behavioural responses. In a complementary fashion, here we reviewed some evidence from neuroscience underlying the phenomenon of HV, as an additional mechanism of adaptation, and proposed an elementary formalization of its core principles using a specific reward formula, which encloses the agent's inter-

nal physiological dynamics as part of the process of internal assessment.

There are numerous examples of this principle of adaptation in the animal world. To cite a few, hibernating animals increase their appreciation for food consumption as temperatures start declining and increase their foraging rhythms likewise, in a more controlled environment, it is well known that lab animals devaluate food as they gradually satiate (Shizgal, 1997). Inspired by these observations and on the data we previously reviewed, this study has proposed a simplified version of the natural mechanisms influencing the internal, subjective perception of value, typically dependent on an overall assessment of the environment and of the agent's internal physiological state, and consequently influencing their behaviour. In particular, we included the agents' internal deficits and its knowledge of the environment as factors biasing the assessment of value used for learning and adaptation. In this manner, our formulation of HV depends on the agent's internal state, but also of the intensity of the physiological effect resulting from a consummatory interaction and of the frequency of object encounter. To test this, we performed a number of tests of adaptation in environments with different availability and distribution of resources with an agent endowed with this formulation of reward. Our main conclusion is that, for the given experimental setup, the manner in which reward is internally perceived exerts a strong influence on the strategy to attain physiological stability. Remarkably, although intuition dictates that it should take longer in the scarce than in the abundant scenario if based on the number of object interactions only, the agent balanced out the learning phase by magnifying the reward value for the fewer interactions by modifying the actor's policy accordingly, and consequently, accelerated the adaptation process. However, it is worth noting that if the agent were also endowed with the ability of dynamically predicting the λ_S as it interacts with the environment, this adaptation could be accelerated even further. The final stationary values of physiological stability are comparable in both scenarios (see figure 6C-D). Furthermore, the advantage of an assessment criterion dependent on the internal physiology comes forth when comparing to the theoretical case of a value formula independent of subjective criteria. An example for the case of consummatory behaviours is the compulsive glutton, in which the lack of sensitivity to the physiological effect of eating leads the subject to consume food continuously beyond satiation limits in a continuous fashion. Although this certainly satisfies the need of nutrition, not being able to exercise some restraint is also detrimental for physiological stability, as the excess of satisfaction of one drive may mean that too little time has been devoted to addressed the remaining ones.

By contrast, as we integrate the dynamics of interaction into the process of internal assessment, we are making the agent's internal scale of value relative to the environment in which it lives and interacts, and importantly, dependent not on one of the agent's homeostatic variables but on each and everyone of them simultaneously. This implementation of external-internal dynamics, generates a process of internal assessment *grounding* the agent's assessment on the environment. Because of this ecological formulation of reward value, it is possible to estimate the boundaries of these policies by the ratio between the rate of decay of the agent's internal resources over the rate of successful encounter with objects in the environment, which again, depends on its distribution of resources. In conclusion, although it is certainly possible to learn adaptive policies with a fixed reward formula, we argue that the natural consequence of a grounded criterion of assessment is the faster adaptation of overall behavioural strategies to changes of the environment. For example, the same architecture we presented yielded strategies ranging from motivation driven to reactive, as a function of the environment (see figure 7C, 7F and 8C). However, our agent could dynamically alternate between these strategies as changes of the environment were presented.

Grounded hedonic value. As opposed to a fixed reward formula, the operation of our specific formulation of hedonic value is influenced by a relative assessment between the rhythms of consumption of the agent's internal resources (τ decay constants) and the frequency of successful encounters with objects of the environment. The calculation of value, in the terms proposed by the reward for-

mula (see eq. 10), results from the mutual interaction across several external and internal factors, often with opposite effect. First, the main contribution results from the external to internal effect that follows a successful behaviour execution ($\underline{d}_{k-1} - \underline{d}_k$). If the execution of the behaviour did not yield any compensatory effect, the value is negative, as the natural decay of the homeostatic variables made $\underline{d}_k$ larger than $\underline{d}_{k-1}$. Second, the value is also sensitive to the overall deficitary state of the agent. Since the value of the overall deficit $\|d_k\|$ ranges between zero and one, the value will be magnified if the physiological effect of that decision approaches the optimal zone (see figure 4A). Third, the value of an interaction is also dependent on the distribution of object affordances (see figure 5) and by the rate of object encounter in the environment, captured by the λ_S parameter (see methods section). The longer the time between successful interactions, the smaller the reward.

