


Learning by appraising: an emotion-based approach to intrinsic reward design

Adaptive Behavior
2014, Vol. 22(5) 330–349
© The Author(s) 2014
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/1059712314543837
adb.sagepub.com


Pedro Sequeira, Francisco S Melo and Ana Paiva

Abstract

In this paper, we investigate the use of emotional information in the learning process of autonomous agents. Inspired by four dimensions that are commonly postulated by appraisal theories of emotions, we construct a set of reward features to guide the learning process and behaviour of a reinforcement learning (RL) agent that inhabits an environment of which it has only limited perception. Much like what occurs in biological agents, each reward feature evaluates a particular aspect of the (history of) interaction of the agent history with the environment, thereby, in a sense, replicating some aspects of appraisal processes observed in humans and other animals. Our experiments in several foraging scenarios demonstrate that by optimising the relative contributions of each reward feature, the resulting “emotional” RL agents perform better than standard goal-oriented agents, particularly in consideration of their inherent perceptual limitations. Our results support the claim that biological evolutionary adaptive mechanisms such as emotions can provide crucial clues in creating robust, general-purpose reward mechanisms for autonomous artificial agents, thereby allowing them to overcome some of the challenges imposed by their inherent limitations.

Keywords

Reinforcement learning, intrinsic motivation, appraisal theories of emotions, reward design, optimal reward problem

1 Introduction

From a computational perspective, reinforcement learning (RL) is concerned with providing efficient algorithms that enable artificial agents to acquire new tasks through trial and error (Kaelbling, Littman, & Moore, 1996; Sutton & Barto, 1998). Through a process of repeated interactions with its environment, an RL agent experiments with different actions, observes their effect on the environment, and receives evaluative feedback (in the form of a numerical reinforcement signal) about how well it is performing with respect to some unknown target task. Inspired by behaviourist psychology theories, RL algorithms are a natural choice when designing autonomous agents that must *adapt* their behaviour to their environment (Sutton & Barto, 1998).

However, in deploying an RL agent, the agent designer is faced with a number of design challenges that critically impact the performance of the agent. The first (and perhaps most fundamental) challenge is an *agent-modelling* challenge: RL agents are characterised by their *state* in the environment, and the state should contain all relevant information for the agent to select the best possible action. Then, at each decision step,

the agent should *observe the current state*, select one possible action from its repertoire, observe the impact of this action in terms of both the resulting state and the received reinforcement signal, and adjust its action selection strategy accordingly (Sutton & Barto, 1998). Unfortunately, it is often not possible for the agent designer to provide the agent with the ability to observe the whole state. Considering this limitation, the designer may decide to either consider a more complex model that explicitly accommodates for the perceptual limitations of the agent (Kaelbling, Littman, & Cassandra, 1998) or to ignore these limitations and treat whatever information is available as the agent’s complete state (Jaakkola, Singh, & Jordan, 1995).

A second challenge that the designer faces is a *task-modelling* challenge: given the adopted representation for the (state of the) agent, the designer must design a reinforcement signal that enables the agent to *learn* the

Instituto Superior Técnico, Universidade de Lisboa, Portugal

Corresponding author:

Pedro Sequeira, Instituto Superior Técnico, Universidade de Lisboa, Av. Prof. Dr. Cavaco Silva, 2744-016 Porto Salvo, Portugal.
Email: pedro.sequeira@gaips.inesc-id.pt

intended task as *efficiently* as possible. The design of reward functions is a difficult endeavour and has been the topic of intense research in the RL literature, which has led to interesting results regarding both *inverse reinforcement learning* (Ng & Russell, 2000; Ramachandran & Amir, 2007) and *reward shaping* (Ng, Harada, & Russel, 1999; Wiewiora, 2003).

Recent research regarding the origin of rewards in nature (Singh, Lewis, & Barto, 2009) and *intrinsically motivated reinforcement learning* (IMRL) (Singh, Lewis, Barto, & Sorg, 2010) has led to the formulation of the *optimal reward problem* (ORP) to address the task-modelling challenge discussed above. Roughly speaking, the ORP involves the discovery of a reward function from a set of possible rewards, which should induce the best “lifelong behaviour” possible for the agent in a set of environments of interest, as measured in terms of a target task.

Interestingly, results have indicated that by selecting the reward that best “solves” the ORP, it is often the case that *both* the agent- and task-modelling challenges can be successfully addressed (Sorg, Singh, & Lewis, 2010a). In particular, it is often possible to select a reward that not only enables the agent to learn the desired task in an efficient manner but also, as part of that process, mitigates the impact of the agent’s inherent limitations on its ability to successfully perform the task. Intuitively, the reward is used to provide the agent with implicit information about (parts of) the state that the agent would be unable to perceive otherwise. The ORP thus provides an appealing framework within which the RL agent designer can reason about rewards while alleviating part of the modelling burden associated with selecting good state representations.

However, the ORP raises a *new* design challenge: that of designing a rich set of possible rewards for the task at hand from which to select such an informative reward. Such an endeavour often involves significant domain knowledge, and several possibilities have been considered in the literature, which require varying levels of manual adjustment (Bratman, Singh, Lewis, & Sorg, 2012; Niekum, Barto, & Spector, 2010; Singh et al., 2010; Sorg, Singh, & Lewis, 2010b).

In this paper, we investigate the *nature of the rewards* to be considered in addressing the ORP. In particular, we want to construct a set of possible rewards that is *general enough* to alleviate the need for excessive adjustments across domains and also *informative enough* to provide useful information for each specific domain. We address the ORP within the framework of IMRL, in which the process of reward optimisation is interpreted as a computational counterpart to the evolutionary process that crafted reinforcement mechanisms in animals (Singh et al., 2010). Drawing inspiration from natural systems, we consider intrinsic reward mechanisms inspired by *appraisal theories of emotions*.

In a previous paper (Sequeira, Melo, & Paiva, 2011), we performed a preliminary study of the impact of emotion-based rewards on intrinsically motivated agents. In this paper, we focus on the design of a general-purpose reward mechanism and its impact on alleviating the demand for having to design specific rewards for different domains. The main technical contribution is the integration of a mechanism within IMRL that provides a reward built from a set of four domain-independent emotion-based features, namely, *novelty*, *valence*, *goal relevance* and *control*, each of which are inspired by a dimension of *appraisal of the emotional significance of events* and are commonly found in the psychology literature (Ellsworth & Scherer, 2003; Lazarus, 2001; Leventhal & Scherer, 1987; Roseman & Smith, 2001; Scherer, 2001). We perform such a mapping regardless of its *validity* in terms of appraisal theories, but we redesigned many of the previously proposed features to focus on emotions as a plausible source of such general-purpose and domain-independent intrinsic reward and discuss possible alternatives for each feature.

We illustrate the usefulness of the proposed reward design by comparing the performance of our “emotion-driven” RL agents with that of standard, goal-driven RL agents in several experiments that feature foraging scenarios. In addition, we extend our previous work by investigating the impact of maladaptive behaviours on the agent’s performance and the emergence of “universal” agents that behave well, on average, in all scenarios.

2 Related work

Early artificial intelligence (AI) research was mostly focused on the reproduction of human reasoning processes, e.g. by building systems that could prove theorems (Newell & Simon, 1956), solve algebra word problems (Bobrow, 1964) or understand English sentences (Winograd, 1971). Pioneering AI researchers also emphasised the role of *emotional processes* as attention-focusing, task-prioritising mechanisms that are crucial to any system that is to be regarded as intelligent (Minsky, 1986; Simon, 1967). Developments in neurophysiology brought prominence to the role of emotions in cognition (Damasio, 1994) and prompted the AI community to develop computational models of emotions, which are usually based on appraisal theories of emotions (Marsella, Gratch, & Petta, 2010). Many works in the area of affective computing (AC) address the impact of emotional processes on decision-making to create more engaging interactive artificial agents (Picard, 2000).

Within the area of AC, several works combine learning and emotional mechanisms in a complementary manner to create artificial agents that exhibit richer behaviour. For example, the model uses RL to build

emotion–object associations and predict the user’s actions (El-Nasr, Yen, & Ioerger, 2000). Another example is the work of Armony, Servan-Schreiber, Cohen, and LeDoux (1997), in which connectionist learning is used to emulate effects that are commonly associated with fear conditioning. The artificial creatures of Cañamero (1997) also use “low-level” emotional signals that drive behaviour selection. Jacobs, Broekens, and Jonker (2014) derive emotions of “joy”, “distress”, “hope” and “fear” from signals generated by the RL algorithm and demonstrate that the results of agent-based simulations are able to replicate the psychological and behavioural dynamics of emotion.

Another line of work, which is more closely related to our own, uses emotions to actually *influence* decision-making within RL. Gadanho (2003) proposed a bottom-up approach to emotion elicitation. Gadanho’s system uses artificial neural networks to determine a dominant emotional state. A measure of “wellbeing” (or valence) is calculated for each state by computing the relative change in the value of a set of homeostatic variables (energy, welfare and activity) and also accounting for predictions associated with that state. RL is used to learn state–behaviour associations, and the rewards are provided by the intensity and valence of the current dominant emotion.

Salichs and Malfaz (2006) proposed a set of basic emotions to control the behaviour of an RL agent. In their model, the reward depends on variations in the agent’s wellbeing. Their behaviour-selection mechanism uses a predefined level of “dare” that determines a preference for conservative (high-valued) actions over bad (low-valued) actions due to “fear”.

Marinier, Laird, and Lewis (2009) proposed an intrinsic reward signal based on the appraisal of *conductiveness*, which determines the sign of the reward value, whereas the intensity of the agent’s current *feeling* determines the magnitude of the signal. An experiment conducted in a grid-world scenario demonstrated that intermediate, emotion-based rewards lead to learning the task faster.

Broekens, Kusters, and Verbeek (2007) proposed associating positive affective states with exploitation, and negative states with exploration. They demonstrated that this model provides adaptive benefits for RL agents in specific scenarios.¹ The affective state and reward depend on the relation between the short- and long-term running averages of past reinforcement signals. Following this work, Hogewoning, Broekens, Eggermont, and Bovenkamp (2007) used a chi-squared statistical test to compute the significance of the differences between these two averages to influence action selection.

The work that is most similar to our proposed approach is that of Ahn and Picard (2006). In that work, the authors considered the use of extrinsic and intrinsic rewards, both to improve the learning performance of the agent and to influence decision-making.

The extrinsic reward relates to external goals, and the paper proposes a model for an *affective anticipatory reward* that is based on *valence* and *arousal* levels.

Much like in the work of Ahn and Picard (2006), our approach complements the (extrinsic) reward signal provided by the environment with an intrinsic reward signal that is constructed from a set of features based on major dimensions of emotional appraisal. In a sense, these features provide, at each time-step, a dynamical representation of the “emotional state” of the agent. This aspect is in contrast with most surveyed works, which rely either on a predefined set of discrete emotions or scalar evaluations of the emotional state of the agent.²

Additionally, unlike Ahn and Picard (2006), we do not treat the external and internal rewards differently. Instead, they are harmoniously combined to yield a single reward signal that guides the agent’s behaviour. The trade-off between such external and internal rewards is “fit” for the class of environments that the agent expects to encounter. A related optimisation can be found in biological agents, which process emotional states differently depending on their survival needs (Frijda & Mesquita, 1998; Roseman & Smith, 2001; Smith & Kirby, 2009). Moreover, we do not rely on predefined associations between emotional states and actions. Instead, the agent learns from the (combined) intrinsic rewards, an action selection rule that optimises the balanced benefit that arises from the environment and the agent’s internal state.

We also refer to our work in Sequeira, Melo, and Paiva (2014), which complements this paper. In that work, we test the emergence of emotion-related rewards by using evolutionary computation mechanisms. In other words, in this paper, we depart from the emotional appraisal literature to manually design reward features that have evaluative characteristics that are similar to those ascribed by some appraisal dimensions, whereas in Sequeira et al. (2014), we used genetic programming to determine a set of domain-independent reward features, and then analysed the dynamical and structural properties of those features in consideration of appraisal theories.

3 Background

This section describes the decision-theoretic framework within which we introduce our contributions. We discuss the models used throughout the paper to describe our agents, and present the basic nomenclature and notation.

