Learning by Appraising:
An Emotion-based Approach to Intrinsic Reward Design

Pedro Sequeira Francisco S. Melo Ana Paiva
INESC-ID / Instituto Superior Técnico, Universidade de Lisboa
Av. Prof. Dr. Cavaco Silva
2744-016 Porto Salvo, Portugal
pedro.sequeira@gaips.inesc-id.pt,
{fmelo,ana.paiva}@inesc-id.pt

Abstract
In this paper, we investigate the use of emotional information in the learning process of autonomous agents. Inspired in four dimensions commonly postulated by appraisal theories of emotions, we construct a set of reward features to guide the learning process and behavior of a reinforcement learning (RL) agent inhabiting 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, in a sense replicating some aspects of appraisal processes observed in humans and other animals. Our experiments in several foraging scenarios show that, by optimizing the relative contributions of each reward feature, the resulting “emotional” RL agents attain superior performance than standard goal-oriented agents, particularly in face of their inherent perceptual limitations. Our results support our claim that biological evolutionary adaptive mechanisms such as emotions can provide a crucial clues in creating robust, general purpose reward mechanisms for autonomous artificial agents, allowing them to overcome some of the challenges imposed by their inherent limitations.

1 Introduction
From a computational perspective, reinforcement learning (RL) is concerned with providing efficient algorithms that enable artificial agents to acquire new tasks by trial-and-error (Kaelbling et al., 1996; Sutton and Barto, 1998). Through a process of repeated interaction with its environment, an RL agent experiments 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 in behaviorist psychology theories, RL algorithms stand as a natural choice when designing autonomous agents that must adapt their behavior to their environment (Sutton and Barto, 1998).

However, in deploying an RL agent, the agent designer is faced with a number of design challenges that will critically impact the performance of the agent. The first (and perhaps most fundamental) challenge is an agent
modeling challenge: RL agents are characterized 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, and observe the impact of this action both in terms of the resulting state and in terms of the received reinforcement signal, adjusting its action selection strategy accordingly (Sutton and Barto, 1998).

Unfortunately, it is often not possible for the agent designer to provide the agent with the ability to observe the whole state. In face of this limitation, the designer may decide either to consider a more complex model that explicitly accommodates for the perceptual limitations of the agent (Kaelbling et al., 1998), or to ignore these limitations and treat whichever information is available as its complete state (Jaakkola et al., 1995).

A second challenge that the designer faces is a task modeling challenge: given the adopted representation for the (state of the) agent, the designer must design the reinforcement signal that enables the agent to learn the intended task as efficiently as possible. The design of reward functions is a difficult endeavor, and has been the topic of intense research in the RL literature, leading to interesting results on both inverse reinforcement learning (Ng and Russell, 2000; Ramachandran and Amir, 2007) and reward shaping (Ng et al., 1999; Wiewiora, 2003).

Recent research on the origin of rewards in nature (Singh et al., 2009) and intrinsically motivated reinforcement learning (IMRL, Singh et al., 2010) has lead to the formulation of the optimal reward problem (ORP) to address the task modeling challenge discussed above. Roughly speaking, the ORP involves the discovery of a reward function, from a set of possible rewards, inducing the best “lifelong behavior” of the agent in a set of environments of interest, as measured in terms of a target task.

Interestingly, results have shown that, by selecting the reward that best “solves” the ORP, it is often the case that both the agent and task modeling challenges can be successfully addressed (Sorg et al., 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, in that process, mitigate the impact of the agent’s inherent limitations in its ability to successfully perform that task. Intuitively, the reward is used to provide the agent with implicit information on (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 modeling 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 endeavor will often involve significant domain knowledge, and several possibilities have been considered in the literature, requiring different levels of handcrafting (Bratman et al., 2012; Nickum et al., 2010; Singh et al., 2010; Sorg et al., 2010b).

In this paper we investigate the nature of 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 handcrafting across domains, and also informative enough to indeed provide useful information for each specific domain. We address the ORP within the framework of intrinsically motivated RL (IMRL), in which the process of reward optimization 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 et al., 2011a) we performed a preliminary study on 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 providing a reward built from a set of four domain-independent emotion-based features, namely novelty, valence, goal relevance and control, each inspired on a dimension of appraisal of the emotional significance of events, commonly found in the psychology literature (Ellsworth and Scherer, 2003; Lazarus, 2001; Leventhal and Scherer, 1987; Roseman and Smith, 2001; Scherer, 2001). We perform such mapping regardless of its validity in terms of appraisal theories, but redesigned many of the previously proposed features in order to focus on emotions as a plausible source of such general-purpose and domain-independent intrinsic reward, discussing possible alternatives for each feature.

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

2 Related Work

Early research within artificial intelligence (AI) was mostly focused on the reproduction of human reasoning processes, e.g., by building systems that could prove theorems (Newell and Simon, 1956), solve algebra word problems (Bobrow, 1964) or understand English sentences (Winograd, 1971). Pioneer AI researchers also emphasized the role of emotional processes as an attention-focusing, task prioritizing mechanism, crucial to any system 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 in the AI community the development of computational models of emotions, usually based on appraisal theories of emotions (Marsella et al., 2010). Many works in the area of affective computing address the impact of emotional processes in decision-making, and contribute to more engaging interactive artificial agents (Picard, 2000).

Within the area of affective computing (AC), several works combine learning and emotional mechanisms in a complementary manner, in order to create artificial agents that exhibit richer behaviors. For example, the FLAME model uses RL to build emotion-object associations and to predict the user’s actions (El-Nasr et al., 2000). Another example is the work of Armony et al. (1997), where connectionist learning is used to emulate effects commonly associated with fear-conditioning. The artificial creatures of Cañamero (1997) also use “low-level” emotional signals that drive behavior selection. Jacobs et al. (2014) derive emotions of “joy”, “distress”, “hope” and “fear” from signals generated by the RL algorithm and show that the results of agent-based simulations are able to replicate psychological and behavioral dynamics of emotion.