Certainly, this is not the only possible formulation of value that could be proposed to ground value in the environment in a dynamic fashion. However, the purpose of this study was to portray the operation of the principle of hedonic value on behaviour, relating the basic external and internal factors, not to provide a detailed implementation of the neural processes and dynamics underlying reward in the brain. Although this should not be considered the last word on the matter, this study provides a significant understanding on the mechanisms underlying adaptation, which depend on the subjective interpretation of reward, and it may be considered as a first step towards a parsimonious formulation of hedonic value for artificial agents.

6 Conclusion

This paper has formulated a process of internal modulation of value as an additional mechanism to extend the agent's adaptivity to difficult or changing environments. It shows that whenever we make reward value dependent on the motivational state and on previous experience about the environment, in our case recorded by the actor-critic policies, this modulation can exert a significant influence on the behavioural cycles generated and on the agent's overall physiological stability. Although further study will be necessary to find specific methodologies to develop mechanisms of adaptation based on affective phenomena, the review and results presented here highlight that this kind of processes are called to play a significant role in behavioural adaptation.

Acknowledgements

We wish to acknowledge George Konidaris for his support and advice and Paul Cisek, Matthew Carland, Mehdi Khamassi and Benoit Girard for their comments on an earlier version of this manuscript. This research was funded by a fellowship of the British Council/La Caixa and a fellowship of the Graduate School at the University of Edinburgh.

References

Ackley, D. and Littman, M. (1991). Interactions between learning and evolution. In Langton, C., Taylor, C., Farmer, J. D., and Rasmussen, S., editors, *Artificial Life II, SFI Studies in the Sciences of Complexity*, volume X, pages 487–509. Addison-Wesley, Menlo Park, CA.

Ahlgren, A. and Halberg, F. (1990). *Cycles of nature: an introduction to biological rhytms*. National Teachers Association.

Alexander, G., DeLong, M., and PL, S. (1990). Basal ganglia-thalamocortical circuits: parallel

substrates for motor, oculomotor, prefrontal and limbic functions. *Progress in Brain Research*, 85:119–146.

Ashby, W. (1965). *Design for a Brain: The Origin of Adaptive Behaviour*. Chapman & Hall, London.

Avila-García, O. and Cañamero, L. (2002). A comparison of behaviour selection architectures using viability indicators. In *Proc. of International Workshop on Biologically-Inspired Robotics: The Legacy of W. Grey Walter*. Bristol HP Labs, UK.

Baldassarre, G. (2002). A modular neural-network model of the basal ganglia's role in learning and selecting motor behaviours. *Journal of Cognitive Systems Research*, 3:5–13.

Balleine, B. W. and O'Doherty, J. P. (2010). Human and rodent homologies in action control: Corticostriatal determinants of goal-directed and habitual action. *Neuropsychopharmacology*, 35:48–69.

Baranes, A. and Oudeyer, P.-Y. (2010). Adaptive motivation in a biomimetic action selection mechanism‘. In *Proceedings of IEEE/RSJ International Conference on Intelligent Robots and Systems (IROS2010)*.

Bloch, F. (2007). Comment on 'making decisions in large worlds' by ken binmore. *Annals of Economics and Statistics*, 86:43–45.

Butz, M. V., Shirinov, E., and Reif, K. L. (2010). Self-organizing sensorimotor maps plus internal motivations yield animal-like behavior. *Adaptive Behavior*, 18(3-4):315–337.

Cañamero, L. D. (1997). Modeling motivations and emotions as a basis for intelligent behavior. In Johnson, W. L., editor, *Proceedings of the First International Symposium on Autonomous Agents (Agents'97)*, pages 148–155. New York, NY: ACM Press.

Cardinal, R., Parkinson, J., Hall, J., and Everitt, B. (2002). Emotion and motivation: the role of the amygdala, ventral striatum, and prefrontal cortex. *Neuroscience and Behavioral Reviews*, 26(3):321–352.

Chib, V., Rangel, A., Shimojo, S., and O'Doherty, J. (2009). Evidence for a common representation of decision values for dissimilar goods in human ventromedial prefrontal cortex. *Journal of Neuroscience*, 29:12315–12320.

