Emotion-Based Intrinsic Motivation for Reinforcement Learning Agents

Pedro Sequeira, Francisco S. Melo, and Ana Paiva

Instituto Superior Técnico
Universidade Técnica de Lisboa
INESC-ID
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 propose an adaptation of four common appraisal dimensions that evaluate the relation of an agent with its environment into reward features within an *intrinsically motivated reinforcement learning* framework. We show that, by optimizing the relative weights of such features for a given environment, the agents attain a greater degree of fitness while overcoming some of their perceptual limitations. This optimization process resembles the evolutionary adaptive process that living organisms are subject to. We illustrate the application of our method in several simulated foraging scenarios.

Keywords: reinforcement learning, intrinsic motivation, appraisal.

1 Introduction

Emotions have often been regarded as detrimental to cognition by impairing rational decision-making. However, as the body of knowledge about the influence of emotions on humans and other animals grows, emotions are increasingly being regarded as a beneficial adaptive mechanism for decision-making [5,9]. Studies in animals and simple organisms showed that, throughout evolution, emotions might have provided animals with an ability to survive longer and procreate more [4,5]. This is done by means of associative learning processes that allow organisms to extend the range of stimuli perceived as hazardous or beneficial and focus their attention in important aspects of the environment while changing their behavior accordingly [4,5]. Emotions also play a fundamental role in learning, by eliciting physiological signals that bias our behavior toward maximizing reward and minimizing punishment [4]. Reinforcement learning mechanisms found in nature thus rely on emotional cues to indicate the advantages or adversity of an event. Without such mechanisms, animals could not know "*whether a behavior never performed by any of its ancestors should be repeated or not*" [5].

One way of explaining how emotions are generated according to one's relationship with the environment is by developing appraisal theories of emotion [7,9].

S. D´Mello et al. (Eds.): ACII 2011, Part I, LNCS 6974, pp. 326–336, 2011.

Appraisal theories posit that emotions are elicited by evaluations (appraisals) of events which characterize aspects of the situation in terms of its significance for the organism's well-being or goals [7]. In order to differentiate between emotional states, several theories propose a set of appraisal dimensions, each of which evaluates specific aspects of the subject-environment relationship.

Given the simplicity and usefulness of learning and emotional-processing skills in nature, we expect that these same mechanisms adapted to artificial agents may lead to more robust and adaptable agents. As such, in this paper we adopt the framework of intrinsically motivated reinforcement learning (IMRL) [16] and propose numerical counterparts for some common appraisal dimensions [7,9]. Each of the adopted dimensions evaluates a certain aspect of the agent-environment relation and is translated into a numerical feature that provides intrinsic reward to the agent in a reinforcement learning (RL) context. The specific way in which the agent interprets these features is optimized to the agent's environment, in a process that relates to the evolutionary environmental conditioning that organisms are subject to in nature. Our results show that the contributions from the different reward features in fact lead to distinct behaviors that allow agents to overcome certain shortcomings in particular environments and attain better performance. Moreover, we show that the absence of such emotion-based processing mechanism may have a significant negative impact on the agent's performance in some scenarios. Finally, we show that the proposed appraisal features can be used as general reward features for IMRL agents.

2 Background and Related Work

In the RL field there are only a few systems that make use of emotional processing. Examples include an emotional model for robots that combines the values of the robot's sensations, feelings and "hormones" to determine a dominant emotional state from a set of four basic emotions [8]. The agent learns state-behavior associations that are reinforced by emotions. In another approach, the agent's affective state is computed based on a statistical analysis of the reward it receives [3]. The results show that associating positive affective states with exploitation and negative affect with exploration strategies provides adaptive benefits for the agent. In [14], three basic emotions control the behavior strategy of an agent in an RL task: *Happiness* and *Sadness* are determined based on the amount of reward received by the agent, while *Fear* is used as a decision mechanism that prevents the agent from choosing low-valued actions. One other work proposes a model for *affective anticipatory reward* based on valence and arousal levels, which in turn influences decision-making in a risk-taking scenario [2]. Finally, in the *FLAME* model [6], RL is used to build associations of emotional states and objects and to predict the user's actions.

