Special issue on Grounding Emotions in Robots

Adaptive Behavior

Adaptive Behavior
2016, Vol. 24(5) 292–305
© The Author(s) 2016
Reprints and permissions:
sagepub.co.uk/journalsPermissions.nav
DOI: 10.1177/1059712316667202
adb.sagepub.com
SAGE

Happiness as an intrinsic motivator in reinforcement learning

Yue Gao[1] and Shimon Edelman[2]

Abstract
Reinforcement learning, a general and universally useful framework for learning from experience, has been broadly recognized as a critically important concept for understanding and shaping adaptive behavior, both in ethology and in artificial intelligence. A key component in reinforcement learning is the reward function, which, according to an emerging consensus, should be intrinsic to the learning agent and a matter of appraisal rather than a simple reflection of external outcomes. We describe an approach to intrinsically motivated reinforcement learning that involves various aspects of happiness, operationalized as dynamic estimates of well-being. In four experiments, in which simulated agents learned to explore and forage in simulated environments, we show that agents whose reward function properly balances momentary (hedonic) and longer-term (eudaimonic) well-being outperform agents equipped with standard fitness-oriented reward functions. Our findings suggest that happiness-based features can be useful in developing robust, general-purpose reward mechanisms for intrinsically motivated autonomous agents.

Keywords
Happiness, well-being, hedonic, eudaimonic, intrinsic motivation, reinforcement learning, reward design

1 Introduction

The evolutionary success of generalist species depends on the ability to learn from experience (Mery, 2013; Tapia et al., 2015); a task at which humans, the ultimate cognitive generalists on this planet, also excel (Nelson, 2015). Reinforcement learning (RL), in which rewards from experiential outcomes guide exploration and the acquisition of knowledge, has also been recognized as an important component in the development of artificial intelligence systems (Mnih et al., 2015; Silver et al., 2016). In this paper, we address the problem of choosing a good reward function, on which the success of RL is predicated. Specifically, we formulate and evaluate several reward functions that involve happiness, operationalized as dynamic evaluation of behaviorally significant outcomes (Gao & Edelman, 2016). We begin by offering some motivation for this research program.

1.1 Why reinforcement learning

Reinforcement learning, which does take into account behavioral outcomes, yet does not require explicit instruction or error correction, is a universal computational framework for learning, in the sense that any learning problem can be cast as an instance of RL (Schmidhuber, 2015). In its reliance on occasionally delivered scalar rewards, RL mimics learning in humans and other animals (Niv, 2009; Singh, Lewis, & Barto, 2009). It is an effective method for learning under uncertainty in dynamic environments: the agent learns "on its own" with no examples of correct behavior being provided. It evaluates its actions in relation to observations and outcomes, through continued interaction with the environment (Sutton & Barto, 1998).

Reinforcement learning is also a useful computational framework for understanding the neural basis of animal learning (Akaishi, Kolling, Brown, & Rushworth, 2016; Doll, Simon, & Daw, 2012; Kakade & Dayan, 2002; Lee, Seo, & Jung, 2012; Mirolli, Santucci, & Baldassarre, 2013; Niv, 2009; Redgrave, Vautrelle, & Reynolds, 2011). In particular, there is growing evidence for the existence of direct neural counterparts of key RL-related computational concepts, such as reward prediction error (Barto, 1995; Knutson & Gibbs, 2007; Pessiglione, Seymour,

[1] Department of Computer Science, Cornell University, USA
[2] Department of Psychology, Cornell University, USA

Corresponding author:
Shimon Edelman, Department of Psychology, Cornell University, Ithaca, NY 14853, USA.
Email: edelman@cornell.edu

Flandin, Dolan, & Frith, 2006). The circuits in the vertebrate brain that participate in RL are being mapped out and analyzed (Lee et al., 2012).