Coninx, A., Guillot, A., and Girard, B. (2008). Adaptive motivation in a biomimetic action selection mechanism‘. In *2nd french conference on computational neuroscience (NeuroComp 2008)*, pages 158–162.

Conover, K. L. and Shizgal, P. (1994). Differential effects of postingestive feedback on the reward value of sucrose and lateral hypothalamic stimulation in rats. *Behavioral Neuroscience*, 108(3):559–572.

Conover, K. L., Woodside, B., and Shizgal, P. (1994). Effects of sodium depletion on competition and summation between rewarding effects of salt and lateral hypothalamic stimulation in the rat. *Behavioral Neuroscience*, 108(3):549–558.

Cos, I., Cañamero, L., and Hayes, G. M. (2010). Learning affordances of consummatory behaviors: Motivation-driven adaptive perception. *Adaptive Behavior*, 18(3-4):285–314.

Damasio, A. R. (2000). *The Feeling of What Happens: Body and Emotion in the Making of Consciousness*. Harcourt Brace.

Damoulas, T., Cos-Aguilera, I., Hayes, G., and Taylor, T. (2005). Valency for adaptive homeostatic agents: Relating evolution and learning. In *In Proceedings of the 8th European Conference on Artificial Life (ECAL2005), Canterbury, UK.*, volume 3630 of *Lecture Notes in Artificial Intelligence (LNAI)*, pages 936–945. Springer ; Berlin/Heidelberg.

Daw, N., Niv, Y., and Dayan, P. (2005). Uncertainty-based competition between prefrontal and dorsolateral striatal systems for behavioral control. *Nature Neuroscience*, 8(12):1704–1711.

Dickinson, A. and Balleine, B. (2001). *The role of learning in motivation*, volume 3 of *Steven's Handbook of Experimental Psychology*, chapter Learning, Motivation and Emotion. New York, NY: Wiley, third edition.

Dittrich, K. and Klauer, K. C. (2011). Does ignoring lead to worse evaluations? a new explanation of the stimulus devaluation effect. *Cognition and Emotion*, 1:1–16.

Eder, A. B. and Rothermund, K. (2008). When do motor behaviors (mis)match affective stimuli? an evaluative coding view of approach and avoidance reactions. *Journal of Experimental Psychology*, 137(2):262–281.

Freud, S. (1940). *An Outline of Psychoanalysis*. Hogarth Press, London.

Gibson, J. (1986). *The Ecological Approach to Visual Perception*. Hillsdale, N.J. ; London.

Girard, B., Tabareau, N., Pham, Q., Berthoz, A., and Slotine, J.-J. (2008). Where neuroscience and dynamic system theory meet autonomous robotics: a contracting basal ganglia model for action selection. *Neural Networks*, 21(4):628–641.

Grabenhorst, F. and Rolls, E. T. (2011). Value, pleasure and choice in the ventral prefrontal cortex. *Trends in Cognitive Sciences*, 15(2):56–67.

Grabenhorst, F., Rolls, E. T., and Paris, B. A. (2008). From affective value to decision-making in the prefrontal cortex. *European Journal of Neuroscience*, 28:1930–1939.

Guitart-Masip, M., Fuentemilla, L., Bach, D. R., Huys, Q. J., Dayan, P., and Dolan, R. J. (2011). Action dominates valence in anticipatory representations in the human striatum and dopaminergic midbrain. *The Journal of Neuroscience*, 31(21):7867–7875.

Gurney, K., Prescott, T. J., and Redgrave, P. (1998). The basal ganglia viewed as an action selection device. In *Proceedings of the Eigth International Conference on Artificial Neural Networks*, pages 1033–1038.

Hare, T. A., O'Doherty, J., Camerer, C. F., Schultz, W., and Rangel, A. (2008). Dissociating the role of the orbitofrontal cortex and the striatum in the computation of goal values and prediction errors. *Journal of Neuroscience*, 28(22):5623–5630.

Hazy, T. E., Frank, M. J., and O'Reilly, R. C. (2007). Towards an executive without a homunculus: computational models of the prefrontal cortex/basal ganglia system. *Philosophical Transactions from the Royal Society. Series B.*, 113:281–292.

Hinde, R. A. (1960). Energy models of motivation. In *Symposium of Society of Experimental Biology*, volume 14, pages 199–213. Company of Biologists on behalf of the Society for Experimental Biology, at Cambridge University Press.

Hinde, R. A. (1971). Critique of energy models of motivation. In Bindra, D. and Stewart, J., editors, *Motivation*, pages 36–48. Penguin.