Another line of work, closer to our own, makes use of emotions to actually influence decision-making within RL. Gadano (2003) propose a bottom-up approach to emotion elicitation. The system uses artificial neural
networks to determine a dominant emotional state. A measure of “well-being” (or valence) is calculated for each state as the relative change in the value of a set of homeostatic variables (energy, welfare and activity), also accounting for predictions associated with that state. RL is used to learn state-behavior associations, where the rewards are provided by the intensity and valence of the current dominant emotion. Salichs and Malfaz (2006) propose a set of basic emotions to control the behavior of an RL agent, where the reward depends on variations of the agent’s well-being. The behavior 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 et al. (2009) proposed an intrinsic reward signal based on the appraisal of conduciveness which determines the sign of the reward value, while the intensity of the agent’s current feeling determines the magnitude of the signal. An experiment conducted in a grid-world scenario showed that intermediate, emotion-based rewards lead to learning the task faster. Broekens et al. (2007) propose associating positive affective states with exploitation and negative states with exploration and show that this provides adaptive benefits for RL agents in specific scenarios.\footnote{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 et al. (2007) use a Chi-square statistical test to compute the significance of the differences between these two averages to influence action-selection.}

The work that is closest to our proposed approach is that of Ahn and Picard (2006). In that work, the authors consider 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 affective anticipatory reward 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 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 is in contrast with most surveyed works, which rely either on a predefined set of discrete emotions or in scalar evaluations of the emotional state of the agent.\footnote{Additionally, unlike Ahn and Picard (2006), we do not handle differently the external and internal rewards. Instead, they are harmoniously combined to yield a single reward signal that guides the agent’s behavior. The trade-off between such external and internal rewards is “fit” for the class of environments that the agent expects to encounter. A related optimization can be found in biological agents, which process emotional states differently depending on their survival needs (Frijda and Mesquita, 1998; Roseman and Smith, 2001; Smith and 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 optimizes the balanced benefit arising from the environment and the agent’s internal state.}

We also make reference to our work in (Sequeira et al., 2014), corresponding to the complement of this paper,\footnote{Interestingly, such results contrast with common RL approaches that rely on the principle of “optimism in the face of uncertainty” such as UCB (Auer et al., 2002), E\textsuperscript{3} (Kearns and Singh, 2002), or R-MAX (Brafman, 2003).\footnote{We note that, in this paper, we are not concerned with labeling the emotional state of the agent as being “happy”, “sad” or “angry”. However, one can envisage a labeling mechanism that partitions the agent’s “emotional space” into regions, each corresponding to a specific emotional label.}}
where we test the emergence of emotion-related rewards by means of evolutionary computation mechanisms. In other words, in this paper we depart from the emotional appraisal literature to manually design reward features having evaluative characteristics similar to those ascribed by some appraisal dimensions—in (Sequeira et al., 2014) we emerge by means of genetic programming a set of domain-independent reward features and then analyze their dynamical and structural properties in light 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, setting up 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 modeled as a partially-observable Markov decision problem (POMDP) (Kaelbling et al., 1998), denoted as a tuple \( \mathcal{M} = (S, A, Z, P, O, r, \gamma) \).

At each discrete time step \( t = 0, 1, 2, 3, \ldots \), the environment is described by some state, represented as a random variable (r.v.) \( S_t \) taking values in a finite set of possible states, \( S \). The agent makes an observation, denoted as a r.v. \( Z_t \) taking values in a set of possible observations, \( Z \), that depends on the state \( S_t \) but that is often insufficient for the agent to unambiguously infer \( S_t \). The agent then performs some action (denoted as a r.v. \( A_t \in A \)) and the environment transitions to state \( S_{t+1} \). This transition is governed by the probabilities \( \mathbb{P}[S_{t+1} = s' \mid S_t = s, A_t = a] = P(s' \mid s, a) \). The agent then receives a numerical reward, \( r(S_t, A_t) \), representing the desirability of having executed action \( A_t \) in state \( S_t \) (in terms of the target task), and makes a new observation \( Z_{t+1} \) and the process repeats. The observations \( Z_t \) of the agent are governed by the probabilities \( \mathbb{P}[Z_{t+1} = z \mid S_{t+1} = s, A_t = a] = O(z \mid s, a) \). Traditional approaches to RL mainly focus on scenarios where the observations \( Z_t \) do allow the agent to unambiguously determine the underlying state \( S_t \). Such scenarios are said to have full observability and the POMDP parameters \( Z \) and \( O \) can be safely ignored. The resulting model, represented as a tuple \( \mathcal{M} = (S, A, P, r, \gamma) \), 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 behavior with respect to the task it must (learn to) perform, acting as a critic residing in the (external) environment, as depicted in Fig. 1(a) (Singh et al., 2010). The goal of the agent is to select its actions so as to gather as much reward as possible during its lifetime, discounted by some factor (Kaelbling et al., 1996; Sutton and Barto, 1998).

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

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

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

\(^3\)When there is no danger of confusion, we abusively refer to a reward function \( r \) simply as a reward.
Figure 1: Comparison between the RL and IMRL frameworks: (a) the traditional RL model in which a critic in the external environment evaluates the behavior 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 behavior of the agent providing intrinsic rewards (adapted from the work of Singh et al., 2010).

verifies the recursive relation:

\[ Q^*(s, a) = r(s, a) + \gamma \sum_{s' \in S} P(s' \mid s, a) \max_{a' \in A} Q^*(s', a'). \]  

(2)

\( Q^* \) 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 (2), in a dynamic programming approach known as value iteration. The 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 relying on data collected from the environment, either online (Watkins, 1989) or offline (Ernst et al., 2005).

In this paper, we consider RL agents that follow the prioritized sweeping algorithm (Moore and Atkeson, 1993). Prioritized sweeping 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 (implemented using a priority queue).

3.2 Intrinsically Motivated RL and the ORP

The RL paradigm described in Section 3.1 departs from a (PO)MDP model, describing a sequential problem faced by a decision-maker in a dynamic and uncertain world and whose task is implicitly encoded in the reward \( r \). The performance of the learning agent will depend on the ability of \( r \) to convey information on the task to be learned, and several works in the literature have addressed the problem of reward design. One common approach relies on the idea of shaping (Mataric, 1994; Ng et al., 1999; Randlov and Alstrom, 1998): given a reward \( r \) encoding some target task, shaping consists in applying some transformation to \( r \), 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 behavior, and feed this reward to the agent (Ng and Russell, 2000). This approach is known as inverse reinforcement learning and has spanned a significant volume of literature (Melo et al., 2010; Neu and Szepesvári, 2009; Ng and Russell, 2000; Ramachandran and Amir, 2007).

One radically different perspective on the problem of reward design arises from recent work on intrinsically motivated reinforcement learning (IMRL, Singh et al., 2010). IMRL seeks to model, within the RL framework, behaviors observed in nature that are not (directly) oriented towards “survival”—such as curiosity-driven behaviors (Singh et al., 2009, 2010). Within IMRL, the rewards arise from an evaluation of an “internal critic” of information both from the external environment and the agent’s “internal environment”, as depicted in Fig. 1(b).
IMRL further proposes a distinction between extrinsic reward (henceforth denoted as $\rho$), which evaluates the behavior of the agent with respect to some environment-imposed task (e.g., survival), and intrinsic reward (henceforth denoted as $r$), which evaluates the behavior 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 “evolutionary-shaped” mechanisms for optimally coping with the environments it expects to encounter.\footnote{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 $\rho$.} Computationally, IMRL was distilled into the optimal reward problem (Sorg et al., 2010a).