3.1 (Partially observable) Markov decision problems

In its most general form, the sequential decision problem faced by RL agents can be modelled as a *partially observable Markov decision problem* (POMDP)

(Kaelbling et al., 1998), which is denoted as a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{Z}, \mathcal{P}, \mathcal{O}, r, \gamma)$. At each discrete time-step $t = 0, 1, 2, 3, \dots$, the environment is described by some state, which is represented as a random variable (r.v.) S_t that takes values from a finite set of possible states, \mathcal{S} . The agent makes an *observation*, which is denoted as an r.v. Z_t that takes values from a set of possible observations, \mathcal{Z} , that depend on the state S_t but that is often *insufficient* for the agent to unambiguously infer S_t . The agent then performs some action (which is denoted as an r.v. $A_t \in \mathcal{A}$), and the environment transitions to state S_{t+1} . This transition is governed by the probabilities $\mathbb{P}[S_{t+1} = s' | S_t = s, A_t = a] = P(s' | s, a)$. The agent then receives a numerical reward, $r(S_t, A_t)$, that represents the desirability of executing action A_t in state S_t (in terms of the target task) and makes a new observation Z_{t+1} , after which the process repeats. The observations Z_t of the agent are governed by the probabilities $\mathbb{P}[Z_{t+1} = z | S_{t+1} = s, A_t = a] = O(z | s, a)$. Traditional approaches to RL mainly focus on scenarios in which the observations Z_t allow the agent to unambiguously determine the underlying state S_t . Such scenarios are said to have *full observability*, and the POMDP parameters \mathcal{Z} and \mathcal{O} can be safely ignored. The resulting model, which is represented as a tuple $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{P}, r, \gamma)$, is simply referred to as a *Markov decision problem* (MDP).

In the traditional view of RL, the reward $r(s, a)^3$ “evaluates” the agent’s behaviour with respect to the task it must (learn to) perform, thereby acting as a *critic* that resides in the (external) environment, as depicted in Figure 1(a) (Singh et al., 2010). The goal of the agent is to select its actions to gather as much reward as possible during its lifetime, where the reward is discounted by some factor (Kaelbling et al., 1996; Sutton & Barto, 1998).

In an MDP, a *policy* is a decision-rule $\pi : \mathcal{S} \rightarrow \mathcal{A}$ that determines the action to be executed in each state $s \in \mathcal{S}$. We can associate with each MDP policy π , a *value function*, $V^\pi : \mathcal{S} \rightarrow \mathbb{R}$, that determines, for each initial state $s \in \mathcal{S}$, the value that the agent expects to receive by choosing its actions according to π . The value function is

$$V^\pi(s) = \mathbb{E}_\pi \left[\sum_t \gamma^t r(S_t, A_t) | S_0 = s \right] \quad (1)$$

where γ is a positive discount value such that $\gamma < 1$. An *optimal policy* is defined as any policy π^* such that $V^{\pi^*}(s) \geq V^\pi(s)$ for any state $s \in \mathcal{S}$ and policy π . The existence of one such policy can be guaranteed under mild assumptions regarding the MDP (Puterman, 1994). We can also associate with π^* a function $Q^* : \mathcal{S} \times \mathcal{A} \rightarrow \mathbb{R}$ that verifies the recursive relation

$$Q^*(s, a) = r(s, a) + \gamma \sum_{s' \in \mathcal{S}} P(s' | s, a) \max_{a' \in \mathcal{A}} Q^*(s', a') \quad (2)$$

where Q^* determines how good (in the long-run) each action is in each possible state faced by the agent, given that the latter performs optimally afterwards, and can be computed by iterating over equation (2) using a dynamic programming approach that is known as *value iteration*. Computation of Q^* using value iteration requires knowledge of the MDP parameters, namely, r and P . Reinforcement learning typically addresses situations in which one or both of these parameters are unknown. In those situations, the agent must learn the optimal policy by relying on data collected from the environment, either online (Watkins, 1989) or offline (Ernst, Geurts, & Wehenkel, 2005).

In this paper, we consider RL agents that follow the *prioritised sweeping algorithm* (Moore & Atkeson, 1993), which is an online RL algorithm that uses the data collected from the environment to construct estimates \hat{P} and \hat{r} of the parameters of the MDP. Such estimates are then used to perform, at each time-step, multiple-value iteration updates using a well-defined update schedule (which is implemented using a *priority queue*).

3.2 Intrinsically motivated reinforcement learning and the optimal reward problem

The RL paradigm described in Section 3.1 departs from a (PO)MDP model by describing a sequential problem faced by a decision-maker in a dynamic and uncertain world in which the task is implicitly encoded in the reward r . The performance of the agent depends on the ability of r to convey information about the task to be learned, and several works in the literature have addressed the problem of *reward design*. One approach relies on the idea of *shaping* (Mataric, 1994; Ng et al., 1999; Randløv & Alstrøm, 1998): given a reward r that encodes some target task, shaping consists of applying some transformation to r , thereby yielding a second reward, r' , that encodes *the same task* but is more informative for a learning agent. A second and more recent approach is to first construct a reward from a demonstration of the desired behaviour and then feed this reward to the agent (Ng & Russell, 2000). This approach is known as *inverse reinforcement learning* and has generated a significant amount of literature (Melo, Lopes, & Ferreira, 2010; Neu & Szepesvári, 2009; Ng & Russell, 2000; Ramachandran & Amir, 2007).

One radically different perspective on the problem of reward design arises from recent work on IMRL (Singh et al., 2010). IMRL seeks to model, within the RL framework, behaviours observed in nature that are not (directly) oriented towards “survival”, such as curiosity-driven behaviours (Singh et al., 2009, 2010). Within IMRL, the rewards arise from an evaluation of an “internal critic” of information both from the

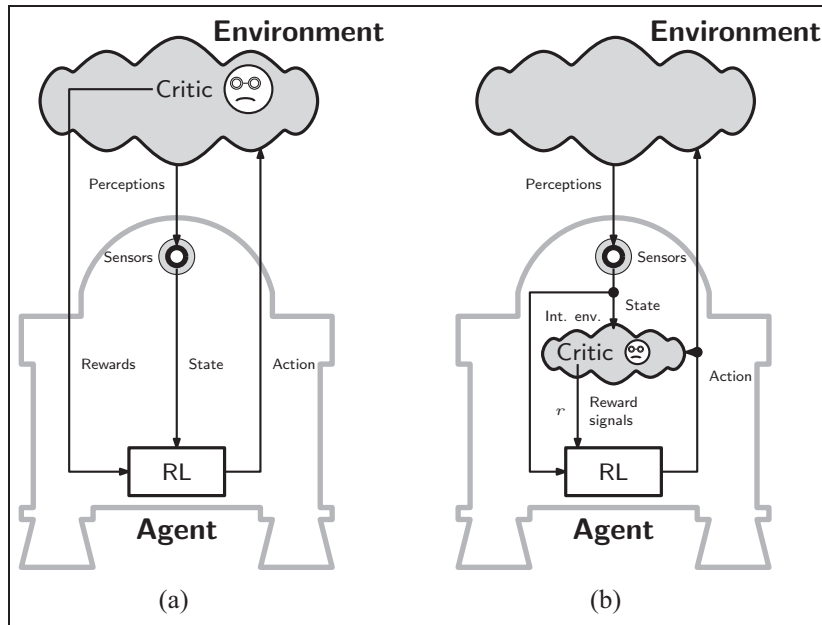


Figure 1. Comparison of the RL and IMRL frameworks: (a) the traditional RL model, in which a critic in the external environment evaluates the behaviour of the agent with respect to some target task; (b) the IMRL model, in which a critic in the agent's "internal environment" evaluates the behaviour of the agent and provides intrinsic rewards.

external environment and the agent's "internal environment", as depicted in Figure 1(b).

IMRL further proposes a distinction between *extrinsic rewards* (henceforth denoted as ρ), which evaluate the behaviour of the agent with respect to some environment-imposed task (e.g. survival), and *intrinsic rewards* (henceforth denoted as r), which evaluate the behaviour of the agent with respect to agent-specific "preferences" (Bratman et al., 2012). The evolutionary perspective discussed by Singh et al. (2010) argues that intrinsic rewards provide the agent with "evolutionarily shaped" mechanisms for optimally coping with the environments it expects to encounter.⁴ Computationally, IMRL was distilled into the ORP (Sorg et al., 2010a).

Definition 1 (Optimal reward problem (ORP)). Given a learning agent U , a set of possible environments \mathcal{E} that agent U may inhabit, and a target task T to be learned, which reward r , among a set \mathcal{R} of possible rewards, induces the best "lifelong performance" in the agent U , measured with respect to the target task T ?

The ORP thus proposes an explicit separation between the *goal of the agent designer*, which concerns the behaviour of the agent with respect to the target task T , and the *goal of the RL agent* itself, which concerns the agent's behaviour with respect to whichever (intrinsic) reward r it receives. Performance with respect to the latter goal, as is standard in RL, is usually measured in terms of the total discounted (intrinsic) reward accumulated over time. As for the former goal, we start by observing that the behaviour of the agent is defined by (i) the POMDP used to model the environment with

which the agent interacts and (ii) the decision algorithm used by the agent. Together, they specify the set of possible interactions that the agent can experience.

Formally, let \mathcal{H} denote the set of all possible *finite histories* that the agent can experience throughout its lifetime. In particular, we consider an element $h \in \mathcal{H}$ as a sequence $h_{1:t} = \{z_1, a_1, \rho_1, z_2, \dots, a_{t-1}, \rho_{t-1}, z_t\}$, where z_τ , a_τ and ρ_τ denote, respectively, the *observation* at time-step τ , the *action* selected at time-step τ , and the *extrinsic reward* at time-step τ . Referring back to Figure 1(b), the internal critic is responsible for processing the agent's perceptions into a history h that contains information about the environment (in the form of a sequence $\{z_1, \dots, z_t\}$) and information about the extrinsic reward (in the form of a sequence $\{\rho_1, \dots, \rho_{t-1}\}$).

Additionally, let r denote the *intrinsic reward* that drives the behaviour of the agent (which is modelled as a POMDP $\mathcal{M} = (\mathcal{S}, \mathcal{A}, \mathcal{Z}, \mathcal{P}, \mathcal{O}, r, \gamma)$). We refer to the remaining parameters of the POMDP as the *environment of interest*, e , and write $\mathbb{P}[h|r, e]$ to denote the probability of observing history $h \in \mathcal{H}$ given r and e .

We define the *fitness function*, $f: \mathcal{H} \rightarrow \mathbb{R}$, that maps each history $h \in \mathcal{H}$ into a numerical value that evaluates the performance of the agent with respect to the target task T . Given a *space of possible rewards*, \mathcal{R} , and a distribution p_{env} over the set of environments, \mathcal{E} , the ORP can thus be formulated as the problem of determining the *optimal reward* $r^* \in \mathcal{R}$ such that

$$r^* = \operatorname{argmax}_{r \in \mathcal{R}} \mathcal{F}(r) \triangleq \mathbb{E}_{e \sim p_{\text{env}}} [f(h) | r, e] \quad (3)$$

where $\mathcal{F}(r)$ is the expected fitness associated with reward r . In this paper, we consider the fitness of an agent U throughout a particular history as the *total extrinsic reward* accumulated therein, i.e.

$$f(h_{1:t}) = \sum_{\tau=1}^t \rho_{\tau} \quad (4)$$

This particular choice of fitness function implicitly indicates that ρ directly measures the fitness of the agent and, therefore, we interchangeably refer to ρ as the *extrinsic reward* and the *fitness-based reward*.

4 Designing emotion-based rewards

This section introduces our main technical contribution, which consists of a set of reward features that are inspired by four common dimensions of emotional appraisal within the IMRL framework.

4.1 Appraisal theories of emotions

Given the potential advantages of emotional processing mechanisms in artificial agents (Picard, 2000), we now investigate how to port such mechanisms to the IMRL framework by considering well-established *appraisal theories of emotion* (ATEs) (Arnold, 1960; Ellsworth & Scherer, 2003; Roseman & Smith, 2001).