All aforementioned approaches rely on a set of discrete emotions that influence the learning and decision-making processes of the agent. In this paper, we propose an approach inspired in appraisal theories which, as seen in Section 1, stresses the importance of emotions in providing intrinsic cues for learning in

dynamic environments. We propose a possible numerical translation of four appraisal dimensions to be used as features of intrinsic reward by an RL agent.

2.1 Intrinsically Motivated Learning

Reinforcement learning (RL) addresses the general problem of an agent faced with a sequential decision problem [18]. By a process of trial-and-error, the agent must learn a mapping that assigns perceptions to actions. Such mapping determines how the agent acts in each possible situation and is commonly known as a *policy*. In single agent scenarios, RL agents can be modeled using *partially observable Markov decision processes* (POMDPs). At every step, depending on its observation, the agent chooses an action a_t from a finite set of possible actions, $\mathcal{A}$, and transitions from state s_t to state s_{t+1} with probability $\mathsf{P}(s_{t+1} \mid s_t, a_t)$. It receives a reward $r(s_t, a_t)$ and makes a new observation z_{t+1} from a set of possible observations, $\mathcal{Z}$, with probability $\mathsf{O}(z_{t+1} \mid s_{t+1}, a_t)$, and the process repeats. The goal of the agent is to choose its actions so as to gather as much reward as possible, discounted by a positive discount factor $\gamma < 1$. Formally, this corresponds to maximizing the value

$$v = \mathbb{E}\left[\sum_t \gamma^t r(s_t, a_t)\right]. \tag{1}$$

The reward function r implicitly encodes the *task* that the agent must learn.

In typical RL scenarios, it is assumed that observations correspond to the actual states of the agent/environment [18]. When this is the case, it is possible to find a *policy* $\pi^* : \mathcal{Z} \to \mathcal{A}$ maximizing the value in (1). However, in many environments, the assumption that $z_t = s_t, \forall t$, is too restrictive, and policies mapping observations directly to actions (called *memoryless policies*) can have arbitrarily poor performance [15]. This means that the perceptual limitations of the agent in fact impair its ability to properly choose its actions. Moreover, computing the best memoryless policy is NP-hard in the worst case [10]. Several algorithmic approaches have been proposed to deal with partial observability in RL settings [1]. One important class of approaches builds into the agent *prior knowledge* that can, somehow, alleviate its perceptual limitations. Examples include approaches based on some form of *memory*. However, such approaches typically require specific learning algorithms tailored to leverage information from particular aspects of the agent's history [1].

Recently, a novel framework for *intrinsically motivated reinforcement learning* was proposed [16]. In this framework, a learning agent interacts with one among a set $\mathcal{E}$ of possible environments, and optimizes its policy with respect to one reward function r among a set $\mathcal{R}$ of possible rewards. An optimal reward function $r^* \in \mathcal{R}$ is such that the *expected fitness* of the agent with respect to a distribution over possible environments is maximized. This fitness is determined by some fitness function $\mathcal{F}$ that maps the history of the interaction of an agent with its environment into a numerical value that, in a sense, measures how well-adapted the agent is to its environment. In the IMRL framework, the learning

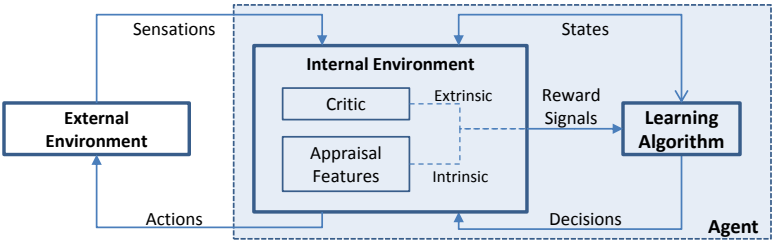


Fig. 1. Proposed framework for emotion-based intrinsic motivation; adapted from [16]

agent receives an *augmented* reward function that incorporates several reward components, herein referred as *reward features* and denoted by $\phi_i, i = 1, \ldots, N$. In this paper, we thus consider $\mathcal{R}$ as the set of all rewards of the form