1.2 What makes for a good RL algorithm

Similar to other models of learning, RL works best when provided with some knowledge about its domain of application. The success of RL, in particular, depends critically on the formulation of the reward function (Sequeira, Pedro, Francisco, & Paiva, 2014), which maps outcomes to reinforcement signals and thus constitutes a critical link in the chain of credit assignment. It is possible to design reward functions by hand, such as in robotics (Feijo, Cornell, & Garzn, 2006), but in complex environments it may be infeasible for the designer to handcraft an effective, let alone optimal, reward function (Schembri, Mirolli, & Baldassarre, 2007; Sorg, Singh, & Lewis, 2010).

1.3 Extrinsic versus intrinsic reward

Outcomes provided by the environment, such as encounters with food or opportunities for procreation, are important in shaping behavior, if only because the individual's survival and the species' evolutionary success depends on them. Such outcomes, however, are typically sparse: the consequence of an action may only become available long after the action has been completed, so that a more immediate source of motivation is called for. Moreover, the value of an outcome to the agent depends on the agent's internal state and current goal. These considerations have prompted researchers to distinguish between intrinsic and extrinsic rewards (Barto & Simsek, 2005; Gottlieb, Oudeyer, Lopes, & Baranes, 2013; Kaplan & Oudeyer, 2004; Oudeyer & Kaplan, 2007; Oudeyer, Kaplan, & Hafner, 2007; Singh, Lewis, Barto, & Sorg, 2010a) and ultimately to propose a formulation of RL in which reward signals are produced exclusively by a "critic" that is part of the learning system: as Singh et al. (2010a) put it, "the sources of all of an animal's reward signals are internal to the animal".

1.4 The role of emotions in shaping intrinsic rewards

Intuitively, intrinsic reward involves an appraisal (Scherer, Schorr, & Johnstone, 2001; Sequeira et al., 2014) of the situation by the agent, mediated by signals that are best thought of as emotions; computational shortcuts that motivate decisions and regulate behavior (Lindquist, Wager, Kober, Bliss-Moreau, & Barrett, 2012; Minsky, 2006; Rogers, 1963). In this sense, the behavioral control in animals is a matter of feelings: evidence suggests that the ability to have sensations of pleasure and pain is strongly connected to basic mechanisms of decision-making (Berridge, 2003; Cohen & Blum, 2002). In artificial systems, the introduction of affective computing, which calls for endowing autonomous agents with emotions, was shown to improve performance robustness and efficiency (Marsella, Gratch, & Petta, 2010; Rumbell, Barnden, Denham, & Wennekers, 2012; Salichs & Malfaz, 2012).

1.5 Our contribution and the plan for the paper

In this paper we propose and test a family of reward functions that involve a simple quantitative model of "happiness"; an ongoing, dynamically maintained, multi-component self-estimate of well-being. Our goal is to formulate a reward function that would be general enough to alleviate the need for handcrafting across domains, and also informative enough to be effective in specific tasks.

The paper is organized as follows. The section 'Background and related work' introduces the required background and notation regarding reinforcement learning and intrinsically motivated RL (IMRL). Section 'The components of happiness as determinants of reward' formulates hedonic and eudaimonic well-being, describes the learning algorithm, and uses these to define intrinsic reward functions. Sections 'Experiment 1' to 'Experiment 4' describe the four experiments in which we have compared the performance of different types of agents, some using happiness-based IMRL, in a variety of exploration and foraging tasks. Section 'Discussion' summarizes the main findings and outlines future work.

2 Background and related work

We shall now briefly introduce the two main ideas that the present work combines: the reinforcement learning (RL) framework (in particular, intrinsically motivated RL, or IMRL), and the use of happiness (in particular, its hedonic and eudaimonic components) in modeling the driving factors that shape behavior. For a tutorial discussion of IMRL, the reader is referred to the chapter by Barto (2013), as well as the entire volume in which it appears (Baldassarre & Mirolli, 2013). The possible role of happiness as a variable that mediates between outcomes and motivation in evolutionary agent-based simulations of foraging behavior has been explored by Gao and Edelman (2016).