ATEs propose that the elicitation of an emotional state is preceded by an *appraisal* of the significance of the individual's situation in terms of its wellbeing and goals (Arnold, 1960). ATEs investigate the *functional aspect of emotions* and seek to explain the effects of

appraisals in decision-making and, more generally, behavioural and cognitive responses to the perceived situation. These responses contribute to focus the individual's attention on significant aspects of its environment (Frijda & Mesquita, 1998; Lazarus, 2001; Leventhal & Scherer, 1987).⁵

Figure 2 provides a high-level illustration of the process of emotional elicitation according to ATEs, which combines information from external stimuli and the individual's internal states—the person–environment relationship—and provides an evaluation of the situation. The outcome of appraisal leads to an emotional state that may induce a set of responses, including the physiological signals and bodily expressions that are responsible for the *subjective feelings* of emotions. Appraisal can take place at *different levels* (Ellsworth & Scherer, 2003; Leventhal & Scherer, 1987): e.g. the basic fight-or-flight evaluation observed in humans and other animals that occurs when a dangerous situation is faced is different from the more cognitive assessment that occurs after the death of a close family member. According to the level at which appraisal takes place, the type of information processed by the subject differs; the information is simpler at low levels and more complex at higher cognitive levels.

Many ATEs in the literature propose *structural models* in which emotions are elicited by evaluations of events through a set of *appraisal variables* (Frijda & Mesquita, 1998; Lazarus, 2001; Roseman, 2001; Scherer, 2001). Each variable is conceptualised as a *dimension* along which appraisal outcomes may vary continuously (Roseman & Smith, 2001). The several dimensions define the criteria used to evaluate a

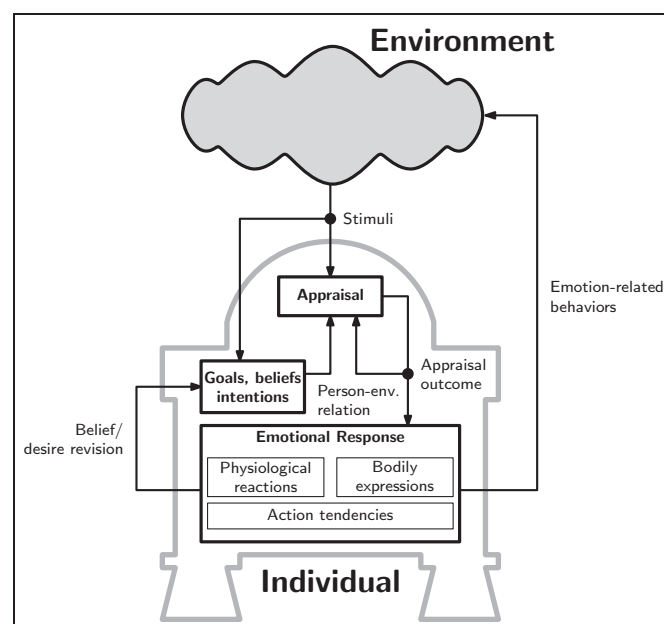


Figure 2. The elicitation of emotional responses as the result of an evaluation of the situation (the stimuli) in relation to the individual's goals, beliefs and intentions from the perspective of appraisal theories of emotion.

situation and ascribe the *structure* or the contents of the appraisal (Ellsworth & Scherer, 2003; Roseman & Smith, 2001).

Most of the appraisal dimensions proposed in the literature address universal, culturally independent evaluations of the personal significance of events. By combining specific values or outcomes of the appraisal dimensions, these theories can model discrete emotions (such as joy, sadness, and fear) and predict the particular physiological responses and action tendencies that are associated with each of them (Ellsworth & Scherer, 2003; Frijda & Mesquita, 1998; Roseman & Smith, 2001). Therefore, most ATEs largely agree regarding which dimensions are necessary to evaluate a given situation. Ellsworth and Scherer (2003) compared the most common ATEs and identified the following set of five *major dimensions of appraisal*, for which there is broad consensus in the community, and on which our approach is based: novelty, pleasantness/valence, goal relevance, power/coping potential and normative/social significance.

4.2 Learning and partial observability

The presentation in Section 3 focused mostly on the benign situation in which the agent, at each time-step, is able to completely observe the state S_t of the environment. However, in real-world scenarios, this is seldom the case. For example, a robot's perception about the state of the world is limited to the accuracy and resolution of its sensors. The POMDP model briefly discussed above enables the agent to reason about information that its observations yield about the actual state of the environment. Unfortunately, POMDP models are significantly more elaborate than MDPs, both conceptually and algorithmically. In fact, while MDPs are efficiently solvable, i.e. an optimal policy for an MDP can be computed rather efficiently (Puterman, 1994), their partially observable counterparts were proven to be undecidable in the worst case (Madani, Hanks, & Condon, 1999).

Given the difficulty inherent in reasoning about partial observability, one possible approach is to ignore partial observability altogether and reason about the observations of the agent as if they were actual states (Jaakkola et al., 1995). Another approach is to rely on the agent observations to track the most likely state of the environment and select the actions accordingly (Littman, Cassandra, & Kaelbling, 1995). In highly structured problems (e.g. robotic navigation), this simple approach can actually yield good results (Cassandra, 1998). However, in general, such simplified solutions are bound to lead to poor performance, as demonstrated by the work of Singh, Jaakkola, and Jordan (1994). Moreover, computing the best such solution is typically a difficult problem (Littman, 1994). Other approaches to address partial

observability in RL settings, that build into the agent's prior knowledge, and can somehow alleviate its perceptual limitations have been proposed (Aberdeen, 2003). Examples include approaches that are based on some form of memory (McCallum, 1995). However, such approaches typically require very specific algorithms that are tailored to leverage information from particular aspects of the agent's history (Aberdeen, 2003).

The IMRL framework discussed in Section 3.2 provides an elegant framework within which it is possible to implicitly "supply" prior knowledge to the learning agent. In fact, by properly tuning the reward, it is possible to induce in the agent, behaviours that may not be directly related to the target task but which, in time, can mitigate the impact of the agent's limitations on its performance (Sorg et al., 2010a). However, as discussed in Section 3.2, an adequately informative reward is critically dependent on the considered set of rewards, \mathcal{R} , which yields a new design challenge—that of designing the set \mathcal{R} of possible rewards for a desired task. As discussed above, this challenge often requires significant domain knowledge, and several possibilities have been considered in the literature (Bratman et al., 2012; Niekum et al., 2010; Sorg et al., 2010b).

Below, we propose a set of domain-independent reward features that are inspired by appraisal theories of emotions and can be used as building blocks to construct richer sets of reward functions.

4.3 Emotion-based reward design

We are now in a position to introduce our main technical contribution. We depart from the discussion of ATEs in Section 4.1 and propose a set of *reward features* that are inspired by each of the major dimensions of appraisal.

Going back to the IMRL agent architecture in Figure 1(b), we recall that the internal critic provides the RL decision-making component with reward information. This reward is, in turn, constructed using information both from the external environment (the sensations, including the extrinsic reward) and the agent's internal environment. Drawing a parallel with the process of appraisal depicted in Figure 2, we can approximately identify the internal critic in our IMRL agent as the module in which appraisal occurs. The reward r used for learning and decision-making approximately corresponds to the outcome of such a process.

We treat the agent's perceptions as states, which is a common simplifying approach that was already discussed in Section 4.2. This approach is equivalent to considering, in the POMDP model, that $\mathcal{S} = \mathcal{Z}$ and $\mathbb{P}[S_t = s | Z_t = s] = \mathbb{P}[Z_t = s | S_t = s] = 1$. Therefore, we henceforth omit any explicit references to observations, with the understanding that "states", as perceived by the agent, actually correspond to POMDP observations. In practice, as discussed above, it is seldom the

case that $\mathcal{S} = \mathcal{Z}$, and our assumption will provide an opportunity to assess the ability of our approach to overcome the impact of disregarding partial observability issues.

We consider a set of possible rewards \mathcal{R} in which each reward r is a linear combination of some predefined *reward features*, $\{\phi_i, i = 1, \dots, N\}$. Each feature ϕ_i maps perception–action–history triplets (which are abusively denoted as (s, a, h) , given our treatment of perceptions as state) to a scalar value $\phi_i(s, a, h) \in \mathbb{R}$. For every $r \in \mathcal{R}$

$$r(s, a, h) = \sum_{i=1}^N \phi_i(s, a, h) \theta_i = \boldsymbol{\phi}^\top(s, a, h) \boldsymbol{\theta}$$

where θ_i is the linear coefficient that is associated with feature ϕ_i in r .

We propose that each appraisal dimension maps to a corresponding reward feature ϕ_i , $i = 1, \dots, N$. Much like appraisal dimensions in biological agents, our reward features evaluate the significance of the agent's current situation for its "wellbeing" according to specific criteria (Singh et al., 2009, 2010). Our approach thus follows the perspective that appraisal corresponds to a multi-dimensional, continuous-valued evaluation (Ellsworth & Scherer, 2003; Scherer, 2001). Given the simplicity of the RL agent model considered here, our emotion-based reward features rely on low-level statistical "summaries" of the agent's history of interaction with the environment.

Because we are focusing on single-agent scenarios, we adopt only four of the aforementioned major dimensions of appraisal, namely, *novelty*, *valence*, *goal relevance* and *control* (Ellsworth & Scherer, 2003).⁶ Our features are constructed from information that is usually available to RL agents and are therefore *general* and *domain-independent*. The value of each reward feature $\phi_i(s, a, h)$ somehow indicates the *degree of activation/significance* of dimension i associated with the execution of action a after perceiving s and given a history of interaction h .

Formally, our set of rewards, \mathcal{R} , is the linear span of the set $\{\phi_n, \phi_t, \phi_c, \phi_v, \phi_p\}$, where

- $\phi_n(s, a, h)$ denotes the *novelty* associated with performing action a after observing s , given the history h ;
- $\phi_t(s, a, h)$ denotes the *goal relevance* of performing action a after observing s , given h ;
- $\phi_c(s, a, h)$ denotes the degree of *control* over the outcome of executing action a after observing s , given h ;
- $\phi_v(s, a, h)$ denotes the *expected valence* of executing a after observing s , given history h ;
- Finally, $\phi_p(s, a, h) = \hat{p}(s, a)$ is not an emotion-based feature. Rather, it corresponds to the *estimated*

fitness-based reward for executing a after observing s , $\hat{p}(s, a) = \mathbb{E}[\rho_t | S_t = s, A_t = a]$.⁷

Below, we describe each of the aforementioned features in detail.

In the RL framework, familiarity with states and actions is directly related to the number of visits to state–action pairs. Let us denote by $n_t(s)$ the number of times that s was perceived up to time-step t and by $n_t(s, a)$ the number of times that action a was selected after perceiving s . We thus quantify the dimension of novelty as

$$\phi_n(s, a, h_{1:t}) = \frac{1}{2} [\lambda_n^{-n_t(s,a)} + \lambda_n^{-n_t(s)}] \quad (5)$$

where λ_n is a positive constant such that $\lambda_n < 1$. The two terms in equation (5) account for the novelty in terms of experienced actions and the novelty in terms of perceived states, respectively. λ_n can be considered a "novelty rate" that determines how novelty decays with experience.

The expression proposed for novelty is related to the *inverse-frequency* feature of Bratman et al. (2012). However, instead of a linearly decaying rate, we consider an exponentially decaying rate that is dependent on, for example, the total number of states and actions that can be experienced, or the agent's lifetime. Additionally, we adopt a frequency-based feature rather than a recency-based feature, because the former better captures the essence of the novelty dimension.⁸ Our calculation of the novelty feature evaluates the amount of past experience only in terms of perceived states and performed actions. However, one can envisage expressions that evaluate the *predictability* of stimuli or the *probability* of actions outcomes that are consistent with the corresponding novelty dimension in biological agents (Ellsworth & Scherer, 2003; Leventhal & Scherer, 1987).

Therefore, goal relevance has a *motivational basis* and is influenced by the importance of the event and the consistency of its outcomes in relation to the goals or needs under consideration (Roseman, 2001). Broadly speaking, the goal relevance of an event increases if such an event is consistent with, or conducive to, the achievement of the individual's goals, and decreases when the consequences of the event are *obstructive* to reaching those goals (Ellsworth & Scherer, 2003; Reisenzein, 2009).