**Definition 1** (Optimal Reward Problem (ORP)). Given a learning agent $U$, a set of possible environments $E$ that agent $U$ may inhabit, and a target task $T$ to be learned, which reward $r$, among a set $R$ 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 behavior of the agent with respect to the target task $T$, and the goal of the RL agent itself, which concerns its behavior 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 behavior 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 $h_{1:t} = \{z_1, a_1, \rho_1, z_2, \ldots, a_{t-1}, \rho_{t-1}, z_t\}$, where $z_\tau$, $a_\tau$ and $\rho_\tau$ denote, respectively, the observation at time-step $\tau$, the action selected at time-step $\tau$, and the extrinsic reward at time-step $\tau$. Referring back to Fig. 1(b), the internal critic is responsible for processing the agent’s perceptions into a history $h$, that contains environment information (in the form of a sequence $\{z_1, \ldots, z_t\}$) and extrinsic reward information (in the form of a sequence $\{\rho_1, \ldots, \rho_{t-1}\}$).

Additionally, let $r$ denote the intrinsic reward driving the behavior of the agent (modeled 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 $\mathbb{P}[h \mid r, e]$ to denote the probability of observing history $h \in \mathcal{H}$ given $r$ and $e$.

We define the fitness function, $f : \mathcal{H} \to \mathbb{R}$, 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, $R$, and a distribution $p_{\text{env}}$ over the set of environments, $E$, the ORP can thus be formulated as the problem of determining the optimal reward $r^* \in R$ such that

$$r^* = \arg\max_{r \in R} \mathcal{F}(r) \triangleq \mathbb{E}_{e \sim p_{\text{env}}} [f(h) \mid r, e],$$

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.,

$$f(h_{1:t}) = \sum_{\tau=1}^{t} \rho_{\tau}.$$
This particular choice of fitness function implicitly indicates that $\rho$ directly measures the fitness of the agent and, therefore, we interchangeably refer to $\rho$ as the extrinsic reward and the fitness-based reward.

4 Designing Emotion-based Rewards

This section introduces our main technical contribution, consisting of a set of reward features inspired in 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, looking into well-established appraisal theories of emotions (Arnold, 1960; Ellsworth and Scherer, 2003; Roseman and Smith, 2001).

Appraisal theories of emotions (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 well-being and goals (Arnold, 1960). ATEs investigate the functional aspect of emotions, and seek to explain the effect of appraisals in decision-making and, more generally, behavioral and cognitive responses to the individual’s perceived situation. These responses contribute to focus the individual’s attention to significant aspects of its environment (Frijda and Mesquita, 1998; Lazarus, 2001; Leventhal and Scherer, 1987).

Figure 2 provides a high-level illustration of the process of emotional elicitation according to ATEs. This process 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 responsible for the subjective feelings of emotions. Appraisal can take place at different levels (Ellsworth and Scherer, 2003; Leventhal and Scherer, 1987)—e.g., the basic fight-or-flight evaluation (observed in humans and other animals) when facing a dangerous situation is different from the more cognitive assessment taking place after the death of a close family member. According to the level at which appraisal takes place, the kind of information processed by the subject will be of a different kind—more simple at a low level, more complex at higher cognitive levels.

Many ATEs in the literature propose structural models in which emotions are elicited by evaluations of events.

---

5 Appraisal is but one of several other theories on 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, focusing on the subjective experience of emotions while ignoring the link between the situation and the individual (Ellsworth and Scherer, 2003; Frijda and Mesquita, 1998; Roseman and Smith, 2001).
through a set of appraisal variables (Frijda and Mesquita, 1998; Lazarus, 2001; Roseman, 2001; Scherer, 2001). Each variable is conceptualized as a dimension along which appraisal outcomes may vary continuously (Roseman and Smith, 2001). The several dimensions define the criteria used to evaluate a situation and ascribe the structure or the contents of the appraisal (Ellsworth and Scherer, 2003; Roseman and Smith, 2001).

Most of the appraisal dimensions proposed in the literature deal with 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 (e.g., joy, sadness, fear, etc.) and predict the particular physiological responses and action tendencies associated with each of them (Ellsworth and Scherer, 2003; Frijda and Mesquita, 1998; Roseman and Smith, 2001). Therefore, most ATEs largely agree in which dimensions are necessary to evaluate of a given situation. Ellsworth and Scherer (2003) have compared the most common ATEs and identify a set of five major dimensions of appraisal, mostly consensual in the community, and in which our approach is based upon: 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 before 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 undecidable in the worst case (Madani et al., 1999).

Given the difficulty inherent to 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 et al., 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 shown in the work of Singh et al. (1994). Moreover, computing the best such solution is, per se, a difficult problem (Littman, 1994).

Other approaches have been proposed to deal with partial observability in RL settings that build into the agent prior knowledge that can, somehow, alleviate its perceptual limitations (Aberdeen, 2003). Examples include approaches based on some form of memory (McCallum, 1995). However, such approaches typically require very specific algorithms 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 behaviors that may not be directly related to the target task but which, in time, can mitigate the impact of the agent’s limitations in 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 brings forth a new design challenge—that of designing the set $\mathcal{R}$ of possible rewards for a desired task. As discussed before, this 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).

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

### 4.3 Emotion-based Reward Design

We are now in position to introduce our main technical contributions. In particular, we depart from the discussion on ATEs in Section 4.1 and propose a set of reward features inspired by each of the major dimensions of appraisal.

Going back to the IMRL agent architecture in Fig. 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 Fig. 2, we can roughly identify the internal critic in our IMRL agent as the module where appraisal will take place. The reward $r$ used for learning and decision-making roughly corresponds to the outcome of such process.

We treat the agent’s perceptions as state, a common simplifying approach already discussed in Section 4.2. This translates into considering, in the POMDP model, that $\mathcal{S} = \mathcal{Z}$ and $\mathcal{P} [S_t = s \mid Z_t = s] = \mathcal{P} [Z_t = s \mid S_t = s] = 1$. 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 before, it will seldom be 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 $\mathcal{R}$, where each reward $r$ is a linear combination of some pre-defined reward features, $\{\phi_i, i = 1, \ldots, N\}$. Each feature $\phi_i$ maps perception-action-history triplets (abusively denoted as $(s, a, h)$, given our treatment of perceptions as state) to a scalar value $\phi_i(s, a, h) \in \mathbb{R}$. Hence, for every $r \in \mathcal{R}$,

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

where $\theta_i$ is the linear coefficient associated with feature $\phi_i$ in $r$.