2.1 Reinforcement learning (RL)

Reinforcement learning is a general approach to sequential decision problems (Sutton & Barto, 1998). By a process of trial-and-error, the agent faced with the decisions must learn a set of policies: a mapping that assigns observed states to actions. At each time step, the agent receives an observation s from its environment E, takes an action a, receives reward r, and repeats this process until a time horizon. The history at time t, denoted by h_t, is the history of interaction $(s_1, a_1, r_1, \ldots, s_t, a_t, r_t)$. Under this formulation, the problem can be cast as a finite state Markov decision process (MDP) — a tuple $M = (S, A, T, R, \gamma)$, where S is a finite set of states; A is a set of actions; $T = P(s'|s, a), s \in S, s' \in S, a \in A$ with $P(s'|s, a)$ is the probability of transitioning to state s' upon taking action a in state s; R specifies the reward distribution; and γ is the temporal discount factor for future rewards.

The goal of the agent is to choose a policy that will maximize its future cumulative rewards, typically the expected discounted sum over an infinite horizon (Sutton & Barto, 1998). Formally, this corresponds to maximizing

$$\sum_{t=0}^{\infty} \gamma^t R_{a_t}(s_t, s_{t+1}) \qquad (1)$$

where action a_t is chosen according to some optimal policy, $R_{a_t}(s_t, s_{t+1})$ is the immediate expected reward received after taking action a_t from state s_t to state s_{t+1}, and $\gamma \in [0, 1]$ is the discount factor.

2.2 Intrinsically motivated reinforcement learning

In the intrinsically-motivated reinforcement learning (IMRL) framework developed by Singh et al. (2009) and Singh et al. (2010a), the rewards that shape behavior are provided by a critic that takes as input the agent's internal representation of its history (see Figure 1 in Singh et al. (2010a)). Their approach is based on searching a space of reward functions for the one that maximizes the expected evolutionary fitness of an RL agent that uses it. Formally, the IMRL framework extends the standard RL approach, as follows:

1. Let E denote some set of environments where we want the agent to perform well (Singh et al., 2009).
2. Let $h_t = s_1 a_1, r_1, \ldots, s_t, a_t, r_t \in H$ denote a particular history of interaction of the agent with the environment up to time-step t, taken from a set H of possible finite histories. The definitions of the variables s, a and r are the same as before.
3. The designer's objectives are encoded in a fitness-based reward function, denoted by $r^F : S \times A \times H \to \mathcal{R}$, which, if used by the agent, prescribes preferences over its behavior during some history of interaction (Bratman, Singh, Sorg, & Lewis, 2012).
4. Let R be some broader space of possible reward functions, which includes, but is not limited to, the fitness component r^F.
5. The agent's objective is then represented by a primary reward function $r : S \times A \times H \to \mathcal{R} \in R$. The agent's behavior under such a reward function takes into account the fitness objective, but is not constrained to follow it at all times. Rather, the agent acts to maximize the cumulative reward as defined by r, which guides its behavior while learning.
6. Finally, $\mathcal{F}(r)$ measures the impact of reward function r on the agent's fitness in the set of environments of interest E.

The optimal reward problem (ORP) is then introduced as an optimization task: choosing an optimal reward function, denoted by $r^* \in R$. ORP maximizes the expected fitness or objective return with respect to a distribution over possible environments E (Bratman et al., 2012; Singh et al., 2009, 2010a). The optimal reward function r^* is defined as

$$r^* = \underset{r \in R}{\operatorname{argmax}} \, \mathcal{F}(r) \qquad (2)$$

2.3 Modeling intrinsic reward with emotions

As in any search-based approach, the success of the above IMRL framework depends on whether or not the space R of reward functions that is being searched does in fact contain good candidate functions. Given the central role of emotions in regulating animal