$$r(s, a) = \sum_i \theta_i \phi_i(s, a), \tag{2}$$

where the weights θ_i determine the contribution of each reward-feature ϕ_i to the overall reward r that the agent must maximize throughout its lifetime. We refer to a *fitness-based reward signal* as a reward-feature r^{ext} that explicitly rewards fitness-maximizing states [16]. For ease of exposition, we henceforth refer to such feature as the *extrinsic reward*, which can be interpreted as corresponding to the fulfillment of some of the agent's basic needs. For example, if the agent is a predator, its task could be to find its prey and feed, and the extrinsic reward would correspond to the predator being satiated. The other reward-features constitute the *intrinsic reward*, which contrasts with the "original" extrinsic reward in that it does not necessarily relate to the task that the agent must accomplish. However, as shown in [16,17], intrinsic rewards can be an effective mechanism to endow the agent with useful information to overcome some of its perceptual limitations (such as memory) and enhance its performance. One important aspect that remains unexplored, however, is concerned with which information should be used to build these intrinsic rewards. In this paper we propose the use of simulated affective features inspired in common appraisal dimensions to build the intrinsic reward.

3 Emotionally Motivated Learning Agents

Figure 1 depicts the proposed framework for emotion-based intrinsically motivated learning, adapted from [16]. In this framework, we follow the perspective that certain affective states may encode useful information that guide an agent during learning and decision-making. In particular, we adopt four common appraisal dimensions: novelty, motivation, valence and control. Inspired by appraisal theories of emotion, we propose for each dimension a possible reward feature that evaluates certain aspects of the agent-environment relationship (corresponding to the internal environment in Fig. 1). These features map the result

of appraisals into scalar values that somehow indicate the degree of activation of each dimension. In our framework, an agent receives a total reward r^{tot} calculated as a linear combination of all the proposed features,

$$r^{\text{tot}}(s,a) = \theta^{\mathfrak{n}}\mathfrak{n}(s,a) + \theta^{\mathfrak{m}}\mathfrak{m}(s,a) + \theta^{\mathfrak{c}}\mathfrak{c}(s,a) + \theta^{\mathfrak{v}}\mathfrak{v}(s,a) + \theta^{\text{ext}}r^{\text{ext}}(s,a), \quad (3)$$

where weights $\theta^{\mathfrak{n}}$, $\theta^{\mathfrak{m}}$, $\theta^{\mathfrak{c}}$, $\theta^{\mathfrak{v}}$ and θ^{ext} are scalar values between 0 and 1 which are initially set for the agent and remain fixed throughout its lifetime. A particular weight set $\boldsymbol{\theta} = [\theta^{\mathfrak{n}}, \theta^{\mathfrak{m}}, \theta^{\mathfrak{c}}, \theta^{\mathfrak{v}}, \theta^{\text{ext}}]$ corresponds to a built-in configuration for the agent that indicates which aspects of its relationship with the environment it gives more attention to. For example, a weight configuration $\boldsymbol{\theta} = [0, 0, 0, 0, 1]$ indicates that the agent is predisposed to value only extrinsic rewards while completely ignoring intrinsic motivation. Each particular weight configuration will yield different degrees of fitness depending on the environment where the learning takes place. Due to this fact, the weight set is optimized to maximize the agent's fitness according to the environment, which will also allow it to overcome some of its perceptual limitations. The optimal weight set is denoted by $\boldsymbol{\theta}^*$.

3.1 Affective Reward Features

Appraisal theories define a set of dimensions to generate affective states in response to events [7]. Our framework adopts four of the *major dimensions of appraisal* [7] that characterize many of the existing appraisal theories. We intentionally do not adapt common *social dimensions* as they are responsible for more complex emotions like shame or guilt that we do not explore here.

Appraisal theories characterize high-level psychological processes for the generation of emotions in humans. Some of the commonly proposed dimensions deal with complex concepts and mental representations such as beliefs, causal attribution or social norms [9,7]. However, in our learning framework, we are concerned with aspects of the agent-environment relationship capable of affecting the agent's behavior. Because of that, one of the main challenges is to map the evaluations made by the appraisal dimensions into low-level numerical features that can be used as intrinsic reward-features. Leventhal and Scherer [9] discuss the possibility of appraising events from the environment at three different levels of processing: sensory-motor, schematic and conceptual. This way, by considering the events at different levels, it is possible to explain emotions as an adaptive mechanism that develops from simple, reflex-like responses into more complex cognitive-emotional patterns through learning [9].