We propose that each appraisal dimension maps to a corresponding reward feature $\phi_i, i = 1, \ldots, N$. Much like appraisal dimensions in biological agents, our reward features evaluate the significance of the agent’s current situation for its “well-being”, 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 and Scherer, 2003; Scherer, 2001). Given the simplicity of the RL agent model considered, our emotion-based reward features rely on low-level statistical “summaries” of the agent’s history of interaction with the environment.

Since 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 and Scherer, 2003).6

---

6We refer to the work of Sequeira et al. (2011b) for a treatment of the multiagent case.
constructed from information 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 $\{\phi_n, \phi_c, \phi_e, \phi_p\}$, where

- $\phi_n(s, a, h)$ denotes the novelty associated with performing action $a$ after observing $s$, given the history $h$;
- $\phi_c(s, a, h)$ denotes the goal relevance of performing action $a$ after observing $s$, given $h$;
- $\phi_e(s, a, h)$ denotes the degree of control over the outcome of executing action $a$ after observing $s$, given $h$;
- $\phi_p(s, a, h)$ denotes the expected valence of executing $a$ after observing $s$, given history $h$;
- Finally, $\phi_p(s, a, h) = \hat{r}(s, a)$ is not an emotion-based feature, and corresponds to the estimated fitness-based reward for executing $a$ after observing $s$.\(^7\) $\hat{r}(s, a) = E \left[p_t \mid S_t = s, A_t = a \right]$. 

In the continuation, we describe each of the above features in detail.

**Novelty** is one of the most basic and low-level dimensions of emotional appraisal of events, usually eliciting focus of attention to important changes occurring in the environment (Ellsworth and Scherer, 2003; Frijda and Mesquita, 1998; Reisenzein, 2009). There are several factors contributing to the evaluation of an event’s novelty, e.g., the level of habituation to a stimulus, the individual’s motivation state or the perception of predictability or expectedness of a situation (Ellsworth and Scherer, 2003; Roseman, 2001). At perception or schematic level, novelty usually refers to the degree of familiarity or matching between the perceived stimuli and the agent’s knowledge structures built so far (Frijda and Mesquita, 1998; Leventhal and Scherer, 1987; Reisenzein, 2009).

In the RL framework, familiarity about 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 action $a$ was selected after perceiving $s$. One can thus quantify the dimension of novelty as

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

where $\lambda_n$ is a positive constant such that $\lambda_n < 1$. The two terms in (5) account, respectively, for the novelty in terms of experienced actions and the novelty in terms of perceived states. $\lambda_n$ can be seen as a “novelty rate” determining how novelty decays with experience.

The expression proposed for novelty is related with the inverse-frequency feature of Bratman et al. (2012). However, instead of a linear decaying rate, we consider an exponential decaying rate that is dependent, for example, on the total number of states and actions that can be experienced or the agent’s lifetime. Also, we adopted a frequency-based feature in favor of recency-based feature, as it captures better the essence of the novelty dimension.\(^8\) Our calculation of the novelty feature includes only information that evaluates the amount of past experience 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, all in line with the corresponding novelty dimension in biological agents (Ellsworth and Scherer, 2003; Leventhal and Scherer, 1987).

\(^7\)This estimate is constructed by the agent as part of its learning process where we consider $\hat{r}(s, a) = \frac{1}{n_t(s, a)} \sum_{t=1}^{t} r(s_t, a_t)$.

\(^8\)States and actions that have not been visited for some time may not be novel, as they have been visited often. On the other hand, recently visited ones may still be novel, as they have seldom been experimented.
Goal Relevance assesses the relevance of a perceived event in terms of the attainment of the agent’s long-term goals or the satisfaction of its needs (Ellsworth and Scherer, 2003; Lazarus, 2001; Leventhal and Scherer, 1987). Also related to the notions of desire-congruence (Reisenzein, 2009) or motive-consistence (Roseman, 2001), goal relevance is essential for the survival and adaptation of an individual to its environment (Ellsworth and Scherer, 2003; Reisenzein, 2009). 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 being considered (Roseman, 2001). Broadly speaking, we can say that the goal relevance of an event increases if such event is consistent or conductive to the achievement of the individual’s goals, and decreases when the consequences of the event are obstructive to reaching those goals (Ellsworth and 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, $\phi_{\rho}$ verifying the fixed point relation $V_{\rho}(t) = \max_{a \in A} Q_{\rho}(t)(s, a)$, 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 $s_{\rho}(t) = \arg\max_{s \in S} V_{\rho}(t)(s)$, 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 translated as

$$\phi_{\rho}(s, a, h_{1:t}) = \frac{1}{1 + \hat{d}_t(s)}.$$  \hspace{1cm} (6)

This expression is coherent 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 such goal. As a measure of distance, we used the estimated number of steps required to reach the goal, which is a generalization 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 that naturally induced by the transition probabilities $P$.

Control usually involves a proactive assessment of the ability of the individual to deal with a particular situation (Ellsworth and Scherer, 2003; Frijda and Mesquita, 1998; Lazarus, 2001; Leventhal and Scherer, 1987). It is considered to be part of a “secondary” level of appraisal, as it requires the individual to evaluate its ability to produce an appropriate response to the event being evaluated (Lazarus, 2001). Such coping potential usually refers to the power the agent has to assess the probability of possible outcomes and change the situation and its consequences Ellsworth and Scherer (2003). At a higher level (i.e., more cognitive) of appraisal, these evaluations often require adjusting either the significance of the situation at hand (Ellsworth and Scherer, 2003; Roseman, 2001), or the individual’s goals, in order to cope with the possible outcomes of the event (Lazarus, 2001; Smith and 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.