At a very low level, the goal of an individual is to attain maximum fitness throughout its lifetime. Let $V_\rho^{(t)}$ denote the estimate, at time-step t , of the value function associated with only the fitness-based reward, ϕ_p , which satisfies the fixed-point relation $V_\rho^{(t)}(s) = \max_{a \in \mathcal{A}} Q_\rho^{(t)}(s, a)$, where $Q_\rho^{(t)}$ denotes the estimate of the action-value function associated with only the fitness-based reward. States for which $V_\rho^{(t)}$ is high should then be

preferable over those with a low value of $V_\rho^{(t)}$. We define the *estimated goal state* at time-step t , $s_\rho^{(t)}$, as $s_\rho^{(t)} = \operatorname{argmax}_{s \in \mathcal{S}} V_\rho^{(t)}(s)$ and let $\hat{d}_t(s)$ denote the estimate, at time-step t , of the number of steps needed to reach $s_\rho^{(t)}$ from s , given the agent's current model of the environment.⁹ In our framework, goal relevance is thus expressed as

$$\phi_t(s, a, h_{1:t}) = \frac{1}{1 + \hat{d}_t(s)} \quad (6)$$

This expression is consistent with the role of goal relevance in biological agents, according to ATEs. In particular, it decreases the relevance of states that are farther from the (perceived) goal and is maximal when the agent reaches the goal. As a measure of distance, we used the estimated number of steps required to reach the goal, which is a generalisation of the Manhattan distance proposed by Bratman et al. (2012). However, unlike the Manhattan distance, \hat{d}_t does not require any specific metric structure in the underlying MDP state-space, aside from the structure that is naturally induced by the transition probabilities P .

At a higher level (i.e. more cognitive) of appraisal, these evaluations often require adjusting either the significance of the situation at hand (Ellsworth & Scherer, 2003; Roseman, 2001) or the individual's goals to cope with the possible outcomes of the event (Lazarus, 2001; Smith & Kirby, 2009). At a lower level of processing, these evaluations simply assess the extent to which an event or its outcomes are *controllable* and whether the individual has the ability to change the situation to its benefit (Frijda & Mesquita, 1998; Roseman, 2001).

We adopt the perspective that control over a situation is often directly related to the degree of *predictability* of the outcomes under consideration (Ellsworth & Scherer, 2003; Leventhal & Scherer, 1987; Roseman, 2001). The ability of an RL agent to control its environment is directly related to the *accuracy* of its world model. Accurate world models allow the agent to reason correctly about which actions maximise its reward/fitness, whereas inaccurate world models may cause the agent to often select suboptimal actions.¹⁰

To measure the accuracy of the agent's world model, we determine how well $Q_\rho^{(t)}$ satisfies the relation in equation (2) given estimates $\hat{\rho}$ of the fitness-based reward. Specifically, we measure how the most recent information perceived by the agent impacts on its current estimate by defining the *prediction error* associated with $Q_\rho^{(t)}(s, a)$ whenever s, a are experienced at time-step t as $\Delta Q_\rho^{(t)}(s, a) = k \cdot |Q_\rho^{(t)}(s, a) - Q_\rho^{(t-1)}(s, a)|$, where k is a normalising constant and $Q_\rho^{(t-1)}(s, a)$ corresponds to the previous value computed for $Q_\rho^{(t)}(s, a)$, i.e. $t-1$ corresponds to the previous time-step in which action a was executed given state s . Denoting by $\mathcal{T}_{s,a}$ the set of all time-steps in which the state-action pair s, a was

experienced, we define the control feature $\phi_c(s, a, h_{1:t})$ according to the negative running average prediction error associated with $Q_\rho^{(t)}(s, a)$, i.e.

$$\phi_c(s, a, h_{1:t}) = 1 - \frac{1}{n_t(s, a)} \sum_{\tau \in \mathcal{T}_{s,a}} \Delta Q_\rho^{(\tau)}(s_\tau, a_\tau) \quad (7)$$

From the above expression, we note that $\phi_c(s, a, h_{1:t})$ is close to 0 for those state-action pairs that, throughout the agent's history, are hardest to "learn". Conversely, $\phi_c(s, a, h_{1:t})$ is close to 1 for those state-action pairs that the agent learns quickly. Note also that as the agent's knowledge of the environment improves, so does the value of $\phi_c(s, a, h_{1:t})$. This feature thus provides a meaningful measure of the agent's predictive ability.

We conclude by noting that the feature ϕ_c is somewhat related to the *quality-of-model* feature proposed by Sorg et al. (2010a), which also accounts for discrepancies in the state transition model. The control feature is also related to works that measure the *model accuracy* and *learning progress*, such as Lopes, Lang, Toussaint, and Oudeyer (2012) and Moulin-Frier and Oudeyer (2013), which cause agents to progressively explore regions of the state-space that seem more complex and interesting. In the AC literature, ϕ_c is also related to the *uncertainty model* proposed by Ahn and Picard (2006), which calculates the level of emotional arousal of an agent by considering discrepancies between the value of the current action in the current state and the expected reward associated with the current action as observed in other states.

At such a low level, in our IMRL framework, valence is perhaps best represented as the fitness-based reward itself, ϕ_ρ , because it provides an immediate direct evaluation of the perceived states and executed actions in terms of the associated fitness. However, as observed in Section 3.2, ϕ_ρ is external to the agent and fails to take into account any experience that the agent may accumulate. Alternatively, we adopt the idea that the *implicit value* of things can change throughout time according to experience (Cardinal, Parkinson, Hall, & Everitt, 2002; Ellsworth & Scherer, 2003; Leventhal & Scherer, 1987). Bearing this idea in mind, and to account for the integration of experience in the valence dimension of appraisal, we evaluate the value of the agent's current situation (with respect to fitness), both in terms of the perceived state and in terms of the experienced action.

Formally, we define valence as

$$\phi_v(s, a, h_{1:t}) = \frac{1}{2} \left[\frac{V_\rho^{(t)}(s) - V_\rho^{(t)}(\min)}{V_\rho^{(t)}(\max) - V_\rho^{(t)}(\min)} + \frac{Q_\rho^{(t)}(s, a) - Q_\rho^{(t)}(\min)}{Q_\rho^{(t)}(\max) - Q_\rho^{(t)}(\min)} \right] \quad (8)$$

where $V_\rho^{(t)}(\max) = \max_{s \in \mathcal{S}} V_\rho^{(t)}(s)$, $Q_\rho^{(t)}(\max) = \max_{a \in \mathcal{A}} Q_\rho^{(t)}(s, a)$, $V_\rho^{(t)}(\min) = \min_{s \in \mathcal{S}} V_\rho^{(t)}(s)$, $Q_\rho^{(t)}(\min) = \min_{a \in \mathcal{A}} Q_\rho^{(t)}(s, a)$.

As observed from the above expression, $\phi_v(s, a, h)$ is maximised when the agent executes the best action in the most valuable state, which implies a *learned preference* for behaviours that the agent believes will lead to a high degree of fitness in the environment.

An alternative for this expression was proposed by Ahn and Picard (2006). In this work, the agent feels “good” or “bad” depending on how the expected reward received after executing an action in the current state compares with rewards that were previously experienced in that state. However, such an expression evaluates only the immediate reward and mostly ignores the agent’s long-term goals. Another alternative formulation can be found in Broekens et al. (2007). In this formulation, a short-term average of received reinforcements is measured against its long-term running average to construct the reward (and valence) with which the agent learns. The analysis of valence is therefore made according to the past actions executed, whereas our proposal for valence reflects possible changes of preference towards stimuli as given by future courses of action. We also refer to the “wellbeing” (or valence) feature proposed in Gadanho (2003), which is calculated for each state as its relative impact and prediction value using a set of homeostatic variables. Positive/negative changes and predictions lead to positive/negative states of wellbeing.

5 Experiments and results

To evaluate our emotion-based intrinsic rewards, we performed a set of experiments in foraging environments inspired by those of Singh et al. (2010). We model our agent as a predator that tries to feed on prey throughout time. The choice of these scenarios is tightly connected with the objectives of the experiments. First, foraging scenarios enable a clear evaluation of the agent’s behaviour in terms of the target task. In particular, the extrinsic reward p reinforces feeding behaviours that are closely related with fitness (as is the case for biological agents).

Second, foraging scenarios also facilitate evaluation of the impact of the different reward features on the behaviour of the agent. As will soon become apparent, different environments require different policies to attain maximal fitness. Our reward features, if properly combined, induce policies that attain maximal fitness and overcome the limitations of the agent. In contrast, if poorly combined, they lead to poor performance, thereby mirroring what can also be observed in nature—poorly adapted individuals usually perform poorly.

Foraging scenarios, given their ease of interpretation, also simplify the assessment of whether our emotion-based reward mechanism yields advantages when designing artificial learning agents. As our results

illustrate, the partial observability of the state, which is enforced in most scenarios, prevents our agents from learning the target task (feeding) given only the extrinsic reward. Therefore, much like biological agents in nature, our agents must engage in behaviours that are not directly related to fitness enhancement but which often (indirectly) lead to a more successful “feeding policy”.

5.1 Experimental setup

We used a total of six scenarios (see Figure 3), which are either taken directly from the IMRL literature or modifications thereof (Singh et al., 2010; Sorg et al., 2010a). We describe these scenarios below.

Hungry–Thirsty scenario: This scenario is adapted from the work of Singh et al. (2010) and is depicted in Figure 3(a): It contains two inexhaustible resources: food and water. The resources can be positioned at any of the corners of the environment (positions (1 : 1), (5 : 1), (1 : 5), and (5 : 5)), thereby yielding a total of 12 possible configurations of food and water (only one of which is depicted in Figure 3(a)). The agent’s fitness is defined as the amount of food consumed. However, the agent can only consume food if it is not thirsty, a condition that is achieved only by consuming the water resource (drinking). At each time-step after drinking, the agent becomes thirsty again with a probability of 0.2. The agent observes its position and thirst status (either thirsty or not thirsty).

Lairs scenario: This scenario is an adaptation of the “boxes” scenario of Singh et al. (2009, 2010). One possible layout of the environment is depicted in Figure 3(b). There are two lairs positioned at different corners of the environment, thereby resulting in six possible configurations. The fitness of the agent is defined as the number of prey captured. Whenever a lair is occupied by a prey, the agent can drive the prey out by means of a *Pull* action. The state of the lair transitions to prey outside, and the agent has exactly one time-step to capture the prey with a *Capture* action before the prey runs away. In either case, the state of the lair transitions to empty. At every time-step, there is a probability of 0.1 that a prey will appear in an empty lair. The agent is able to observe its position and the state of both lairs (occupied, empty, or prey outside).

Moving-Prey scenario: This scenario is also adapted from the work of Singh et al. (2010), and one possible configuration is depicted in Figure 3(c). In this scenario, at any time-step, there is *exactly one prey* available, and the prey is located at one of the end-of-corridor locations (positions (3 : 1), (3 : 3) or (3 : 5)). The agent’s fitness is again defined as the number of prey captured. Whenever the agent captures a prey, the latter disappears from the current location and a new prey

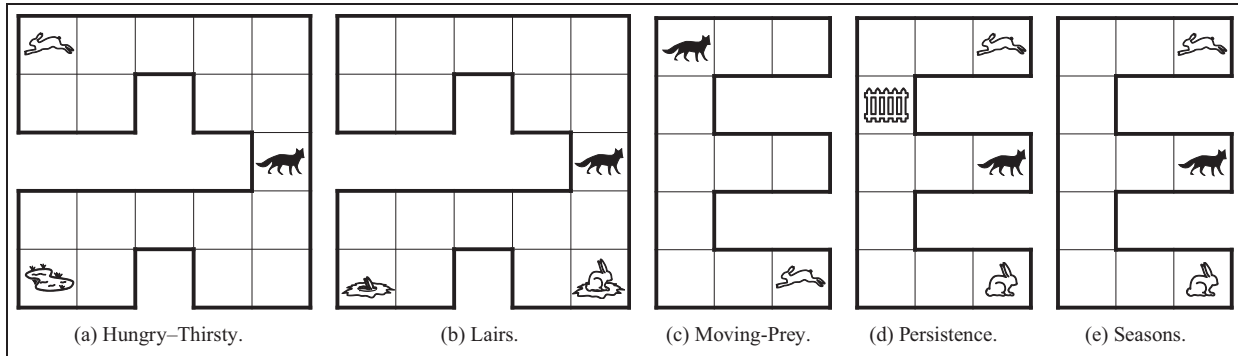


Figure 3. Possible environment configuration for the several foraging scenarios used in the experiments. In all environments, we represent our agent and its corresponding starting position by the dark fox figure. We refer to a cell in column x and row y as having position $(x : y)$. See the text for specific descriptions of the dynamics of each scenario.

randomly appears at one of the two other possible prey locations.