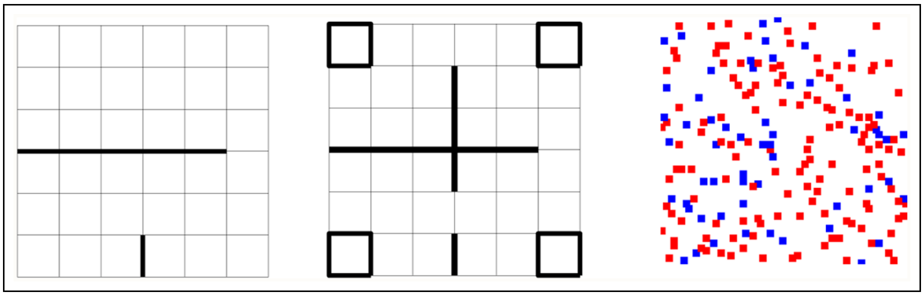

Figure 1. The three types of environments used in our experiments. *Left*: the map for Experiment 1, which involved exploration only (no food resources). Thick black lines are impassable walls. *Middle*: the map for Experiments 2 and 3, in which food could be found in the four boxes situated in the corners. *Right*: in Experiment 4, the 100 × 100 map contained small (3 × 3) food patches, shown in red, and also some poison patches, shown in blue.

behavior, it makes sense to incorporate emotion-based components into the reward functions that populate the space R.

Emotion-based intrinsic reward has been studied within the RL framework (Ahn & Picard, 2006; Sequeira, Melo, & Paiva, 2011). One possible approach here is to define the agent's affective state in terms of a statistical analysis of the outcomes it attains and to associate positive affect with exploitation and negative affect with exploration (Broekens, 2007). Another model (Salichs & Malfaz, 2006) utilized three basic emotions, fear, happiness, and sadness, to control the behavioral strategy of the agent. Happiness and sadness were determined by the outcomes, while fear corresponded to the agent's aversion to low-valued actions. Happiness was also part of the RL framework proposed by Daswani and Leike (2015),who defined it as the sum of payout and good news.

In the IMRL framework, Sequeira et al. (2011, 2014) made important contributions to the incorporation of affect into intrinsic reward. Their work took inspiration from appraisal theory, including the dimensions of novelty, goal relevance, degree of control, and expected valence. Sequeira et al. (2011, 2014) showed that assigning different weights to each dimension led to learning that was more effective in terms of fitness than traditional RL.

In this paper, we build on the insights of Singh et al. (2010a) and Sequeira et al. (2014) in putting forward an IMRL method that searches a space of reward functions that involve emotion-like variables. Critically, some of these reward functions include the components of happiness — hedonic and eudaimonic well-being — that are motivated by psychological considerations and are provably effective in an evolutionary setting (Gao & Edelman, 2016).

2.4 *The components of happiness and their role in shaping behavior*

In the psychological literature, there is a standard distinction between two major components of happiness, each operationalized as subjective well-being: *hedonic*, expected through responses to questions such as "How happy are you right now?", and *eudaimonic*, based on responses to questions such as "How happy are you with your life in general?" (Henderson & Knight, 2012; Rutledge, Skandali, Dayan, & Dolan, 2014). In earlier work (Gao & Edelman, 2016), we investigated several possible formulations for happiness that involve these two aspects of well-being, as well as their role in evolutionary success of agents in a variety of simulated environments.

We found that (i) the effects of attaching more weight to eudaimonic (longer-term) than to momentary (hedonic) happiness and of extending the memory for past happiness are both stronger in an environment where food is scarce; (ii) "relative consumption", in which the agent's well-being is diminished by that of its neighbors, is more detrimental to survival when food is scarce; and (iii) agents with a positive outlook, whose longer-term happiness gets increased more from positive events than decreased from negative ones, is generally advantageous. In the present paper, we use those findings as a conceptual foundation for designing IMRL agents.

3 The components of happiness as determinants of reward

In this section, we introduce a set of reward features that are designed to integrate hedonic and eudaimonic well-being factors into the IMRL framework. Specifically, we formulate the reward function $r \in R$ as a linear combination of hedonic and eudaimonic well-being functions, $\mathcal{H}$ and $\mathcal{E}$, each of which maps history to a scalar value

$$r(h_t) = \theta_{he}\mathcal{H}(h_t) + (1 - \theta_{he})\mathcal{E}(h_t) \tag{3}$$

where the hedonic/eudaimonic balance parameter θ_{he} controls the relative contribution of each of these two types of well-being to the intrinsic reward.