\(^9\)Note that $s_{\rho}(t)$ is not known beforehand, and depends on the time-step $t$. It is updated whenever the agent perceives a state in which the expected value is larger in comparison with all previously visited states. Note also that the distance estimate $\hat{d}$ will often be inaccurate, since it is build from the agent’s estimated model of the environment. However, we still expect it to convey useful information about “spatial relations between states”.
We adopt the perspective that control over a situation is often directly related to the degree of predictability of the outcomes being considered (Ellsworth and Scherer, 2003; Leventhal and Scherer, 1987; Roseman, 2001). The ability of an RL agent to control its environment is directly related with the accuracy of its world-model. Accurate world models allow the agent to reason correctly about which actions maximize its reward/fitness, while inaccurate world models may cause the agent to often select suboptimal actions.\footnote{Naturally, other features relating the coping potential or power available may be suitable. We opted for this interpretation of control as related to prediction error due to the nature of our RL agents and the kind of information they have access to.}

To measure the accuracy of the agent’s world model, we determine how accurately $Q^{(t)}_\rho$ verifies the relation (2) given estimates $\hat{\rho}$ of the fitness-based reward. Specifically, we measure how the most recent information perceived by the agent impacts its current estimate, defining the prediction error associated with $Q^{(t)}_\rho(s,a)$ whenever $s,a$ are experienced at time-step $t$ as $\Delta Q^{(t)}_\rho(s,a) = k \cdot |Q^{(t)}_\rho(s,a) - Q^{(t-1)}_\rho(s,a)|$, where $k$ is a normalizing constant and $Q^{(t-1)}_\rho(s,a)$ corresponds to the previous value computed for $Q_\rho(s,a)$, i.e., $t = 1$ corresponds to the previous time step in which action $a$ was executed given state $s$. We then define the control feature $\phi_c(s,a,h_{1:t})$ according to the negative running average prediction error associated with $Q^{(t)}_\rho(s,a)$, i.e.,

$$\phi_c(s,a,h_{1:t}) = 1 - \frac{1}{n_t(s,a)} \sum_{\tau=1}^{t_{s,a}} \Delta Q^{(\tau)}_\rho(s_\tau,a_\tau),$$

(7)

where $t_{s,a}$ corresponds to those time steps in which the state-action pair $s,a$ was experienced. From the above expression, we note that $\phi_c(s,a,h_{1:t})$ will be close to 0 for those state-action pairs that, throughout the agent’s history, are hardest to “learn”. Conversely, $\phi_c(s,a,h_{1:t})$ will be close to 1 in those state-action pairs that the agent quickly learns. Note also that, as the agent’s knowledge of the environment improves, so does the value of $\phi_c(s,a,h_{1:t})$. The feature thus provides a meaningful measure of the agent’s predictive ability.

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

Valence measures how intrinsically pleasant a given situation is (Ellsworth and Scherer, 2003). It is considered a low-level, automatic appraisal dimension, generated from both innate detectors and learned preferences, and indicates whether a stimulus is “positive” or “negative” in terms of biological significance to the individual (Leventhal and Scherer, 1987). Unlike other dimensions, valence is considered a feature of the stimulus itself, mostly independent of the momentary situation of the individual (Ellsworth and Scherer, 2003).

At such low-level, in our IMRL framework valence is perhaps best represented as the fitness-based reward itself, $\phi_\rho$, as it provides an immediate direct evaluation of the perceived states and executed actions in terms of fitness.
However, as seen in Section 3.2, $\phi_\rho$ 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 et al., 2002; Ellsworth and Scherer, 2003; Leventhal and Scherer, 1987). Bearing this idea in mind, and in order 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

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

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

As can be seen from the expression above, $\phi_v(s, a, h)$ is maximized when the agent executes the best action in the most valuable state, denoting a learned preference towards a behavior 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), where the agent feels “good” or “bad” depending on how the expected reward received after executing an action in the current state compares with previous rewards previously experienced in that state. However, such expression evaluates only immediate reward, mostly ignoring the agent’s long-term goals. Another alternative can be found in (Broekens et al., 2007), where 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 taken according to the past actions executed while our proposal for valence reflects possible changes of preference towards stimuli as given by future courses of action. We also refer to the “well-being” (or valence) feature proposed in (Gadanho, 2003), calculated for each state as its relative impact and prediction value in a set of homeostatic variables. Positive/negative changes and predictions lead to positive/negative states of well-being.

5 Experiments and Results

To evaluate our emotion-based intrinsic rewards, we run a set of experiments in foraging environments inspired by those of Singh et al. (2010), where we model our agent as a predator trying to feed on preys throughout time. The choice of these scenarios is tightly connected with the objectives of the experiments. First of all, foraging scenarios enable a clear evaluation of the agent’s behavior in terms of the target task. In particular, the extrinsic reward $\rho$ will reinforce feeding behaviors, closely related with fitness (as is the case in biological agents).

Secondly, foraging scenarios also facilitate an evaluation of the impact of the different reward features in the behavior of the agent. As will soon become apparent, different environments will 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. On the other hand, if poorly combined, they will lead to poor performance, 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 brings advantages when designing artificial learning agents. As our results illustrate,
Figure 3: Possible environment configuration for the several foraging scenarios used in the experiments. In all environment we represent our agent and its corresponding start-position by the “dark fox” figure. We refer to a cell in column \( x \), row \( y \) as a position \((x:y)\). See text for specific descriptions of the dynamics of each scenario.

the partial observability of the state, 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 behaviors not directly related to fitness enhancement but which will often (indirectly) lead to a more successful “feeding policy”.

5.1 Experimental setup

We used a total of 6 scenarios (see Fig. 3), either from the IMRL literature or modifications thereof (Singh et al., 2010; Sorg et al., 2010a), that we describe in the continuation.

Hungry-Thirsty scenario: This scenario is adapted from the work of Singh et al. (2010), and is depicted in Fig. 3(a). It contains two inexhaustible resources, corresponding to food and water. Resources can be positioned in any of the environment corners (positions \((1:1)\), \((5:1)\), \((1:5)\), and \((5:5)\)), leading to a total of 12 possible configurations of food and water (only one of which is depicted in Fig. 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 achievable 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 Fig. 3(b). There are two lairs positioned in different corners of the environment, resulting in 6 possible configurations. The fitness of the agent is defined as the number of preys 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 0.1 probability that a prey will appear in an empty lair. The agent is able to observe its position and the state of both lairs (either occupied, empty, or prey outside).

Moving-Preys scenario: This scenario is also adapted from the work of Singh et al. (2010), and one possible configuration is depicted in Fig. 3(c). In this scenario, at any time-step, there is exactly one prey available, located in 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 preys captured. Whenever the agent captures a prey, the latter disappears from the current location and a new prey randomly appears in one of the two other possible prey locations.
**Persistence scenario:** The environment used in this scenario is depicted in Fig. 3(d). In this scenario, the environment contains two types of preys always available. *Hares* are located in position (3 : 1) and contribute to the fitness of the agent with a value of 1 when captured, while *Rabbits* are located in (3 : 5), contributing with a value of 0.01. Whenever it captures a prey, the agent’s position is reset to the initial position at (3 : 3). The environment also contains a *fence*, located in (1 : 2), that prevents the agent from easily capturing hares. To cross over the fence towards the hare location at time \( t \), the agent must perform action \( N \) for \( N \) consecutive time-steps, after which the fence is reinforced, requiring an increasing number of actions \( N \) to be crossed.\(^{11} \) The agent does not know how many steps it takes to cross the fence (or whether crossing is at all possible).