Considering this multi-level view of emotional appraisal, we now describe a possible mapping of the adapted appraisal dimensions into scalar values, corresponding to the reward features in the IMRL framework. We are not claiming these mappings to be the only ones possible. We propose low-level features that, in our framework, make similar evaluations to those expected from the respective appraisal dimensions. We henceforth denote by $n_t(s)$ the number of visits to state s up to time-step t, and by $n_t(s,a)$ the number of times that action a has been experienced in state s.

Novelty usually refers to the degree of familiarity of the perceived stimuli in relation to the agent's knowledge structures built so far [7,9]. Statistically, at a low-level, this is directly correlated to the number of visits to state-action pairs. Therefore, in our framework, this dimension is quantified as the reward-feature

$$\mathfrak{n}(s,a) = \frac{\lambda^{n_t(s,a)} + \lambda^{n_t(s)}}{2},$$

where λ is a positive constant such that $\lambda < 1$. λ can be seen as a "novelty rate", determining how the novelty dimension decays with experience.

Motivation asserts the *relevance* of a perceived event in terms of the agent's goals or needs [7,9]. As such, motivation increases as the agent approaches its perceived goals, and decreases otherwise. For the purposes of our model, we assume that, at each time-step t, the agent has access to an estimated distance, $\hat{d}(s_t, s^*)$, that returns an estimate of the number of actions needed to move from its current state s_t to a goal-state s^* where the reward is maximal. This distance estimate needs not be accurate, but should be coherent with the true values, *i.e.*, if $d(s_1, s^*) > d(s_2, s^*)$ then $\hat{d}(s_1, s^*) > \hat{d}(s_2, s^*)$, where $d(\cdot, \cdot)$ denotes the actual distance. This distance estimate is not computed from the model, but perceived by the agent. Motivation is thus translated in terms of the numerical value

$$\mathfrak{m}(s,a) = \frac{1}{1 + \hat{d}(s, s^*)}.$$

Control, depending on the level at which the appraisal is being made, indicates the *potential* of an agent in coping with the situation being evaluated [7,9] or the degree of *correctness/clarity* of the world-model that the organism has built of its own environment. Statistically, this is highly correlated with the number of visits to state-action pairs in a reverse manner to that of novelty. Therefore, for simplicity, we represent the amount of control as $\mathfrak{c}(s,a) = 1 - \mathfrak{n}(s,a)$. Because the agent becomes more familiarized with its environment overtime, the states, actions and even rewards that are not novel enhance the degree of *correctness* of its world-model.

Valence measures how pleasant a given situation is. It is a product of innate detectors or learned preferences/aversions that basically indicate whether a stimulus is "positive" or "negative" in terms of biological significance for the organism [7,9]. In our framework, we translate this as a measure of how much the current extrinsic reward r^{ext} contributes to the overall goal of the agent, which corresponds to the value

$$\mathfrak{v}(s,a) = \frac{r^{\text{ext}}(s,a)}{V^{\text{ext}}(s) - \gamma \sum_{s'} \mathsf{P}(s' \mid s,a) V^{\text{ext}}(s')},$$

where $V^{\text{ext}}(s)$ denotes the value (in terms of total discounted extrinsic reward) that the agent currently associates with state s. The expression for $\mathfrak{v}$ essentially follows the extrinsic reward-feature, but weighting it with respect to its contribution to the overall value that the agent expects to achieve[1].

[1] This expression is related to a well-known *shaping function* [12].

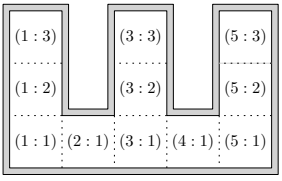


Fig. 2. The foraging environment used in the experiments. Each square marked by $(x : y)$ coordinates represents a possible location for the agent.