Houk, J. C., Adams, J. L., and Barto, A. G. (1995). Models of information processing in the basal ganglia. In Houk, J. C., Davis, J. L., and G., B. D., editors, *A Model of How the Basal Ganglia Generate and Use Neural Signals That Predict Reinforcement*, A Bradford Book, chapter 13, pages 249–270. MIT Press, 2nd. edition (1998) edition.

Huang, X. and Weng, J. (2002). Novelty and reinforcement learning in the value system of developmental robots. In *2nd International Workshop on Epigenetic Robotics*, pages 57–55.

Hull, C. (1943). *Principles of Behaviour: an Introduction to Behaviour Theory*. D. Appleton-Century Company Inc., New York.

Jin, X. and Costa, R. M. (2010). Start/stop signals emerge in nigrostriatal circuits during sequence learning. *Nature*, 466(22):457–462.

Kable, J. and Glimcher, P. (2007). The neural correlates of subjective value during intertemporal choice. *European Journal of Neuroscience*, 10(12):1625–1633.

Kennerley, S. W., Behrens, T. E., and Wallis, J. D. (2011). Double dissociation of value computations in orbitofrontal and anterior cingulate neurons. *Nature Neuroscience*, 14(12):1581–1589.

Khamassi, M., Lachèze, L., Girard, B., and Guillot, A. (2005). Actor-critic models of reinforcement learning in the basal ganglia: From natural to artificial rats. *Adaptive Behavior*, 13(2):131–148.

Konidaris, G. and Barto, A. (2006). An adaptive robot motivational system. In Nolfi, S., Baldassarre, G., Calabretta, R., Hallam, J. C. T., Marocco, D., Meyer, J.-A., Miglino, O., and Parisi, D., editors, *Animals to Animats 9: Proceedings of the 9th International Conference on Simulation of Adaptive Behavior*, pages 346–356. Springer Verlag.

Konidaris, G., Kuindersma, S., Barto, A., and Grupen, R. (2010). Constructing skill trees for reinforcement learning agents from demonstration trajectories. *Advances in Neural Information Processing Systems*, 23:2402–2410.

Konidaris, G. D. and Hayes, G. M. (2005). An architecture for behavior-based reinforcement learning. *Adaptive Behavior*, (13(1)):5–32.

Li, J. and Daw, N. D. (2012). Signals in human striatum are appropriate for policy update rather than value prediction. *Journal of Neuroscience*, 31(14):5504–5511.

Lorenz, K. (1966). *Evolution and Modification of Behaviour*. Methuen & Co Ltd, London.

Matarić, M. and Brooks, R. (1990). *Cambrian intelligence: The early history of the new AI*, chapter Learning a distributed map representation based on navigation behaviors. The MIT Press.

McClure, S. M., Daw, N. D., and Montague, P. R. (2003). A computational substrate for incentive salience. *TRENDS in Neurosciences*, 26(8).

McDougall, W. (1913). The sources and direction of psychophysical energy. *American Journal of Insanity*.

McFarland, D. and Spier, E. (1997). Basic cycles, utility and opportunism in self-sufficient robots. *Robotics and Autonomous Systems*, (20):179–190.

McFarland, D. J. and Sibly, R. M. (1975). The behavioural final common path. *Proceedings of the Royal Society of London*, 270:265–293.

Merrick, K. E. (2010). Modelling behavior cycles as a value system of developmental robots. *Adaptive Behavior*, 18:237–257.

Peters, J. and Buechel, C. (2010). Neural representations of subjective reward value. *Behavioural Brain Research*, 213:135–141.

Pfeifer, R. (1996). Building fungus eaters: Design principles of autonomous agents. In Maes, P., Mataric, M. J., Meyer, J.-A., Pollack, J., and Wilson, S. W., editors, *FROM ANIMALS TO ANIMATS 4: Fourth International Conference on Simulation of Adaptive Behavior*, Complex Adaptive Systems, pages 3–12, Cambridge, MA. The MIT Press/Bradford Books.

Quilodran, R., Roth, M., and Procyk, E. (2008). Behavioral shifts and action valuation in the anterior cingulate cortex. *Neuron*, 57(2):314–325.

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

Reynolds, J. R. and O'Reilly, R. (2009). Developping pfc representations using reinforcement learning. *Cognition*, 113:281–292.