**Seasons scenario:** The environment used in this scenario is portrayed in Fig. 3(e) and contains two types of prey. *Hares* appear in (3 : 1) and contribute to the agent’s fitness with a value of 1, while *Rabbits* appear in (3 : 5) and contribute with a value of 0.1. As with 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 5,000 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 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 with a positive amount of 1 and when poisoned, with a negative value of −1. As in the Seasons scenario, the health status of hares changes every 5,000 steps. Again, the agent knows neither the current season nor whether a prey is poisoned or not.

**Agent Description**

In all scenarios, the agent is modeled as a POMDP whose state dynamics follow from the descriptions above. In all scenarios we treat observations as states and use prioritized sweeping to learn a policy mapping observations to actions (Moore and Atkeson, 1993). As discussed in Section 3, prioritized sweeping constructs a model of the environment and uses this model to perform value-iteration updates. Specifically, our agent maintains an estimate \( \hat{P}^{(t)}(s' \mid s, a) \) of the transition probabilities as perceived by the agent, given by

\[
\hat{P}^{(t)}(s' \mid s, a) = \frac{1}{n_t(s, a)} \sum_{\tau=1}^T \mathbb{1}(s, a, s') (s_\tau, a_\tau, s_{\tau+1}).
\]

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, prioritized 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 life-time, the agent uses an \( \varepsilon \)-greedy exploration strategy with a decaying exploration parameter \( \varepsilon_t = \lambda^t \), where

\(^{11} \) Denoting by \( n_t(\text{fence}) \) 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 as either Hare Season or Rabbit Season with equal probability.
The comparative results of the experiments in the Hungry-Thirsty 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 value of $p < 0.02$. Figure 4(a) further supports our conclusions, providing a depiction of the

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Optimal Parameter Vector</th>
<th>Mean Fitness</th>
<th>Standard ($r_E$)</th>
<th>Random ($r_0$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hungry-Thirsty</td>
<td>$\theta^* = [-0.4, 0.0, 0.0, 0.5, 0.1]^T$</td>
<td>9,505.6 ± 7,303.6</td>
<td>7,783.7 ± 6,930.1</td>
<td>35.6 ± 40.6</td>
</tr>
<tr>
<td>Lairs</td>
<td>$\theta^* = [0.1, 0.0, -0.2, 0.0, 0.7]^T$</td>
<td>8,635.8 ± 1,133.3</td>
<td>7,536.7 ± 944.8</td>
<td>173.3 ± 13.5</td>
</tr>
<tr>
<td>Moving-Preys</td>
<td>$\theta^* = [0.4, 0.0, -0.1, 0.2, -0.3]^T$</td>
<td>1,986.9 ± 110.0</td>
<td>381.3 ± 17.2</td>
<td>683.1 ± 25.7</td>
</tr>
<tr>
<td>Persistence</td>
<td>$\theta^* = [-0.1, 0.1, -0.1, 0.1, 0.6]^T$</td>
<td>1,879.8 ± 11.2</td>
<td>136.3 ± 1.4</td>
<td>17.1 ± 0.7</td>
</tr>
<tr>
<td>Seasons</td>
<td>$\theta^* = [0.0, 0.1, 0.6, 0.0, 0.3]^T$</td>
<td>6,142.3 ± 1,336.3</td>
<td>4,959.3 ± 1,862.4</td>
<td>105.7 ± 24.4</td>
</tr>
<tr>
<td>Poisoned-Prey</td>
<td>$\theta^* = [0.1, -0.2, 0.1, 0.0, 0.6]^T$</td>
<td>5,237.6 ± 77.2</td>
<td>1,284.3 ± 4.3</td>
<td>80.6 ± 24.9</td>
</tr>
</tbody>
</table>

$\lambda = 0.999$. We use a novelty rate $\lambda_n = 1.001$ for the computation of the novelty reward-feature in (5). In all experiments, we consider a discount of $\gamma = 0.9$.

**Reward parameter optimization**

We consider the space of rewards, $\mathcal{R}$, as the set of all rewards of the form $r(s, a, h) = \phi(s, a, h)^T \theta$, where $\phi(s, a, h)$ is the set of all reward features described in Section 4.3 and $\theta$ is the vector containing the corresponding parameters representing the weight or contribution of each feature to the overall reward. In order to determine the optimal reward function $r^*$ (or, equivalently, the corresponding optimal parameter vector $\theta^*$) 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 5-dimensional hypercube $I = [-1; 1]^5$, and sample a total of $K = 14,003$ uniformly distributed parameter vectors from $I$, enforcing $||\theta_k||_1 = 1, k = 1, \ldots, K$.

As discussed in Section 3.2, we consider the fitness function defined in (4). To evaluate the fitness of an agent driven by reward $r_k = \phi^T \theta_k, k = 1, \ldots, K$, we run a total of $N = 200$ independent Monte-Carlo trials, each of which consisting 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, $\theta^*$, such that $\theta^* = \arg\max_{\theta_k, k=1,\ldots,K} F(r_k)$.

### 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, $\theta^*$, resulting from the parameter optimization procedure described in Section 5.1. We then compare the fitness attained by our “emotion-driven” RL agent, driven by $r^* = \phi^T \theta^*$, that of a “standard” RL agent, driven by a reward $r_E = \phi^T \theta_E$, with $\theta_E = [0, 0, 0, 0, 1]^T$, that considers only the extrinsic component, and a “random” RL agent, driven by a reward $r_0 = \phi^T \theta_0$, with $\theta_0 = [0, 0, 0, 0, 0]^T$, that ignores all reward information. The objective is to assess the usefulness of the proposed emotion-based features when compared against an agent driven only by the designer’s extrinsic reward. The random agent provides a baseline for comparison.\(^{13}\)

The illustrative videos of the observed behaviors in different stages of the learning process in all scenarios are available online at [http://gaips.inesc-id.pt/~psequeira/emot-design/].

\(^{14}\)The high standard deviation of the mean cumulative fitness observed in both agents is due to the different environment configurations leading to very different fitness. For example, when the food and water are both located on the left, the agent must traverse
learning performance of all agents. We also emphasize that the behavior 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 contrasts with previous approach that relied on domain-dependent state information to construct the reward (Singh et al., 2009).