Persistence scenario: The environment used in this scenario is depicted in Figure 3(d). In this scenario, the environment contains two types of prey that are always available. *Hares* are located at $(3 : 1)$ and contribute to the fitness of the agent with a value of 1 when captured, whereas *Rabbits* are located at $(3 : 5)$ and contribute with a value of 0.01. Whenever the agent captures a prey, its position is reset to the initial position, $(3 : 3)$. The environment also contains a *fence*, which is located at $(1 : 2)$, that prevents the agent from easily capturing hares. To cross the fence towards the hare location at time t , the agent must perform action N for N_t consecutive time-steps, after which the fence is reinforced, thereby requiring an increasing number of actions N for it to be crossed.¹¹ The agent does not know how many steps it takes to cross the fence (or whether crossing is possible).

Seasons scenario: The environment used in this scenario is portrayed in Figure 3(e) and contains two types of prey. *Hares* appear at $(3 : 1)$ and contribute to the agent's fitness with a value of 1, whereas *Rabbits* appear at $(3 : 5)$ and contribute with a value of 0.1. As in the Persistence scenario, the agent's position is reset to $(3 : 3)$ upon capturing any prey. However, in this scenario, only one prey is available at each time-step, depending on the season, which changes every 5000 time-steps.¹² Additionally, in the Rabbit season, for every 10 rabbits that it captures, the agent is attacked by the farmer, which negatively impacts its fitness by a value of -1 . The agent knows neither the current season nor how many rabbits it has consumed since it was last attacked.

Poisoned-Prey scenario: This scenario is a variation of the the Seasons scenario. The layout and prey positions are the same, but both rabbits and hares are always available to the agent. Rabbits contribute to the fitness of the agent with a value of 0.1. Hares, when healthy, contribute a positive amount of 1. When poisoned, they

contribute a negative value of -1 . As in the Seasons scenario, the health status of hares changes every 5000 steps. Again, the agent knows neither the current season nor whether a prey is poisoned.

5.1.1 Agent description. In all scenarios, the agent is modelled as a POMDP whose state dynamics follow from the above descriptions. In all scenarios, the agent has four available actions, $\mathcal{A} = \{N, S, E, W\}$, that deterministically move it in the corresponding direction. In the Lairs scenario, the agent has also *Pull* and *Capture* actions available. Prey are captured automatically whenever they are co-located with the agent. In all but the Hungry-Thirsty and Lairs scenarios, the agent is only able to observe its current $(x : y)$ position and whether it is co-located with a *prey*.

In all scenarios, we treat observations as states and use prioritised sweeping to learn a policy that maps observations to actions (Moore & Atkeson, 1993). As discussed in Section 3, prioritised sweeping constructs a model of the environment and uses this model to perform value-iteration updates. Specifically, our agent maintains an estimate $\hat{P}^{(t)}(s' | s, a)$ of the transition probabilities as perceived by the agent, which is given by $\hat{P}^{(t)}(s' | s, a) = \frac{1}{n_t(s, a)} \sum_{\tau=1}^t \mathbb{I}(s, a, s') (s_\tau, a_\tau, s_{\tau+1})$. The reward features discussed in Section 4.3 are then used to build the intrinsic reward and thus compute the associated optimal Q -function, Q^* . In our experiments, prioritised sweeping updates the Q -values of up to 10 state-action pairs in each iteration using a learning rate of $\alpha = 0.3$. During its lifetime, the agent uses an ϵ -greedy exploration strategy with a decaying exploration parameter $\epsilon_t = \lambda^t$, where $\lambda = 0.999$. We use a novelty rate $\lambda_n = 1.001$ for the computation of the novelty reward-feature in equation (5). In all experiments, we consider a discount of $\gamma = 0.9$.

5.1.2 Reward parameter optimisation. We consider the space of rewards, \mathcal{R} , as the set of all rewards of the

form $r(s, a, h) = \phi(s, a, h)^\top \theta$, where $\phi(s, a, h)$ is the set of all reward features described in Section 4.3 and θ is the vector that contains the corresponding parameters that represent the *weight* or *contribution* of each feature to the overall reward. To determine the optimal reward function r^* (or, equivalently, the corresponding optimal parameter vector θ^*) for each of the different (set of) environments considered, we adopt the simple approach of Singh et al. (2010). In particular, we restrict the parameter vector to lie in the five-dimensional hypercube $I = [-1; 1]^5$ and sample a total of $K = 14,003$ uniformly distributed parameter vectors from I , where we enforce $\|\theta_k\|_1 = 1, k = 1, \dots, K$.

As discussed in Section 3.2, we consider the fitness function defined in equation (4). To evaluate the fitness of an agent driven by reward $r_k = \phi^\top \theta_k, k = 1, \dots, K$, we perform a total of $N = 200$ independent Monte Carlo trials, each of which consists of a continuous run of 100,000 learning steps. During each trial, the agent is allowed to interact with and learn from the environment. The fitness of the agent given a reward function r_k is then measured as the average fitness across the N trials, i.e. $F(r_k) \approx \frac{1}{N} \sum_{i=1}^N f(h^i)$, where h^i is the history of the agent at trial i . Finally, we select the optimal parameter vector, θ^* , such that $\theta^* = \operatorname{argmax}_{\theta_k, k=1, \dots, K} F(r_k)$.

5.2 Results

We now describe the results of our experiments, which are detailed in Table 1. For each scenario, we indicate the optimal parameter vector, θ^* , that results from the parameter optimisation procedure described in Section 5.1. We then compare the fitness attained by our “emotion-driven” RL agent, which is driven by $r^* = \phi^\top \theta^*$, with that of a “standard” RL agent, which is driven by a reward $r_E = \phi^\top \theta_E$ (where $\theta_E = [0, 0, 0, 0, 1]^\top$) that considers only the extrinsic component, and a “random” RL agent, which is driven by a reward $r_0 = \phi^\top \theta_0$ (where $\theta_0 = [0, 0, 0, 0, 0]^\top$) that ignores all reward information. The objective is to assess the usefulness of the proposed emotion-based features through

comparison with an agent that is driven only by the designer’s extrinsic reward. The random agent provides a baseline for comparison.¹³

The comparative results of the experiments in the Hungry–Thirsty scenario in Table 1 show that the emotion-driven agent clearly outperforms the standard RL agent. The difference in performance between the two agents is statistically significant for a p value of < 0.02 .¹⁴ Figure 4(a) further supports our conclusions by providing a depiction of the learning performance of all agents. We also emphasise that the behaviour of our agent is driven by a combination of reward features that evaluate aspects of its interaction with the environment, most of which have little relation with the semantics of the domain (namely, with the agent’s hunger or thirst status or the presence of food or water in its position; see Table 1). This result contrasts with a previous approach that relied on domain-dependent state information to construct the reward (Singh et al. 2009).¹⁵

A comparison of the results of the experiments in the Lairs scenario is presented in Table 1 and Figure 4(b). Again, the results indicate a statistically significant difference (at $p = 10^{-4}$) in the performance of the “emotion-driven” and “standard” RL agents. By considering the learned policies of both agents, we observe that the emotion-driven agent learned to go from lair to lair and successively pull and capture rabbits as soon as a lair became empty. In comparison, the standard RL agent typically focused on one of the two lairs and captured only the rabbits in that lair. In spite of the small probability of each lair transitioning from empty to occupied, waiting for a rabbit at a single lair is not the best strategy. The observed policy is also consistent with the findings in the “boxes” experiments of Singh et al. (2010); again, the main difference between that work and ours is that our agent relies on domain-independent reward features. By analysing the optimal parameter vector θ^* in Table 1, we can observe a small preference for exploratory behaviour (expressed as a positive weight in *novelty*) and less predictable states (expressed as a negative weight in *control*). Given the dynamics of

Table 1. The mean cumulative fitness and optimal parameter vector (θ^*) for each foraging scenario. We compare the performance of the optimal “emotion-driven” agent using r^* with that of a “standard” RL agent using r_E and an agent acting “randomly” using r_0 . The fitness results correspond to averages calculated over 200 independent Monte Carlo trials.

Scenario	Optimal parameter vector						Mean fitness		
	$\theta^* = [$	$\theta_n,$	$\theta_t,$	$\theta_c,$	$\theta_v,$	$\theta_p]^\top$	Emotion Opt. (r^*)	Standard (r_E)	Random (r_0)
Hungry–Thirsty	$\theta^* = [$	−0.4,	0.0,	0.0,	0.5,	0.1] [⊤]	9505.6 ± 7303.6	7783.7 ± 6930.1	35.6 ± 40.6
Lairs	$\theta^* = [$	0.1,	0.0,	−0.2,	0.0,	0.7] [⊤]	8635.8 ± 1133.3	7536.7 ± 944.8	173.3 ± 13.5
Moving-Prey	$\theta^* = [$	0.4,	0.0,	−0.1,	0.2,	−0.3] [⊤]	1986.9 ± 110.0	381.3 ± 17.2	683.1 ± 25.7
Persistence	$\theta^* = [$	−0.1,	0.1,	−0.1,	0.1,	0.6] [⊤]	1879.8 ± 11.2	136.3 ± 1.4	17.1 ± 0.7
Seasons	$\theta^* = [$	0.0,	0.1,	0.6,	0.0,	0.3] [⊤]	6142.3 ± 1336.3	4959.3 ± 1862.4	105.7 ± 24.4
Poisoned-Prey	$\theta^* = [$	0.1,	−0.2,	0.1,	0.0,	0.6] [⊤]	5237.6 ± 77.2	1284.3 ± 4.3	80.6 ± 24.9

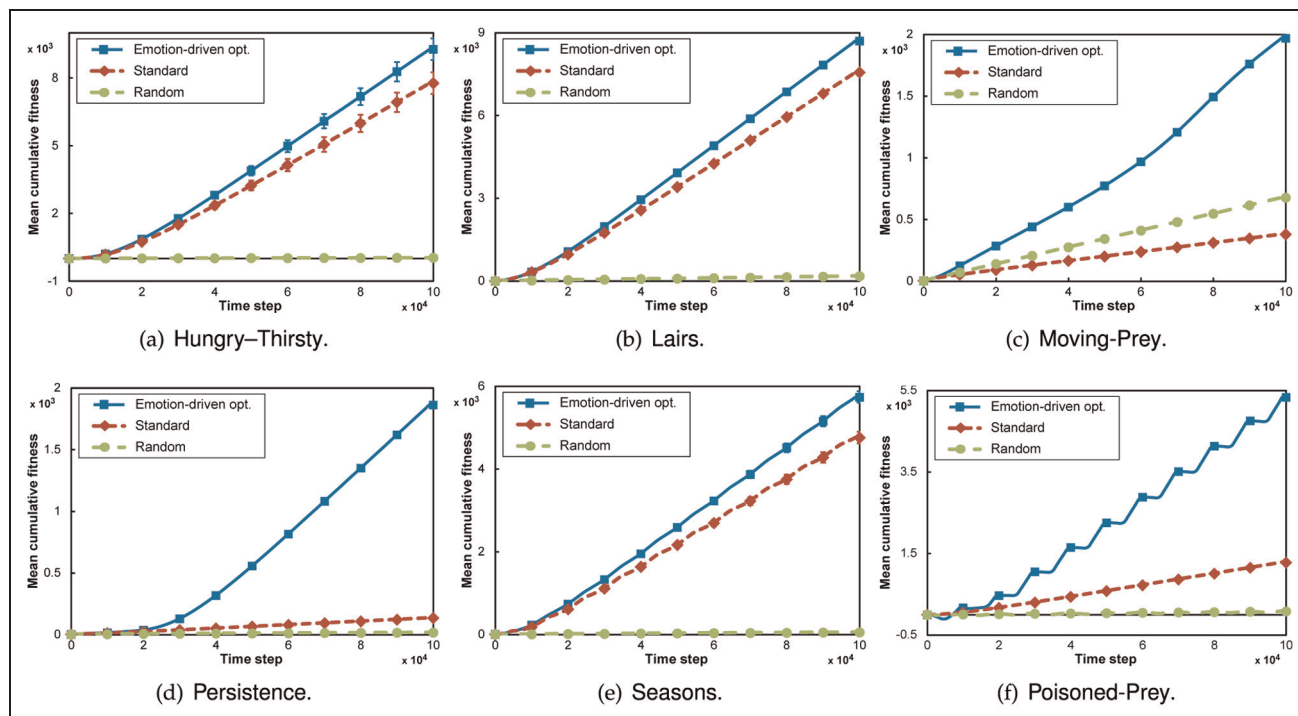


Figure 4. The evolution of the cumulative fitness in the foraging scenarios. We compare the performance of the optimal “emotion-driven” agent with that of a “standard” RL agent and an agent that acts “randomly”. The results correspond to averages calculated over 200 independent Monte Carlo trials, the standard error of which is calculated at 10,000 time-step intervals (see the text for more details).