Rolls, E. T. (2004). The functions of the orbitofrontal cortex. *Brain and Cognition*, 55:11–29.

Rolls, E. T. (2005). *Emotion Explained*. Oxford University Press.

Rolls, E. T. and Grabenhorst, F. (2008). The orbitofrontal cortex and beyond: From affect to decision-making. *Progress in Neurobiology*, 86:216–244.

Rumelhart, D., Hinton, G., and Williams, R. (1986). Learning internal representations by error propagation. In Rumelhart, D. and McClelland, J., editors, *Parallel Distributed Processing: Explorations in the Microstructure of Cognition*, volume I. MIT Press.

Rushworth, M. F., Mars, R. B., and Summerfield, C. (2009). General mechanisms for making decisions? *Current Opinion in Neurobiology*, 19:75–83.

Samejima, K., Ueda, Y., Doya, K., and Kimura, M. (2005). Representation of action-specific reward values in the striatum. *Science*, 310:1337–1340.

Schaeffer, M. and Rotte, M. (2007). Favorite brands as cultural objects modulate reward circuit. *Neuroreport*, 18(2):141–145.

Schultz, W., Apicella, P., and Ljungberg, T. (1993). Responses of monkey dopamine neurons to reward and conditioned stimuli during successive steps of learning a delayed response task. *Journal of Neuroscience*, (13):900–913.

Schultz, W., tremblay, K., and Hollerman, J. (2000). Reward processing in primate orbitofrontal cortex and basal ganglia. *Cerebral Cortex*, 10:272–741.

Seth, A. K. (2000). Agent-based modelling and the environmental complexity thesis. In Hallam, B., Floreano, D., Hallam, J., Hayes, G., and Meyer, J., editors, *From animals to animats 7: Proc. Seventh Int. Conf. on the simulation of Adaptive Behavior*, pages 13–24. Cambridge, MA: MIT Press.

Shizgal, P. (1997). Neural basis of utility estimation. *Current Opinion in Neurobiology*, 7:198–208.

Smith, K. S., Berridge, K. C., and Aldridge, J. W. (2011). Disentangling pleasure from incentive salience and learning signals in brain reward circuitry. *PNAS*, 108(27):255–264.

Sutton, R. and Barto, A. (1981). Towards a modern theory of adaptive networks: Expectation and prediction. *Psychological Review*, 198(88):135–170.

Sutton, R. S. and Barto, A. G. (1998). *Reinforcement Learning*. MIT Press.

Tanji, J. and Hoshi, E. (2001). Behavioural planning in the prefrontal cortex. *Current Opinion in Neurobiology*, 11:164–170.

Tinbergen, N. (1951). *The Study of Instinct*. Oxford University Press.

Toussaint, M. (2003). Learning a world model and planning with a self-organizing, dynamic neural system. In *Proceedings of NIPS' 2003*.

Turner, R. S. and Desmurget, M. (2010). Basal ganglia contributions to motor control: a vigour tutor. *Current Opinion in Neurobiology*, 20:704–716.

Velásquez, J. (1998). Modeling emotion-based decision-making. In *Proceedings of the 1998 AAAI Fall Symposium Emotional and Intelligent: The Tangled Knot of Cognition*, Technical REport FS-98-03, pages 164–169. Orlando, FL: AAAI Press.

Wallis, J. D. (2012). Cross-species studies of orbitofrontal cortex and value-based decision-making. *Nature Neuroscience*, 15(1):13–19.

Wallis, J. D. and Miller, E. (2003). Neuronal activity in primate dorsolateral and orbital prefrontal cortex during performance of a reward preference task. *European Journal of Neuroscience*, 18(7):2069–81.

Wilson, S. W. (1991). The animat path to ai. In Meyer, J.-A. and Wilson, S., editors, *From Animals to Animats: Proceedings of the First International Conference on Simulation of Adaptive Behavior*. The MIT Press.

Figure Captions

Figure 1. Architecture for Behaviour Selection and Learning. It consists of a module of internal physiology (top right), including a set of homeostatic variables and internal drives, a sensory module delivering the behaviour affordances for the closest object (bottom right), the value function (centre top), and the actor-critic module (centre), learning behavioural policies as a function of the value obtained by interaction with the environment (left).

Figure 2. Schematic of a 2D physiological space. The blue dots indicate a sequence of possible behavioural executions leading to the optimal zone. As consummatory behaviours are executed the state shifts towards the optimal zone. The viability zone is enclosed by the lethal boundary.