As appropriate for an MDP setting, our RL reward features involve statistical "summaries" of the agent's history of interaction with the environment. Formally, let H denote the set of all possible finite histories that the agent can experience throughout its lifetime. In particular, we consider an element $h \in H$ as a sequence $h_t = \{s_1, a_1, n_1, \ldots, s_t, a_t, n_t\}$, where s_t, a_t and n_t denote, respectively, the agent's state at time-step t, the agent's action at time t, and the affectively important outcomes achieved from the agent environment interaction at time t. For instance, the affectively important outcomes could include obtaining water, food, friendship, etc (Ahn & Picard, 2006; Salichs & Malfaz, 2012; Sequeira et al., 2011). The internal critic of the agent is responsible for processing the agent's perceptions into the history h_t.

The hedonic well-being $\mathcal{H}$ consists of three factors: $\mathcal{H}_s$ based on agent's current state s; $\mathcal{H}_a$ based on agent's action a and its current state s; and $\mathcal{H}_n$ based on affectively important outcomes achieved during the agent's interaction with the environment

$$\mathcal{H} = \theta_s\mathcal{H}_s + \theta_a\mathcal{H}_a + (1 - \theta_s - \theta_a)\mathcal{H}_n \tag{4}$$

The observation-based component is computed as follows

$$\mathcal{H}^t_s(s_t, h_t) = \left(\frac{1}{\lambda_s}\right)^{-c(s_t, h_t, \Delta t)} \tag{5}$$

where λ_s is a positive constant in the range $(0, 1]$ and Δt is the duration of the agent's memory window. The novelty function $c(s_t, h_t, \Delta t)$ tracks the number of time steps since the agent previously visited state s_t during history h_t in the past Δt time window. An inverse-frequency account of novelty has been utilized in other models (Bratman et al., 2012; Singh, Lewis, Barto, & Sorg, 2010b; Sorg et al., 2010). Here, we chose to only consider the recent history of duration Δt instead of the entire past interaction history.

The action decision-based hedonic component is computed as follows

$$\mathcal{H}_a^t(a_t, s_t, h_t) = \left(\frac{1}{\lambda_a}\right)^{-c(a_t, s_t, h_t, \Delta t)} \tag{6}$$

where the λ_a and Δt parameters are as in equation (5) above and the novelty function $c(a_t, s_t, h_t, \Delta t)$ is the number of time steps since the agent previously executed action a_t while in state s_t during history h_t in the past Δt time window.

The hedonic component associated with extrinsic outcomes is computed as follows

$$\mathcal{H}_n^t = n_t \tag{7}$$

where n_t represents the affectively important outcomes obtained from the agent-environment interaction at time t. In our experiments that involve foraging, we used n_t to represent the agent's encounters with food or poison.

The agent's eudaimonic well-being $\mathcal{E}$ is then computed from its present value of $\mathcal{H}$, its memory of past values of $\mathcal{H}$ extending over a number of cycles, and the rates of rise and fall of $\mathcal{H}$. The present formulation of eudaimonic well-being is based on our previous work (Gao & Edelman, 2016). Specifically

$$\mathcal{E}^t = (1 - \theta_p - \theta_n)\left(\mathcal{H}^t - \underbrace{\frac{1}{\Delta t}\sum_{i=t-\Delta t}^{t-1}\mathcal{H}^i}_{\text{expectation of hedonic}}\right) + \sum_{i=t-\Delta t}^{t}\left(s_p\left(\frac{d\mathcal{H}^i}{di}\right) + s_n\left(\frac{d\mathcal{H}^i}{di}\right)\right)$$
$$s_p(x) = \begin{cases} \theta_p \cdot x & x \geq 0 \\ 0 & x<0 \end{cases} \qquad s_n(x) = \begin{cases} 0 & x \geq 0 \\ \theta_n \cdot x & x<0 \end{cases} \tag{8}$$

where Δt is the extent of the memory window. The step function s_p selects the weight assigned to upswings of H and s_n, and to downswings; θ_p and θ_n are the respective weights (Gao & Edelman, 2016). Thus, agents can in principle value positive and negative events differently. The intuition behind this formulation is that the agent may have a subjective expectation based on the past interaction with the environment. Its eudaimonic well-being $\mathcal{E}$ is based on its current hedonic well-being and its expectation, as well as on its evaluation of the rise and fall of hedonic well-being. The three terms could be weighted differently.