the environment, the less controllable states correspond to the lair positions but *not immediately after a rabbit is consumed*. Therefore, the optimal parameterisation drives the agent to change location after capturing a rabbit. Such “nomad” behaviour is not, in itself, directly related to fitness maximisation. Instead, it is an intrinsic preference of the agent for certain situations, a preference for which the agent was conditioned by its environment and one that, in the long run, ends up enhancing its overall fitness—which is exactly what we intend with our approach.

The results for the Moving-Prey scenario are compared in Table 1 and in Figure 4(c). In this scenario, we can observe a much larger difference in the performance of the “emotion-driven” and “standard” RL agent, which is caused by the impact of partial observability: the RL agent keeps looking for prey at the position at which it last found one, but the dynamics of the environment ensure that no prey exists there. This aspect of the scenario makes the performance of the standard RL agent inferior to that of the random agent. By analysing the optimal parameter vector θ^* in Table 1, we observe that the extrinsic reward has a *negative weight* which, unlike for the standard RL agent, drives the agent *away* from the position at which it last found a prey. In contrast, there is a large positive weight assigned to *novelty*, which prompts the agent to pursue an exploratory policy.¹⁶

The Persistence scenario tests the potential of our agent to cope with short-term difficulties in pursuit of larger long-term (fitness-based) rewards. In particular, the fence at position (2 : 1) acts as an apparent obstacle that, if successfully overcome, leads to an improved performance in terms of fitness. The results for this scenario are compared in Table 1 and in Figure 4(d). The emotion-driven agent clearly outperforms the standard RL agent. Further analysis of the learned policies indicates that the standard RL agent is able to capture some “higher valued” hares in the initial stages of the simulation. However, as the fence becomes more difficult to cross and exploration decays, the standard RL agent finally settles on capturing the “lower-valued” rabbits because they provide a more accessible reward. This behaviour can also be perceived from the learning curves in Figure 4(d), for which the difference between the emotion-driven and the standard RL agent only becomes apparent after approximately 20,000 time-steps, when the maximum number of actions was already required to cross the fence.¹⁷ In contrast, the emotion-driven agent “stubbornly” prefers to pursue the hares rather than the rabbits. By analysing the optimal parameter vector θ^* in Table 1, we note that *goal relevance* and *valence* both have positive weights. *Goal relevance* drives the agent to approach high-valued (“goal”) states—in this case, the cell with the hares. In contrast, *valence* rewards states and actions with

Table 2. Optimal parameter vectors and corresponding mean cumulative fitness for the performance, in the Persistence scenario, of the different optimal emotion-based agents. We also include the results of the “standard” and “random” agents for comparison. The results correspond to averages calculated over 200 independent Monte Carlo trials.

Fit to Scenario	Optimal Parameter Vector						Mean Fitness
	$\theta^* = [$	$\theta_n,$	$\theta_t,$	$\theta_c,$	$\theta_v,$	$\theta_p]^\top$	
Hungry–Thirsty	$\theta^* = [$	−0.4,	0.0,	0.0,	0.5,	0.1] $^\top$	14.5 ± 5.8
Lairs	$\theta^* = [$	0.1,	0.0,	−0.2,	0.0,	0.7] $^\top$	67.3 ± 2.1
Moving–Prey	$\theta^* = [$	0.4,	0.0,	−0.1,	0.2,	−0.3] $^\top$	47.3 ± 4.2
Persistence	$\theta^* = [$	−0.1,	0.1,	−0.1,	0.1,	0.6] $^\top$	1879.8 ± 11.2
Seasons	$\theta^* = [$	0.0,	0.1,	0.6,	0.0,	0.3] $^\top$	11.5 ± 1.5
Poisoned–Prey	$\theta^* = [$	0.1,	−0.2,	0.1,	0.0,	0.6] $^\top$	60.5 ± 1.5
Standard (r_E)	$\theta_E = [$	0.0,	0.0,	0.0,	0.0,	1.0] $^\top$	136.3 ± 1.4
Random (r_0)	$\theta_0 = [$	0.0,	0.0,	0.0,	0.0,	0.0] $^\top$	17.1 ± 0.7

“above-average” values also pushing the agent towards the hare. Overall, a balanced combination of the different features provides the best policy and motivates the agent to cross the fence and attain higher rewards, even when the cross-the-fence behaviour, in itself, has no direct impact on fitness.

The results of the experiments for the Seasons scenario, which are compared in Table 1 and depicted in Figure 4(e), again exhibit a statistically significant difference in performance between the emotion-driven agent and the standard RL agent ($p < 10^{-4}$). Analysing the correspondent policies, we observe that both agents learn the same “safe policy”, i.e. eat the hares when available and ignore the rabbits. The observed difference is due to the impact of the *control* feature in the emotion-driven agent, which discourages the venture into less predictable states and enables the agent to settle for the hares *sooner* than the standard RL agent. This scenario provides an interesting example in which following a “safe” behaviour leads to a better adaptation to the environment, unlike some of the previous scenarios, in which exploratory behaviours led to an increased fitness.

The results of the Poisoned-Prey scenario are detailed in Table 1. In spite of the apparent difference in performance, both the emotion-driven and standard RL agents engage in fitness-enhancing policies, as is indicated by the positive slopes of both learning curves of Figure 4(f). In this scenario, the difference in attained fitness is due to the fact the standard RL agent, which relies on extrinsic reward only, preferred to capture rabbits throughout. From its perspective, this is a sensible behaviour because, on average, eating a hare results in a (fitness-based) reward of 0. The emotion-driven agent, in contrast, preferred to capture hares throughout time. Interestingly, as observed in Figure 4(f), it was able to survive the poisoned seasons. In fact, the fitness curve indicates that in the healthy season, the agent was able to consume a large number of hares. In the poisoned

season, in contrast, after suffering some initial loss (as indicated by the small peaks), the agent mostly stopped capturing hares.¹⁸ By analysing the corresponding parameter vector θ^* , we observe that the agent is driven by an interesting balance of positive *novelty* and *control*; whereas *novelty* fosters exploration (and hence the agent’s ability to track the seasons), *control* causes the agent to remain in states that it can easily predict (which results in the agent’s ability to effectively capture hares in the *healthy* season). To conclude, it is interesting to note that in this last scenario, the emotion-based features enabled the emergence of a relatively complex behaviour that allowed the agent to track the non-Markovian dynamics that result from the season changes. The results for this scenario thus support our hypothesis that emotion-based rewards enable learning agents to better adapt to their environment.

5.3 Maladaptation impacts fitness

We continue our results section and investigate what happens when an agent fit to a certain class of environments \mathcal{E}_1 is deployed in an environment e_2 that is significantly different from those in \mathcal{E}_1 . Much like what occurs in natural systems, we expect such “maladapted” agents to generally perform poorly in terms of fitness. To investigate this question, we deployed the optimal emotion-driven agents that were previously fit to each of the six foraging environments, in the Persistence scenario. As evidenced by the results in Table 2, there was a substantial difference in performance among the agents: the “maladapted” agents were unable to cope with the difficulties posed by this scenario and, as such, were unable to effectively capture prey. In fact, these agents were “conditioned” by the corresponding environments to address the emotional reward features in a specific way that allowed them to thrive therein. However, in the Persistence scenario, those reward

Table 3. Mean rank values for the “universal” emotion-based agent, which uses θ_U , the “standard” agent, which uses θ_E , and the “random” agent, which uses θ_0 calculated across all the foraging scenarios; see the text for details.

Parameter vector	$\theta = [$	$\theta_n,$	$\theta_t,$	$\theta_c,$	$\theta_v,$	$\theta_p]^\top$	Mean rank
Universal	$\theta_U = [$	0.0,	0.0,	-0.3,	0.0,	0.7] $^\top$	522.7 \pm 460.2
Standard	$\theta_E = [$	0.0,	0.0,	0.0,	0.0,	1.0] $^\top$	779.8 \pm 602.0
Random	$\theta_0 = [$	0.0,	0.0,	0.0,	0.0,	0.0] $^\top$	6243.0 \pm 2996.7

Table 4. Comparison of the mean cumulative fitness attained by the “universal” emotion-based agent, which uses θ_U , the “optimal” agent, which uses θ^* , and a standard RL agent, which uses θ_E , in each foraging scenario.

Scenario	Mean fitness		
	Universal (θ_U)	Optimal (θ^*)	Standard (θ_E)
Hungry-Thirsty	8297.8 \pm 5933.5	9505.6 \pm 7,303.6	7783.7 \pm 6930.1
Lairs	8798.0 \pm 1576.6	8635.8 \pm 1133.3	7536.7 \pm 944.8
Moving-Prey	460.8 \pm 49.2	1986.9 \pm 110.0	381.3 \pm 17.2
Persistence	470.8 \pm 59.0	1879.8 \pm 11.2	136.3 \pm 1.4
Seasons	4912.0 \pm 2606.3	6142.3 \pm 1336.3	4959.3 \pm 1862.4
Poisoned-Prey	1279.7 \pm 5.2	5237.6 \pm 77.2	1284.3 \pm 4.3

features actually distract the agent from the prey, and therefore these agents performed even worse than the standard RL agent (and some even performed worse than the random agent).

5.4 “Universal” agent

It is also important to assess the existence of a “universal” or “good enough” parameter configuration, i.e. one that is better *on average* than the fitness-based agent for all scenarios. For that purpose, we measured the average “rank” of each parameter vector across all foraging scenarios.¹⁹ The rank value for a specific scenario was calculated by sorting all the tested parameter vectors θ_k , $k = 1, \dots, K$ in descending order according to the respective mean cumulative fitness attained in that scenario. This ranking means that the optimal parameter vector θ^* for a scenario corresponds to the highest ranked vector in that scenario, which has a rank value of 0. We then averaged the rankings of all the tested parameter vectors across all scenarios and selected the one with highest mean value, which corresponds to the *universal parameter vector* and is denoted by θ_U .

Table 3 presents the configuration obtained for θ_U and a comparison between the resulting rankings for the universal, “standard” RL and “random” agents. The configuration of θ_U indicates that a combination of fitness-based rewards and negative *control* allowed the universal agent to attain a good overall performance. This result is explained by the non-stationarity of most of the foraging scenarios, which leads the agent to favour “less controllable” situations. However, as one would expect given the rank value of θ_U , when the performance of the universal parameter vector is calculated

individually in each scenario, the results are only marginal when compared with the corresponding optimal parameter vectors, as indicated in Table 4. Nonetheless, the performance of the universal agent was consistent with that of the standard agent, and in all but the Hungry-Thirsty, Seasons and Poisoned-Prey scenarios, it even performed significantly better ($p < 10^{-4}$).