Figure 3. Depiction of the motivation-driven actor-critic RL schematic. Its state consists of the set of perceived affordances and of the values of the agent's internal drives. This information is used to calculate the behaviour preferences for the current state. The actor is composed of a set of multi-layer feed-forward neural networks to implement the behaviour preference functions (as part of the actor). Furthermore, the state-value function is also estimated by a separate neural network as part of the critic.

Figure 4. A. Depiction of a simplified 2-D physiological space. The x and y axis represent the value for Drive 1 (D1) and Drive 2 (D2). The two red sequences represent the time-course of the agent's physiological state during two random behavioural cycles, which sequentially compensate one drive after another until reaching the optimal zone. B. Depiction of the agent's 3-D physiological space. The x, y and z dimensions stand for the levels of hunger, tiredness and restlessness of the agent. The four colour dots represent the four initial states considered to test the agent's behavioural responses and the lines directed towards the origin the theoretically optimal physiological transitions towards the optimal zone (see related results in figure 9).

Figure 5. **Abundant** and **scarce** distribution of affordances as a function of object width. We parametrised the affordances offered by each object as a function of its simulated width parameter, which ranges between 0 and 1. The different grey areas indicate the size intervals for an object to offer the affordance labelled no the y-axis. For example, objects of width between 0.4 and 1.0 afford to rest in the abundant scenario (left), while this is reduced widths between 0.4 and 0.7 in the scarce one (right).

Figure 6. A. Behavioural cycle length for the abundant scenario (see figure 5A) with respect to baseline (ideal scenario), as a function of simulation-time (#Compensatory Cycle), during a typical simulation run.B. Same metric for the scarce scenario (see figure 5B). C. Time-course of the agent's physiological stability during a simulation run (ideal scenario) for the abundant distribution of affordances (see figure 5A) with respect to baseline. D. Likewise for the scarce distribution of affordances (see figure 5B)

Figure 7. **Analysis of behavioural patterns learned by the agent in the ideal and abundant environments, which exhibit a mainly motivation-driven pattern.** The simulation

time unit on the x-axis is the decision # during the proving phase. A. Sample behavioural cycles in an ideal environment at the beginning of a typical simulation run(cycle initiation is signaled by vertical dashed lines). The top traces show the drive values (D1, red-hunger; D2, green-tiredness; D3, blue-restlessness), and the drive expressing the highest urge (WTA-D) during a few cycles. The two bottom graphs show the policy values for each behaviour during the same behavioural cycles (B1, red-eat; B2, green-rest; B3, blue-interact), and the behaviour # exhibiting the highest activation (WTA-B). B. Like A, but at the end of a typical simulation run in an ideal scenario. C. Ideal Environment. Percentage of agreement of the agent's decisions with a purely motivation-driven policy. The x-axis scale stands for the probing block at which the metric was calculated (see methods). D-F. Like A-C, respectively, but referring to the abundant environment.

Figure 8. **Analysis of behavioural patterns in the scarce environment, which exhibit a stimulus-driven pattern**. A. Sample behavioural cycles in an ideal environment at the beginning of a typical simulation run. The top traces show the affordance values (A1, red-eating; A2, green-rest; A3, blue-interact), and the affordance perceived with the highest intensity (WTA-A) during a few cycles. The two bottom graphs show the policy values for each behaviour during the same behavioural cycles (B1, red-eat; B2, green-rest; B3, blue-interact), and the number of the behaviour exhibiting the highest activation (WTA-B). B. Like A, but at the end of a typical simulation run in an ideal scenario. C. Scarce Environment. Time-course of motivation-driven and stimulus/incentive driven behaviour, averaged over 20 simulation runs each. The x-axis scale stands for the probing block at which the metric was calculated (see methods).

Figure 9. Typical behavioural cycles for the case of the *abundant* (top) and *scarce* (bottom) environments, recorded during the stationary regime. The 3D space represents the physiological space of the agent; the initial position (blue dot) is the initial physiological state of the cycle. The green and red traces showing the effect of the behaviour executions on the agent's physiological state, starting each at three possible physiological states (hunger, tiredness, boredom), from left to right, respectively: (0.9, 0.8, 0.7), (0.5, 0.8, 0.9) and (0.9, 0.5, 0.8), and end in the optimal zone (black area).