3.1 The learning algorithm

Because the goal of the present study is to evaluate the happiness-based approach to intrinsic reward, we chose to combine the candidate reward functions with a standard IMRL algorithm. One such algorithm is ϵ-greedy Q-learning (Singh et al., 2010b; Sutton & Barto, 1998). In this algorithm, the Q-table stores the learned values for state–action pairs, α is the learning rate, and γ is the temporal discounting factor. The parameter ϵ controls the balance between exploitation and exploration: at each time step, the agent executes a random action with probability ϵ and the greedy action with respect to the current Q-table with probability $(1 - \epsilon)$. The Q-table is updated as follows: $Q_{t+1}(s_t, a_t) = (1 - \alpha)Q_t(s_t, a_t) + \alpha[r_t + \gamma \max_{b \in A}(Q_t(s_{t+1}, b))]$, where r_t is the reward at time step t.

We compared several formulations for intrinsic reward, with the following design considerations in mind:

- a balance should be maintained between exploration and exploitation (not necessarily by means of an explicit probabilistic choice as in the ϵ-greedy algorithm);
- each action decision should consider not only the agent's current knowledge of the environment, but also its expected well-being gain;
- the difference between actual and expected well-being should be considered when updating the agent's knowledge of the environment.

The result, Algorithm 1, is intended to be used in conjunction with a variety of reward functions, including notably intrinsic happiness-based ones, listed in section 'The different types of agents'. At each step, the agent selects its next action a_t, given its past history h_t. If the exploration parameter ϵ is nonzero, there is a finite probability of taking a random action; otherwise the action a_t is chosen that maximizes the sum of the expected reward $E(r(h_{t+1}))$ and the value associated with new state s_{t+1}. After taking action a_t and extending the history h_{t+1}, the agent updates its Q-table entry for state s_t and action a_t, taking into account the difference between the actual reward $r(h_{t+1})$ and expected reward $E(r(h_{t+1}))$.

3.2 The different types of agents

In our experiments, we compared the performance of several types of agents, each defined by a combination of a learning algorithm and a reward function, as

Algorithm 1 The ϵ-greedy Q-learning algorithm chooses a random action (an exploration step) with probability $0 \leqslant \epsilon \leqslant 1$ and a greedy action (an exploitation step, which maximizes the expected reward) with probability $1 - \epsilon$. When ϵ is set to 0, the choice is always greedy; the balance between exploration and exploitation in that case depends on the structure of the reward function.

```
Initialize environment $E$ and matrix $Q$
for $t = 1$ to NumTimeSteps do
    if rand(0, 1)$<\epsilon$ then
        Randomly select and carry out action $a_t$
    else
        Carry out the expected best action
        $a_t = \text{argmax}_{a \in A} \, (E(r(h_{t+1})) + Q_t(s_{t+1}, a_t))$
    end if
    Observe reward $r(h_{t+1})$ and state $s_{t+1}$
    Update $Q$-table:
    $Q_{t+1}(s_t, a_t) = (1 - \alpha)Q_t(s_t, a_t) + \alpha(r(h_{t+1}) - E(r(h_{t+1}))$
                $+ \gamma \max_{b \in A} (Q_t(s_{t+1}, b))$
end for
```

detailed below. For each agent type, the optimal reward function $r^* \in R$ is found (Singh et al., 2010b) via approximately exhaustive search of the parameter space, which defines the space R of possible reward functions. To make this search tractable, we appropriately discretize the parameter Δt in the range $[1, 1000]$, and the parameters $(\theta_{he}, \theta_s, \theta_a, \Delta t, \theta_p, \theta_n)$, as well as ϵ and α.