The comparative results of the experiments in the Lairs scenario are depicted in Table 1 and Fig. 4(b). Again, the results show a statistically significant difference (at $p = 10^{-4}$) in performance between the “emotion-driven” and the “standard” RL agents. By looking at the learned policies of both agents, we observe that the emotion-driven agent learned to go from lair to lair, successively pulling and capturing rabbits as soon as a lair becomes empty. In comparison, the standard RL agent typically focuses on one of the two lairs, capturing only the rabbits in that lair. In spite of the small probability of each lair transitioning from empty to occupied, waiting for a rabbit in a single lair proves to be not the best strategy. The observed policy is also consistent with the findings in the “boxes” experiments of Singh et al. (2010), the main difference again lying in the fact that our agent relies on domain-independent reward features. Analyzing the optimal parameter vector $\theta^*$ in Table 1, we can observe a small preference for exploratory behavior (expressed as a positive weight in novelty) and less predictable states (expressed as a negative weight in control). Given the dynamics of the environment, the less controllable states correspond to the lair positions, but not immediately after a rabbit is consumed. Therefore, the optimal parameterization drives the agent to change location after capturing a rabbit. Such “nomad” behavior is not, in itself, directly related with fitness maximization. Instead, it is an intrinsic preference of the agent for certain

---

15We note, however, that the purpose of our experiments is mostly distinct from that of Singh et al. (2009).
situations, a preference to 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 comparative results for the Moving-Preys are depicted in Table 1 and in Fig. 4(c), where we can observe a much larger difference in the performance of the “emotion-driven” and “standard” RL agent, caused by the impact of partial observability—the RL agent keeps looking for preys in the position where it last found one, but the dynamics of the environment ensure that no prey exists there. This makes the performance of the standard RL agent to be inferior even than that of the random agent. Analyzing the optimal parameter vector $\theta^*$ in Table 1, we emphasize the fact that the extrinsic reward has a negative weight which, unlike the standard RL agent, drives the agent away from the position where it last found a prey. On the other hand, there is a large positive weight assigned to novelty, prompting the agent to pursue an exploratory policy.\footnote{The results in this scenario are also in line with the findings of Singh et al. (2010) in a similar setting.}

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 in position (2 : 1) acts as an apparent obstacle that, if successfully overcome, will lead to an improved performance in terms of fitness. The comparative results of this scenario are depicted in Table 1 and in Fig. 4(d). The emotion-driven agent clearly outperforms the standard RL agent. Further analysis over the learned policies shows that the standard RL agent is able to capture some “higher-valued” hares in the initial stages of the simulation, but as the fence becomes more difficult to cross and exploration decays, it finally settles on capturing the “lower-valued” rabbits as they provide a more accessible reward. This behavior can also be perceived from the learning curves in Fig. 4(d), where the difference between the emotion-driven and the standard RL agent only becomes apparent after around 20,000 time-steps, when the maximum number of actions was already required to cross the fence.\footnote{This value was confirmed experimentally.} In contrast, the emotion-driven agent “stubbornly” prefers to pursue the hares in favor of the rabbits. By analyzing the optimal parameter vector $\theta^*$ 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. On the other hand, valence rewards states and actions with a value “above average”, also pushing the agent towards the hare. Overall, a balanced combination of the different features provides the best policy, motivating the agent to cross the fence and attain higher rewards—even when the cross-the-fence behavior, in itself, has no direct impact on fitness.

The comparative results of the experiments in the Seasons scenario, detailed in Table 1 and depicted in Fig. 4(e), again show a statistically significant difference in performance between the emotion-driven agent and the standard RL agent ($p < 10^{-4}$). Analyzing 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 venturing 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” behavior leads to a better adaptation to the environment, unlike some of the previous scenarios in which exploratory behaviors 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 the standard RL agent engage in fitness-enhancing policies, as can be seen by noting the positive slope in both learning curves of Fig. 4(f). In this scenario, the difference in attained fitness is due to the fact the standard RL agent, relying on extrinsic reward only, preferred to capture rabbits throughout. From its perspective, this is a sensible behavior since, on average, eating a hare results in a (fitness-based) reward of 0. The emotion-driven agent, on the other hand, prefers to capture hares throughout. Interestingly, as seen in Fig. 4(f), it is able to handle the poisoned seasons. In fact, the fitness curve indicates that in the healthy season the agent is able to consume a large amount of hares. In the poisoned season, on the other hand, after suffering some initial loss (as evidenced by the small peaks), the agent mostly stops capturing hares.\textsuperscript{18} By analyzing the corresponding parameter vector $\theta^*$, we notice that the agent is driven by an interesting balance of positive novelty and control—while novelty fosters exploration (hence the agent’s ability to track the seasons), control fosters the agent to remain in states that it can easily predict (hence 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 behavior that allows the agent to track the non-Markovian dynamics resulting from the season changes. The results in this scenario thus support our hypothesis that emotion-based rewards enable learning agents to better adapt to its environment.

### 5.3 Maladaptation Impacts Fitness

We continue our experimental section and investigate what happens when an agent fit to a certain class of environments $\mathcal{E}_1$ is deployed in an environment $e_2$ significantly different from those in $\mathcal{E}_1$. Much like 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, previously fit to each of the six foraging environments, in the Persistence scenario. As evidenced by the results in Table 2, there is a substantial difference in performance between the several agents—the “maladapted” agents were unable to cope with the difficulties posed by this scenario and, as such, are unable to effectively capture preys. In fact, these agents are “conditioned” by the corresponding environments to handle the emotional reward features in a specific way that allows them to thrive therein. However, in this scenario, those reward features actually distract the agent from the preys, and they end up performing even worse than the standard RL agent (some even worse than the random agent).

---

\textsuperscript{18}Interestingly, the slightly negative slope of the curve indicates that, every now and then, it goes back and attempts to capture a hare again, allowing it to effectively monitor the season changes.
Table 3: Mean rank values for the “universal” emotion-based agent using $\theta_U$, “standard” agent using $\theta_E$ and “random” agent using $\theta_0$, across all the foraging scenarios; see text for details.

<table>
<thead>
<tr>
<th>Parameter Vector $\theta$</th>
<th>Mean Rank</th>
</tr>
</thead>
<tbody>
<tr>
<td>Universal $\theta_U$</td>
<td>522.7 ± 460.2</td>
</tr>
<tr>
<td>Standard $\theta_E$</td>
<td>779.8 ± 602.0</td>
</tr>
<tr>
<td>Random $\theta_0$</td>
<td>6243.0 ± 2996.7</td>
</tr>
</tbody>
</table>

5.4 “Universal” Agent

On the other hand, 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 on all scenarios. For that purpose, we measured the average “rank” of each parameter vector across all the foraging scenarios. The rank value for a specific scenario is taken by sorting all the tested parameter vectors $\theta_k, k = 1, \ldots, K$ in descending order, according to the respective mean cumulative fitness attained in that scenario. This means that the optimal parameter vector $\theta^*$ for a scenario corresponds to the highest ranked vector in that scenario, having 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, corresponding to the universal parameter vector, which is denoted by $\theta_U$.