4 Experiments and Results

We designed a set of experiments in foraging environments inspired in [16] to illustrate the emergence of specific behaviors that overcome perceptual limitations of the agent and ultimately lead to a better fitness. In our experiments, the agent is a predator moving in the environment depicted in Fig. 2. At each time-step, the agent is able to observe its position in the environment and whether it is collocated with a prey. The agent has available 5 possible actions, $\{N, S, E, W, \text{Eat}\}$. The four directional actions move the agent deterministically to the adjacent cell in the corresponding direction; the Eat action consumes a prey if one is present at the agent's location, and does nothing otherwise. We use the Dyna-Q/prioritized sweeping algorithm [11] to learn a memoryless policy that treats observations of the agent as states. The agent follows an ε-greedy policy with decaying exploration rate $\varepsilon_t = \lambda^t$. We use a learning rate $\alpha = 0.3$, a novelty rate $\lambda = 1.0001$ and $\gamma = 0.9$. We ran four different experiments that differ in the particular distribution of preys and on the outcome of the agent's actions.

Exploration scenario: This scenario closely resembles the foraging scenario in [16]. At each time step, a prey can appear in any one the three end-of-corridor locations, $(1 : 3)$, $(3 : 3)$ or $(5 : 3)$. Whenever the agent eats a prey, it receives an extrinsic reward of 1. The prey disappears from its current location and randomly reappears in one of the two other locations.

Persistence scenario: In this scenario two preys are always available at $(1 : 3)$ and $(5 : 3)$. The prey in $(1 : 3)$ is a hare and corresponds to an extrinsic reward of 1. The prey in $(5 : 3)$ is a rabbit and corresponds to an extrinsic reward of 0.01. Every time the predator eats one of the preys, it returns to its initial position in $(3 : 3)$ and the prey is replaced in the corresponding location. We assume that eating is *automatic*: every time the predator is collocated with a prey automatically eats it. Additionally, in $(2 : 1)$ there is a fence that prevents the predator from moving from $(2 : 1)$ to $(1 : 1)$ in a single time-step. It will take the predator n successive E actions to "break" the fence for the nth time and move from $(2 : 1)$ to $(1 : 1)$. Every time that the fence is broken, it is rebuilt more solidly, requiring the agent to take $n+1$ actions to break it the next time[2].

[2] We note that the fence only prevents the agent from moving from $(2 : 1)$ to $(1 : 1)$, and not in the opposite direction.

Prey-season scenario: In this scenario only one of two kinds of prey is available at each time-step. In the "hare season", a hare is available in $(5:3)$ at every time-step. Every time the predator eats the hare it receives a reward of 1 and returns to its start location in $(3:3)$. In the "rabbit season", a rabbit is available in $(1:3)$, providing a reward of 0.1 when eaten. However, in each rabbit season, after eating 9 rabbits, a rabbit breeder shoots the predator whenever it tries to eat another, and our agent receives a punishment reward of -1, returning to the start position. Seasons switch every 10 000 steps.

Different rewards scenario: Finally, this is a rather simple scenario in which two preys are always available at $(1:3)$ (a rabbit, worth a reward of 0.1) and at $(5:3)$ (a hare, worth a reward of 1). Like with the previous scenarios, the agent returns to $(3:3)$ every time it eats a prey.

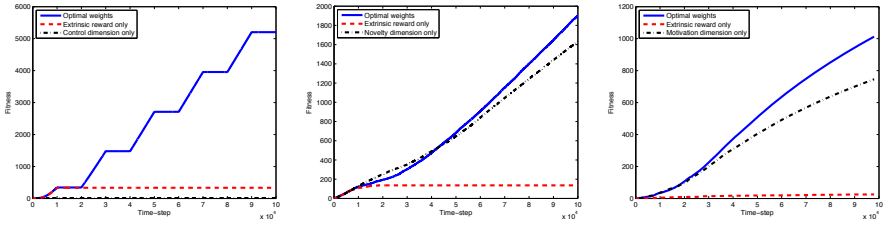
We note that, from the agent's perspective, the scenarios are non-Markovian, since the information about the location of the preys is not directly observable. This emulates some of the challenges that predators face in nature, where their observations do not provide all the necessary information for the best choice of action. The different scenarios were designed with two goals in mind: (i) to test whether the affective reward-features lead to distinct behaviors; and (ii) to determine if, by using them as intrinsic rewards, the agent improves its overall fitness. The weight vector $\boldsymbol{\theta} = [\theta^{\mathfrak{n}}, \theta^{\mathfrak{m}}, \theta^{\mathfrak{c}}, \theta^{\mathfrak{v}}, \theta^{\text{ext}}]$ is optimized for each environment to maximize the agent's fitness, using an adaptive sampling approach similar to the one in [16][3]. This optimization process is, in a sense, similar to the adaptive processes animals experience throughout their evolution.