The results of this experiment thus show the existence of a parameter vector that, despite not being “specialised” for any particular environment, is “good enough” across all scenarios, especially when compared with an agent that learns using only the external task reward.²⁰ In the context of our study, this result thus points towards the general purpose and usefulness of emotion-based rewards for solving complex learning tasks. We note, however, that such a universal configuration is still *dependent* on the particular set of foraging environments in which the learning occurred. It is therefore expected that in scenarios that have dynamics and challenges that are quite distinct from those presented by our foraging scenarios, the discovered universal agent would perform worse than the agent that is best adapted to such environments and possibly even the standard fitness-based agent.

6 Discussion and conclusions

This paper addresses the problem of reward design for IMRL agents in the context of the ORP. Departing from ATEs, we propose a set of four reward features that are inspired by major dimensions of appraisal. Such reward features, together with an additional *extrinsic reward* feature (that encodes an “environment-imposed” task), are used to construct an “environment-

adjusted”*intrinsic reward* that guides the decision-making process of each agent in that environment. Much like emotions in biological agents, our emotion-based features evaluate specific aspects of the history of interaction of the agent with its environment, thereby providing additional information that complements the agent’s perceptual information. For several scenarios, our results indicate that the proposed reward features enable the emergence of complex behaviours that allow our agents to significantly outperform standard RL agents in terms of their ability to cope with the multiple difficulties posed by each environment.

To conclude our paper, we consider additional links between our emotion-driven RL agents and biological agents in nature. One observation is concerned with the natural intrinsic motivation mechanisms that, from a physiological point of view, do not address any specific tissue deficit such as hunger or thirst (Ryan & Deci, 2000). Instead, theories of cognitive dissonance assert that organisms are motivated to reduce the incompatibility between perceived situations and cognitive structures built from past experience. Moreover, individuals find an equilibrium between the search for *novel stimuli* through exploration and the comfort of familiar situations, which provide a sense of *control* or *competence* over the external environment (Ryan & Deci, 2000). This equilibrium is also found in our emotion-driven agents. As can be observed from the results presented in Table 1, a well-balanced equilibrium between the different reward features (that measure, among other things, novelty and control) is fundamental for the agent’s ability to succeed in its environment.

A second interesting observation concerns the role of emotions in biological organisms, in which emotions play a major role in the processing of external events by involving primitive circuits within the limbic system that have been conserved throughout mammalian evolution (LeDoux, 2000). Emotions have thus provided animals with an ability to adapt their behaviours to survive longer and procreate more (Cardinal et al., 2002; Dawkins, 2000). Studies of the neural basis of emotions claim that these anticipatory mechanisms can be explained by simple associative learning processes that provide an ability to change behaviour in response to arbitrary stimuli and an ability to extend the range of stimuli that are perceived as hazardous or beneficial (Cardinal et al., 2002; Dawkins, 2000; LeDoux, 2007). These studies show that biological reinforcement processes rely on *emotional cues* to indicate the pleasantness or adversity of events and identify advantageous acting opportunities or harmful behaviours (Cardinal et al., 2002; Dawkins, 2000; Leventhal & Scherer, 1987). This emotion-based “evolutionary conditioning” of organisms finds a parallel in our process of parameter optimisation. Approximately speaking, this process “hardwires” in our agents associations between our reward features and the agent’s fitness, thereby endowing

the agent with the ability to learn complex behaviours that provide adaptive advantages in the environment.

It is also noteworthy that, as evidenced from Sections 5.3 and 5.4, we do not argue that emotion-based agents are universally superior to standard RL agents. In fact, much like biological agents, our emotion-based RL agents are often unable to perform satisfactorily when deployed in an environment to which they are not “adapted”. Moreover, the “universal” parameter vector that was discovered behaves well, on average, in the tested foraging scenarios. Had we used a different set of scenarios that demanded a completely different set of strategies to obtain fitness, perhaps the universal agent would have performed poorly under such conditions.

Another important observation is related to the level at which emotional appraisals occur. Commonly proposed ATEs focus on appraisals that rely on high-level cognitive concepts and mental representations (Ellsworth & Scherer, 2003; Lazarus, 2001; Leventhal & Scherer, 1987; Scherer, 2001). However, appraisal theorists also suggest that many appraisals, especially in the case of young children and non-human animals, require little cognitive processing or even simple judgments of the event (Frijda & Mesquita, 1998; Leventhal & Scherer, 1987; Scherer, 2001). Such multi-level ATEs explain emotions as an adaptive mechanism that develops from simple, reflex-like innate responses into more complex cognitive patterns (Leventhal & Scherer, 1987). Our emotion-based reward features rely on rather low-level statistical “summaries” of the agent’s history of interaction with the environment. In this multi-level perspective of appraisal, our emotion-based reward features in fact perform low-level evaluations that are similar to those made by different appraisal dimensions. In spite of their simplicity, however, they still allow for the individual and cross-cultural differences that are observed in human emotional experience (see Section 4.1): our features depend both on the *individual characteristics* of the agent, i.e. the particular parameter vector used to construct the intrinsic reward, and on *experience*, because the reward features are constructed from the agent’s particular history of interaction.

In the future, we plan to extend this research to multi-agent settings. In particular, we are interested in addressing the ORP in multi-agent settings within IMRL. For that purpose, we intend to design domain-independent reward features that assess the social acceptability of behaviours to achieve cooperation between learning agents in the context of resource-sharing scenarios.

In conclusion, we believe that the success of our approach stems from the fact that—much like the emotional processes in biological agents—our emotion-based rewards accommodate both the specificity of the agent (its learning algorithm and exploration policy, for

example) and its environment to complement the agent's perceptions. In this sense, the optimisation procedure that is required in the context of the ORP to determine the optimal reward resembles the environmental pressures that biological organisms have been subject to throughout evolution. And, in the case of our emotion-based rewards, just as evolution favours behaviours that seem to enhance the fitness of the agent in the long run, the optimisation of our biologically inspired reward mechanism enables our agents to behave, learn, and "live" much like biological organisms do.

Funding

This work was partially supported by the Portuguese para a e a Tecnologia (project number PEst-OE/EEI/LA0021/2013).

Notes

1. Interestingly, such results contrast with common RL approaches, such as UCB (Auer, Cesa-Binachi, & Fischer, 2002), E^3 (Kearns & Singh, 2002), and R-MAX (Brafman, 2003), that rely on the principle of "optimism in the face of uncertainty".
2. We note that, in this paper, we are not concerned with labelling the emotional state of the agent as "happy", "sad" or "angry". However, one can envisage a labelling mechanism that partitions the agent's "emotional space" into regions, with each corresponding to a specific emotional label.
3. When there is no danger of confusion, we abusively refer to a reward function r simply as a reward.
4. Agent-specific preferences may also accommodate the environment-imposed task. In particular, it is often the case that the intrinsic reward r depends on the extrinsic reward p .
5. ATEs are only one class of several classes of theories about the elicitation of emotions. In particular, ATEs contrast with other theories that do not consider such an evaluative and relational process, and with stimulus-response and other physiological and expressive theories, which focus on the subjective experience of emotions while ignoring the link between the situation and the individual (Ellsworth & Scherer, 2003; Frijda & Mesquita, 1998; Roseman & Smith, 2001).
6. We refer to the work of Sequeira, Melo, Prada, and Paiva (2011) for a treatment of the multi-agent case.
7. This estimate is constructed by the agent as part of its learning process, where we consider $\hat{\rho}^{(i)}(s, a) = \frac{1}{n_i(s, a)} \sum_{\tau=1}^i \rho(s_\tau, a_\tau)$.
8. States and actions that have not been visited for some time may not be novel because they have been visited often. On the other hand, recently visited states may still be novel because they have seldom been experienced.
9. Note that $s_p^{(i)}$ is unknown beforehand and depends on the time-step i . It is updated whenever the agent perceives a state in which the expected value is larger than that of all previously visited states. Note also that the distance estimate \hat{d} will often be inaccurate because it is built from the agent's estimated model of the

environment. However, we still expect it to convey useful information about "spatial relations between states".

10. Naturally, other features that relate the coping potential or power available may be suitable. We opted for this interpretation of control as related to prediction error because of the nature of our RL agents and the type of information to which they have access.
11. Denoting by n_t (fence) the number of times that the agent crossed the fence upwards up to time-step t , N_t is given by $N_t = \min\{n_t(\text{fence}) + 1; 30\}$. The fence is only an obstacle when the agent is moving upward from position (1: 2).
12. The initial season is randomly selected to be either Hare Season or Rabbit Season with equal probability.
13. Illustrative videos of the observed behaviours at different stages of the learning process in all scenarios are available as Online Supplementary Material.
14. The high standard deviation of the mean cumulative fitness observed in both agents is due to the different environment configurations, which lead to very different fitness values. For example, when the food and water are both located on the left, the agent must traverse the environment to move from the water to the food supply, which does not occur in other configurations (see Figure 3(a)).
15. We note, however, that the purpose of our experiments is mostly distinct from that of Singh et al. (2009).
16. The results for this scenario are also in agreement with the findings of Singh et al. (2010) for a similar setting.
17. This value was confirmed experimentally.
18. Interestingly, the slightly negative slope of the curve indicates that, every now and then, the agent returned and attempted to capture a hare again, thereby allowing it to effectively monitor the season changes.
19. We tested different methods of determining the "universal" parameter vector and determined this ranking procedure to be the best at selecting a configuration that, on average, enables the agent to perform well in all scenarios. In particular, we tested procedures that relied on the average fitness calculated across all scenarios, which yielded agents that performed very well in scenarios that provide a high degree of fitness, e.g. Hungry-Thirsty, but behaved poorly in lower maximal-fitness scenarios, e.g. Persistence (see the fitness in Table 1).
20. The difference between the rank value of the universal and "standard" agent reported in Table 3 is significant ($p = 8 \times 10^{-4}$).

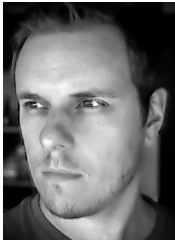
References

- Aberdeen, D. (2003). *A (revised) survey of approximate methods for solving partially observable Markov decision processes*. Eveleigh, Australia: National ICT Australia.
- Ahn, H., & Picard, R. (2006). Affective cognitive learning and decision making: The role of emotions. In R. Trappl (Ed.), *Proceedings of the 18th European meeting on cybernetics and systems research* (pp. 1–6). Vienna: Austrian Society for Cybernetic Studies.
- Armony, J., Servan-Schreiber, D., Cohen, J., & LeDoux, J. (1997). Computational modeling of emotion: Explorations