Table 3 presents the configuration obtained for $\theta_U$ and a comparison between the resulting rankings for the universal, “standard” RL and “random” agents. The configuration of $\theta_U$ indicates that a combination of fitness-based reward and negative control allows the universal agent of attaining a good overall performance, which is justified by the non-stationarity of most of the foraging scenarios, therefore favoring “less controllable” situations. However, as one would expect given the rank value of $\theta_U$, when the performance of the universal parameter vector is taken individually in each scenario, the results are only marginal when compared to the corresponding optimal parameter vectors, as indicated in Table 4. Nonetheless, when compared to the standard agent, the performance of the universal agent was in line with that of the standard agent, and in all but the Hungry-Thirsty, Prey 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 “specialized” in any particular environment, is “good enough” across all scenarios, especially when compared to an agent learning only with the external task reward. In the context of our study this result thus point towards the general-purpose and usefulness of emotion-based rewards in solving complex learning tasks. We note however that such universal configuration is still dependent on the particular set of foraging environments in which learning took place. It is therefore expected that, in scenarios that have dynamics and challenges quite distinct from those presented by our foraging scenarios, the discovered universal agent performs worse than the agent best adapted to such environments or even the standard fitness-based agent.

---

19 We tested different methods of determining the “universal” parameter vector and discovered this ranking procedure to be the best in selecting a configuration which, on average, provides the agent a good performance in all scenarios. In particular, we tested procedures relying on the average fitness across all scenarios, which yielded agents performing very well in scenarios providing a high degree of fitness, e.g., Hungry-Thirsty, but behaved poorly in lower maximal fitness scenarios, e.g., Persistence (see fitness in Table 1).

20 The difference between the rank value of the universal and “standard” agent reported in Table 3 is significant ($p = 8 \times 10^{-4}$).
Table 4: Comparison of the mean cumulative fitness attained by the “universal” emotion-based agent, using $\theta_U$, the “optimal” agent, using $\theta^*$ and a standard RL agent, using $\theta_E$, in each foraging scenario.

<table>
<thead>
<tr>
<th>Scenario</th>
<th>Universal ($\theta_U$)</th>
<th>Mean Fitness</th>
<th>Optimal ($\theta^*$)</th>
<th>Standard ($\theta_E$)</th>
</tr>
</thead>
<tbody>
<tr>
<td>Hungry-Thirsty</td>
<td>8,297.8 ± 5,933.5</td>
<td>9,505.6 ± 7,303.6</td>
<td>7,783.7 ± 6,930.1</td>
<td></td>
</tr>
<tr>
<td>Lairs</td>
<td>8,798.0 ± 1,576.6</td>
<td>8,635.8 ± 1,133.3</td>
<td>7,536.7 ± 944.8</td>
<td></td>
</tr>
<tr>
<td>Moving Preys</td>
<td>460.8 ± 49.2</td>
<td>1,986.9 ± 110.0</td>
<td>381.3 ± 17.2</td>
<td></td>
</tr>
<tr>
<td>Persistence</td>
<td>470.8 ± 59.0</td>
<td>1,879.8 ± 11.2</td>
<td>136.3 ± 1.4</td>
<td></td>
</tr>
<tr>
<td>Prey Seasons</td>
<td>4,912.0 ± 2,606.3</td>
<td>6,142.3 ± 1,336.3</td>
<td>4,959.3 ± 1,862.4</td>
<td></td>
</tr>
<tr>
<td>Poisoned Prey</td>
<td>1,279.7 ± 5.2</td>
<td>5,257.6 ± 77.2</td>
<td>1,284.3 ± 4.3</td>
<td></td>
</tr>
</tbody>
</table>

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 inspired in major dimensions of appraisal. Such reward features, together with an additional extrinsic reward feature (encoding an “environment-imposed” task), are used to construct an “environment-adjusted” intrinsic reward that will guide the process of decision-making of each agent in that environment. Much like emotions in biological agents, our emotion-based reward features evaluate specific aspects of the history of interaction of the agent with its environment, providing additional information that complements its perceptual information. Our results in several scenarios show that the proposed reward features enable the emergence of complex behaviors that allow our agents to largely outperform standard RL agents in their ability to cope with the multiple difficulties posed by each environment.

To conclude our paper, we look into additional links between our emotion-driven RL agents and biological agents in nature. One first observation is concerned with the natural intrinsic motivation mechanisms which, from a physiological point of view, do not address any specific tissue deficit like hunger or thirst (Ryan and 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 that provide an idea of control or competence over the external environment (Ryan and Deci, 2000). This equilibrium is also found in our emotion-driven agents. As seen from the results 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 is concerned with the role of emotions in biological organisms, where 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 behaviors in order to survive longer and procreate more (Cardinal et al., 2002; Dawkins, 2000). Studies on the neural basis of emotions claim that these anticipatory mechanisms can be explained by simple associative learning processes providing an ability to change behavior in response to arbitrary stimuli, and an ability to extend the range of stimuli 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 behaviors (Cardinal...
This emotion-based “evolutionary conditioning” of organisms finds a parallel in our process of parameter optimization. Roughly speaking, this process “hardwires” in our agents associations between our reward features and the agent’s fitness, endowing the agent with the ability to learn complex behaviors that provide adaptive advantages in the environment.

It is also noteworthy that, as evidenced in Secs. 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, demanding a completely different set of strategies in order to obtain fitness, and perhaps the universal agent would have performed poorly under such conditions.

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

In the future we plan to extend this research to multiagent settings. In particular, we are interested in addressing the ORP in multiagent settings within IMRL. For that purpose, we intend to design domain-independent reward features that assess the social-acceptability of behaviors in order 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 for both the specificity of the agent (its learning algorithm, exploration policy, etc.) and its environment to complement the agent’s perceptions. In this sense, the optimization procedure that is required in the context of the ORP to determine the optimal reward, resembles the environmental pressures that biological organisms are subject to throughout evolution. And, in the case of our emotion-based rewards, like evolution favors behaviors that seem to enhance the fitness of the agent in the long-run, the optimization of our biologically-inspired reward mechanism enables our agents to behave, learn, and “live” much like biological organisms do.
Acknowledgments

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

References