Table 1. Agent fitness results for each scenario. The first column indicates the optimal weight set $\boldsymbol{\theta}^*$ obtained for each scenario. The column marked "Optimal" corresponds to the amount of fitness resulting from the optimized weight vector, while the "Extrinsic" column corresponds to the standard Dyna-Q agent ($\boldsymbol{\theta} = [0, 0, 0, 0, 1]$).

Scenario	$\boldsymbol{\theta}^* = [\theta^{\mathfrak{n}}, \theta^{\mathfrak{m}}, \theta^{\mathfrak{c}}, \theta^{\mathfrak{v}}, \theta^{\text{ext}}]$	Optimal	Extrinsic
Prey-season	$\boldsymbol{\theta}^* = [0.00, 0.00, 0.50, 0.00, 0.50]$	5 203.5	334.2
Exploration	$\boldsymbol{\theta}^* = [0.40, 0.00, 0.20, 0.00, 0.40]$	1 902.2	135.9
Persistence	$\boldsymbol{\theta}^* = [0.13, 0.29, 0.29, 0.00, 0.29]$	1 020.8	25.4
Dif. rewards	$\boldsymbol{\theta}^* = [0.00, 0.50, 0.00, 0.25, 0.25]$	87 925.7	87 890.8

Table 1 presents the obtained results of simulating the agent for 100 000 learning steps and correspond to averages of 100 independent Monte-Carlo trials. We present the optimal weight set $\boldsymbol{\theta}^*$ obtained for each scenario and the corresponding amount of fitness attained by the agent. For comparison purposes, we also present the fitness of a standard Dyna-Q agent receiving only extrinsic reward, which corresponds to the weight set $\boldsymbol{\theta} = [0, 0, 0, 0, 1]$. From the results

[3] Although more efficient methods are possible [13], we are not concerned with the computational efficiency of the process of reward optimization.



(a) Prey-season scenario (b) Exploration scenario (c) Persistence scenario

Fig. 3. Cumulative fitness attained in the prey-season, exploration and persistence scenarios. The results correspond to averages over 100 independent Monte-Carlo trials. We included the results for the optimal weight set, for an agent receiving only extrinsic reward, and for an agent receiving only the intrinsic reward component corresponding to the highest weight in the optimal weight set.

in the table, it is clear that our agent outperforms the standard Dyna-Q agent in all tested scenarios, supporting our claim that the features adapted from the emotional appraisal theories provide useful information that allows the agent to overcome some of its limitations. Our results also prompt several other interesting observations. Note, for example, that in the exploration scenario, the intrinsic reward arising from novelty is sufficient to significantly outperform the agent only pursuing extrinsic reward. These results are in accordance with those reported in [16,17] in a similar scenario. In general, depending on the scenario, the weight optimization procedure yields a distinct configuration that is related with specific aspects of the environment. This is also an important observation that supports our claim that different emotion-based features foster different behaviors. The prey-season scenario also provides a very interesting result: individually, neither the extrinsic reward or the control alone are sufficient for the agent to attain a significant performance. However, when combined, they lead to a boost in performance of at least one order of magnitude. This is due to the fact that each feature provides a different strategy: on one hand, the agent must consider extrinsic reward provided by the hares; on the other hand, it should choose more familiar actions during the rabbit season, as eating too much rabbits will result in negative reward. Fig. 3 depicts the learning performance of our agent in three of the test scenarios, against the performance of a "greedy" agent and an agent receiving reward only from one of the intrinsic features. This figure helps to further illustrate the behavior of our approach, showing that not having an emotional mechanism guiding the agent can severely impair its learning performance. It also shows that generally, a combination of the different proposed features is important to attain the best result, *i.e.*, it does not suffice receiving intrinsic reward from only one of the emotional features.