- through the anatomy and physiology of fear conditioning. *Trends in Cognitive Sciences*, 1, 28–34.
- Arnold, M. (1960). *Emotion and personality*. New York, NY: Columbia University Press.
- Auer, P., Cesa-Bianchi, N., & Fischer, P. (2002). Finite-time analysis of the multiarmed bandit problem. *Machine Learning*, 47, 235–256.
- Bobrow, G. (1964). *Natural language input for a computer problem solving system*. PhD Thesis, Massachusetts Institute of Technology, MA.
- Brafman, R. (2003). R-MAX: A general polynomial time algorithm for near-optimal reinforcement learning. *Journal of Machine Learning Research*, 3, 213–231.
- Bratman, J., Singh, S., Lewis, R., & Sorg, J. (2012). Strong mitigation: Nesting search for good policies within search for good reward. In *11th international joint conference on autonomous agents and multiagent systems* (pp. 407–414).
- Broekens, D., Kusters, W., & Verbeek, F. (2007). On affect and self-adaptation: Potential benefits of valence-controlled action-selection. In J. Mira and J. Álvarez (Eds.), *Bio-inspired modeling of cognitive tasks*. New York, NY: Springer.
- Cañamero, D. (1997). Modeling motivations and emotions as a basis for intelligent behavior. In *Proceedings of the 1st international conference on autonomous agents* (pp. 148–155).
- Cardinal, R., Parkinson, J., Hall, J., & Everitt, B. (2002). Emotion and motivation: The role of the amygdala, ventral striatum, and prefrontal cortex. *Neuroscience and Biobehavioral Reviews*, 26, 321–352.
- Cassandra, A. (1998). *Exact and approximate algorithms for partially observable Markov decision processes*. PhD Thesis, Brown University, RI.
- Damasio, A. (1994). *Descartes' error: Emotion, reason, and the human brain*. G. P. Putnam's Sons.
- Dawkins, M. (2000). Animal minds and animal emotions. *American Zoologist*, 40, 883–888.
- El-Nasr, M., Yen, J., & Ioerger, T. (2000). FLAME: Fuzzy logic adaptive model of emotions. *Autonomous Agents and Multiagent Systems*, 3, 219–257.
- Ellsworth, P., & Scherer, K. (2003). Appraisal processes in emotion. In R. Davidson, K. Scherer, & H. Goldsmith (Eds.), *Handbook of the affective sciences*. Oxford, UK: Oxford University Press.
- Ernst, D., Geurts, P., & Wehenkel, L. (2005). Tree-based batch mode reinforcement learning. *Journal of Machine Learning Research*, 6, 503–556.
- Frijda, N., & Mesquita, B. (1998). The analysis of emotions: Dimensions of variation. In M. Mascolo & S. Griffin (Eds.), *What develops in emotional development? Emotions, personality, and psychotherapy*. Plenum Press.
- Gadano, S. (2003). Learning behavior-selection by emotions and cognition in a multi-goal robot task. *Journal of Machine Learning Research*, 4, 385–412.
- Hogewoning, E., Broekens, D.J., Eggermont, J., & Bovenkamp, E. (2007). Strategies for affect-controlled action-selection in Soar-RL. *Nature Inspired Problem-Solving Methods in Knowledge Engineering*, 4528, 501–510.
- Jaakkola, T., Singh, S., & Jordan, M. (1995). Reinforcement learning algorithm for partially observable Markov decision problems. In *Advances in neural information systems 7* (pp. 345–352). Cambridge, MA: MIT Press.
- Jacobs, E., Broekens, J., & Jonker, C. (2014). Emergent dynamics of joy, distress, hope and fear in reinforcement learning agents. In *AAMAS workshop on adaptive learning agents*.
- Kaelbling, L., Littman, M., & Cassandra, A. (1998). Planning and acting in partially observable stochastic domains. *Artificial Intelligence*, 101, 99–134.
- Kaelbling, L., Littman, M., & Moore, A. (1996). Reinforcement learning: A survey. *Journal of Artificial Intelligence Research*, 4, 237–285.
- Kearns, M., & Singh, S. (2002). Near-optimal reinforcement learning in polynomial time. *Machine Learning*, 49, 209–232.
- Lazarus, R. (2001). Relational meaning and discrete emotions. In K. Scherer, A. Schorr, & T. Johnstone (Eds.), *Appraisal processes in emotion: Theory, methods, research* (pp. 37–67). Oxford, UK: Oxford University Press.
- LeDoux, J. (2000). Emotion circuits in the brain. *Annual Review of Neuroscience*, 23, 155–184.
- LeDoux, J. (2007). The amygdala. *Current Biology*, 17, 868–874.
- Leventhal, H., & Scherer, K. (1987). The relationship of emotion to cognition: A functional approach to a semantic controversy. *Cognition & Emotion*, 1, 3–28.
- Littman, M. (1994). Memoryless policies: Theoretical limitations and practical results. In *Proceedings of the 3rd international conference on simulation of adaptive behavior – from animals to animats* (pp. 238–245).
- Littman, M., Cassandra, A., & Kaelbling, L. (1995). Learning policies for partially observable environments: Scaling up. In *Proceedings of the 12th international conference on machine learning* (pp. 362–370).
- Lopes, M., Lang, T., Toussaint, M., & Oudeyer, P.-Y. (2012). Exploration in model-based reinforcement learning by empirically estimating learning progress. In *Advances in neural information systems* (pp. 206–214). Cambridge, MA: MIT Press.
- Madani, O., Hanks, S., & Condon, A. (1999). On the undecidability of probabilistic planning and infinite-horizon partially observable Markov decision problems. In *Proceedings of the 16th AAAI conference on artificial intelligence* (pp. 541–548).
- Marinier, R. P., Laird, J. E., & Lewis, R. L. (2009, March). A computational unification of cognitive behavior and emotion. *Cognitive Systems Research*, 10, 48–69.
- Marsella, S., Gratch, J., & Petta, P. (2010). Computational models of emotion. In K. Scherer, T. Banziger, & E. Roesch (Eds.), *Blueprint for affective computing* (pp. 21–44). Oxford, UK: Oxford University Press.
- Mataric, M. (1994). Reward functions for accelerated learning. In *Proceedings of the 11th international conference on machine learning* (pp. 157–164).
- McCallum, A. (1995). Instance-based utile distinctions for reinforcement learning with hidden state. In *Proceedings of the 12th international conference on machine learning* (pp. 387–395).
- Melo, F., Lopes, M., & Ferreira, R. (2010). Analysis of inverse reinforcement learning with perturbed demonstrations. In *Proceedings of the 19th European conference on artificial intelligence* (pp. 349–354).
- Minsky, M. (1986). *The society of mind*. Simon & Schuster.
- Moore, A., & Atkeson, C. (1993). Prioritized sweeping: Reinforcement learning with less data and less real time. *Machine Learning*, 13, 103–130.

- Moulin-Frier, C., & Oudeyer, P. (2013). Exploration strategies in developmental robotics: a unified probabilistic framework. In *Proceedings of the 3rd international joint conference on development and learning and epigenetic robotics* (pp. 1–6).
- Neu, G., & Szepesvári, C. (2009). Training parsers by inverse reinforcement learning. *Machine Learning*, 77, 303–337.
- Newell, A., & Simon, H. (1956). *The logic theory machine: A complex information processing system*. RAND Corporation.
- Ng, A., Harada, D., & Russel, S. (1999). Policy invariance under reward transformations: Theory and application to reward shaping. In *Proceedings of the 16th international conference on machine learning* (pp. 278–287).
- Ng, A., & Russell, S. (2000). Algorithms for inverse reinforcement learning. In *Proceedings of the 17th international conference on machine learning* (pp. 663–670).
- Niekum, S., Barto, A., & Spector, L. (2010). Genetic programming for reward function search. *IEEE Transactions on Autonomous Mental Development*, 2, 83–90.
- Picard, R. (2000). *Affective computing*. Cambridge, MA: MIT Press.
- Puterman, M. (1994). *Markov decision processes: Discrete stochastic dynamic programming*. John Wiley & Sons.
- Ramachandran, D., & Amir, E. (2007). Bayesian inverse reinforcement learning. In *Proceedings of the 20th international joint conference on artificial intelligence* (pp. 2586–2591).
- Randløv, J., & Alstrøm, P. (1998). Learning to drive a bicycle using reinforcement learning and shaping. In *Proceedings of the 15th international conference on machine learning*.
- Reisenzein, R. (2009). Emotions as metarepresentational states of mind: Naturalizing the belief–desire theory of emotion. *Cognitive Systems Research*, 10, 6–20.
- Roseman, I. (2001). A model of appraisal in the emotion system: Integrating theory, research, and applications. In K. Scherer, A. Schorr, & T. Johnstone (Eds.), *Appraisal processes in emotion: Theory, methods, research* (pp. 68–91). Oxford, UK: Oxford University Press.
- Roseman, I., & Smith, C. (2001). Appraisal theory: Overview, assumptions, varieties, controversies. In K. Scherer, A. Schorr, & T. Johnstone (Eds.), *Appraisal processes in emotion: Theory, methods, research* (pp. 3–19). Oxford, UK: Oxford University Press.
- Ryan, R., & Deci, E. (2000). Intrinsic and extrinsic motivations: Classic definitions and new directions. *Contemporary Educational Psychology*, 25, 54–67.
- Salichs, M., & Malfaz, M. (2006). Using emotions on autonomous agents: The role of happiness, sadness and fear. In *Proceedings of the annual convention on ambient intelligence and simulated behavior* (pp. 157–164).
- Scherer, K. (2001). Appraisal considered as a process of multilevel sequential checking. In K. Scherer, A. Schorr, & T. Johnstone (Eds.), *Appraisal processes in emotion: Theory, methods, research* (pp. 92–120). Oxford, UK: Oxford University Press.
- Sequeira, P., Melo, F., & Paiva, A. (2011). Emotion-based intrinsic motivation for reinforcement learning agents. In *Proceedings of the 4th international conference on affective computing and intelligent interaction* (pp. 326–336).
- Sequeira, P., Melo, F.S., & Paiva, A. (2014). Emergence of emotional appraisal signals in reinforcement learning agents. *Autonomous Agents and Multiagent Systems*.
- Sequeira, P., Melo, F., Prada, R., & Paiva, A. (2011). Emerging social awareness: Exploring intrinsic motivation in multiagent learning. In *Proceedings of the 1st international joint conference on development and learning and epigenetic robotics* (pp. 1–6).
- Simon, H. (1967). Motivational and emotional controls of cognition. *Psychological Review*, 74, 29–39.
- Singh, S., Jaakkola, T., & Jordan, M. (1994). Learning without state-estimation in partially observable Markovian decision processes. In *Proceedings of the 11th international conference on machine learning* (pp. 284–292).
- Singh, S., Lewis, R., & Barto, A. (2009). Where do rewards come from? In N. Taatgen & H. van Rijn (Eds.), *Proceedings of the annual conference of the Cognitive Science Society* (pp. 2601–2606).
- Singh, S., Lewis, R., Barto, A., & Sorg, J. (2010). Intrinsically motivated reinforcement learning: An evolutionary perspective. *IEEE Transactions on Autonomous Mental Development*, 2, 70–82.
- Smith, C., & Kirby, L. (2009). Putting appraisal in context: Toward a relational model of appraisal and emotion. *Cognition & Emotion*, 23, 1352–1372.
- Sorg, J., Singh, S., & Lewis, R. (2010a). Internal rewards mitigate agent boundedness. In *Proceedings of the 27th international conference on machine learning* (pp. 1007–1014).
- Sorg, J., Singh, S., & Lewis, R. (2010b). Reward design via online gradient ascent. In *Advances in neural information systems* 23 (pp. 1–9). Cambridge, MA: MIT Press.
- Sutton, R., & Barto, A. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT Press.
- Watkins, C. (1989). *Learning from delayed rewards*. PhD Thesis, Cambridge University, UK.
- Wiewiora, E. (2003). Potential-based shaping and *Q*-value initialization are equivalent. *Journal of Artificial Intelligence Research*, 19, 205–208.
- Winograd, T. (1971). Procedures as a representation for data in a computer program for understanding natural language. Cambridge, MA: MIT Press.

About the Authors



Pedro Sequeira is an associate researcher at the Intelligent Agents and Synthetic Characters Group (GAIPS) / Inesc-ID in Lisbon, Portugal. He has completed the PhD Program in information systems and computer engineering at the Instituto Superior Técnico of the University of Lisbon. His thesis focused on building more flexible and robust reward mechanisms for autonomous reinforcement learning (RL) agents. His interests are in the area of artificial intelligence, particularly related with autonomous and robotic agents and multi-agent systems involving RL with intrinsic motivation and also evolutionary/adaptive mechanisms.



Francisco S Melo is an assistant professor at Instituto Superior Técnico, University of Lisbon, and a researcher in the GAIPS group of INESC-ID. He received his PhD in electrical and computer engineering at Instituto Superior Técnico in 2007. Since then he has held appointments in the Computer Vision Lab of the Institute for Systems and Robotics, in Lisbon, and in the Computer Science Department of Carnegie Mellon University, in the USA. His research addresses problems within machine learning, particularly on reinforcement learning, planning under uncertainty, multi-agent and multi-robot systems, developmental robotics, and sensor networks.



Ana Paiva is a research group leader of GAIPS at INESC-ID and a full professor at Instituto Superior Técnico, University of Lisbon. She is also an adjunct scientist in the Artificial Intelligence Research Institute (IIIA) from the Spanish National Research Council (CSIC) in Barcelona. She is well known in the area of intelligent agents and multi-agent systems, artificial intelligence, human-robot interaction and affective computing. After her PhD in the UK, she returned to Portugal where she created a group on intelligent agents and synthetic characters (GAIPS). Her research is focused on the affective elements in the interactions between users and machines. She served as a member of numerous international conference and workshops. She has (co)authored over 100 publications in refereed journals, conferences and books.