5 Discussion

In this paper we proposed a framework for generating intrinsic rewards for learning agents. The intrinsic reward is derived from four appraisal dimensions adapted from literature that map into reward-features. We modeled our agents within an IMRL framework and designed a series of experiments to test the validity of our approach. Our results show that the proposed affective features guide the agent in finding a right balance between different behavior strategies in order to attain the maximal fitness in each scenario. Our objective in this work was not to find general feature-weight configurations, but generic features that could be used to produce intrinsic reward. We believe that the success of this approach may be due to the fact that the reward-features, much like the appraisal dimensions they correspond to, characterize aspects of the agent's relationship with its environment. Because the features are embedded in the reward, they indirectly focus the agent in different aspects, bringing out attention to advantageous states while ignoring others that do not seem so favorable. In the future we would like to extend our framework to multiagent scenarios in order to test the appearance of socially-aware behaviors by the agents. This could be done by adding a social intrinsic reward-feature that evaluates whether certain behaviors by the agents are considerate in relation to the overall fitness of the population instead of considering only their own individual fitness.

Acknowledgments. This work was partially supported by the Portuguese Fundação para a Ciência e a Tecnologia (INESC-ID multiannual funding) through the PIDDAC Program funds. The first author acknowledges the PhD grant SFRH/BD/38681/2007 from the Fundação para a Ciência e a Tecnologia.

References

1. Aberdeen, D.: A (revised) survey of approximate methods for solving partially observable Markov decision processes. Technical report, NICTA (2003)
2. Ahn, H., Picard, R.: Affective cognitive learning and decision making: The role of emotions. In: EMCSR 2006: The 18th Europ. Meet. on Cyber. and Syst. Res. (2006)
3. Broekens, D.: Affect and learning: a computational analysis. Doctoral Thesis, Leiden University (2007)
4. Cardinal, R., Parkinson, J., Hall, J., Everitt, B.: Emotion and motivation: The role of the amygdala, ventral striatum, and prefrontal cortex. Neuroscience and Biobehavioral Reviews 26(3), 321–352 (2002)
5. Dawkins, M.: Animal minds and animal emotions. American Zoologist 40(6), 883–888 (2000)
6. El-Nasr, M., Yen, J., Ioerger, T.: FLAME - Fuzzy logic adaptive model of emotions. Auton. Agents and Multiagent Systems 3(3), 219–257 (2000)
7. Ellsworth, P., Scherer, K.: Appraisal processes in emotion. In: Handbook of Affective Sciences, pp. 572–595. Oxford University Press, Oxford (2003)
8. Gadanho, S., Hallam, J.: Robot learning driven by emotions. Adaptive Behavior 9(1), 42–64 (2001)

9. Leventhal, H., Scherer, K.: The relationship of emotion to cognition: A functional approach to a semantic controversy. Cognition & Emotion 1(1), 3–28 (1987)
10. Littman, M.: Memoryless policies: Theoretical limitations and practical results. From Animals to Animats 3, 238–245 (1994)
11. Moore, A., Atkeson, C.: Prioritized sweeping: Reinforcement learning with less data and less real time. Machine Learning 13, 103–130 (1993)
12. Ng, A., Harada, D., Russel, S.: Policy invariance under reward transformations: Theory and application to reward shaping. In: Proc. 16th Int. Conf. Machine Learning, pp. 278–287 (1999)
13. Niekum, S., Barto, A., Spector, L.: Genetic programming for reward function search. IEEE Trans. Autonomous Mental Development 2(2), 83–90 (2010)
14. Salichs, M., Malfaz, M.: Using emotions on autonomous agents. The role of Happiness, Sadness and Fear. In: AISB 2006: Adaption in Artificial and Biological Systems, pp. 157–164 (2006)
15. Singh, S., Jaakkola, T., Jordan, M.: Learning without state-estimation in partially observable Markovian decision processes. In: Proc. 11th Int. Conf. Machine Learning, pp. 284–292 (1994)
16. Singh, S., Lewis, R., Barto, A., Sorg, J.: Intrinsically motivated reinforcement learning: An evolutionary perspective. IEEE Trans. Autonomous Mental Development 2(2), 70–82 (2010)
17. Sorg, J., Singh, S., Lewis, R.: Internal rewards mitigate agent boundedness. In: Proc. 27th Int. Conf. Machine Learning, pp. 1007–1014 (2010)
18. Sutton, R., Barto, A.: Reinforcement Learning: An Introduction. MIT Press, Cambridge (1998)