Qh **Q-learning with hedonic reward.** This type of agent uses the greedy version ($\epsilon = 0$) of Q-learning (Algorithm 1) with a hedonic-only happiness-based reward (equation (4)). To enforce hedonic-only reward, θ_{he} is set to 1. The other parameters, $(\theta_s, \theta_a, \Delta t, \alpha)$, are discretized; Δt is in the range $[1, 1000]$ and $(\theta_s, \theta_a, \alpha)$ are in the range $[0, 1]$.

Qe **Q-learning with eudaimonic reward.** This type of agent uses the greedy version of Q-learning (Algorithm 1) with an eudaimonic-only happiness-based reward (equation (4)). To enforce eudaimonic-only reward, θ_{he} is set to 0. The other parameters, $(\theta_s, \theta_a, \Delta t, \theta_p, \theta_n, \alpha)$, are discretized; Δt is in the range $[1, 1000]$ and $(\theta_s, \theta_a, \theta_p, \theta_n, \alpha)$ are in the range $[0, 1]$.

Qc **Q-learning with combined well-being reward.** This type of agent uses the greedy version of Q-learning (Algorithm 1) with a combined happiness-based reward (equation (4)). To enforce combined well-being reward that includes both hedonic and eudaimonic components, θ_{he} is set to lie in the range $[0.1, 0.9]$. The other parameters, $(\theta_s, \theta_a, \Delta t, \theta_p, \theta_n, \alpha)$, are discretized; Δt is in the range $[1, 1000]$ and $(\theta_s, \theta_a, \theta_p, \theta_n, \alpha)$ are in the range of $[0, 1]$.

Qwb **Q-learning with well-being reward.** This type of agent is a union of the above three types: it uses the greedy version of Q-learning (Algorithm 1) with a happiness-based reward (equation (4)) in which the parameter θ_{he} takes values in the full range $[0, 1]$.

ϵQwb **ϵ-greedy Q-learning with well-being reward.** As with the previous type, this agent uses the full range of happiness-based rewards (equation (4)), with the parameter θ_{he} taking values in $[0, 1]$. However, instead of greedy it uses the ϵ-greedy version of Q-learning (Algorithm 1). The parameters $(\theta_s, \theta_a, \Delta t, \theta_p, \theta_n, \epsilon, \alpha)$ are discretized. Δt is in the range $[1, 1000]$, ϵ is in the range $[0.1, 0.99]$. The rest of the parameters are in the range $[0, 1]$.

ϵQr **ϵ-greedy Q-learning with random reward.** This type of agent, whose performance serves as a simple baseline for comparisons, uses ϵ-greedy Q-learning in conjunction with rewards that are chosen at random from the interval $[0, 1]$ at every time step. The parameters of ϵ in the range $[0.1, 0.99]$ and α in the range $[0, 1]$ are discretized.

ϵQf **ϵ-greedy Q-learning with fitness-based reward.** This type of agent uses ϵ-greedy Q learning with a reward defined exclusively in terms of fitness increments or decrements. A food item carries a reward of 10, a poison item -10. The ϵ parameter in the range $[0.1, 0.99]$ and α in the range $[0, 1]$ are discretized.

We note that while the range of the hedonic-eudaimonic balance parameter θ_{he} for agent type Qwb is the union of the ranges for types Qh, Qe, and Qc (that is, the full $[0, 1]$ interval), the optimal combination of the other parameter values for Qwb does not necessarily coincide with any of the optima for those three types (for which θ_{he} is forced to take on either of the extreme values of 0 or 1, or a strictly intermediate value). The reason for this is the complex interaction between the hedonic and eudaimonic components.

3.3 The three types of environments

The computational experiments described below in sections 'Experiment 1' through 'Experiment 4' compare the performance of agents in three types of environments, illustrated in Figure 1. In the first of these, the task is simply exploration, as it contains no resources to be discovered. In the second environment, the task is foraging for food in a relatively small space that includes obstacles. In the third environment, the space is much larger and may contain both food and poison.

In each case, the partial observability of the state space puts agents that rely too much on fitness in

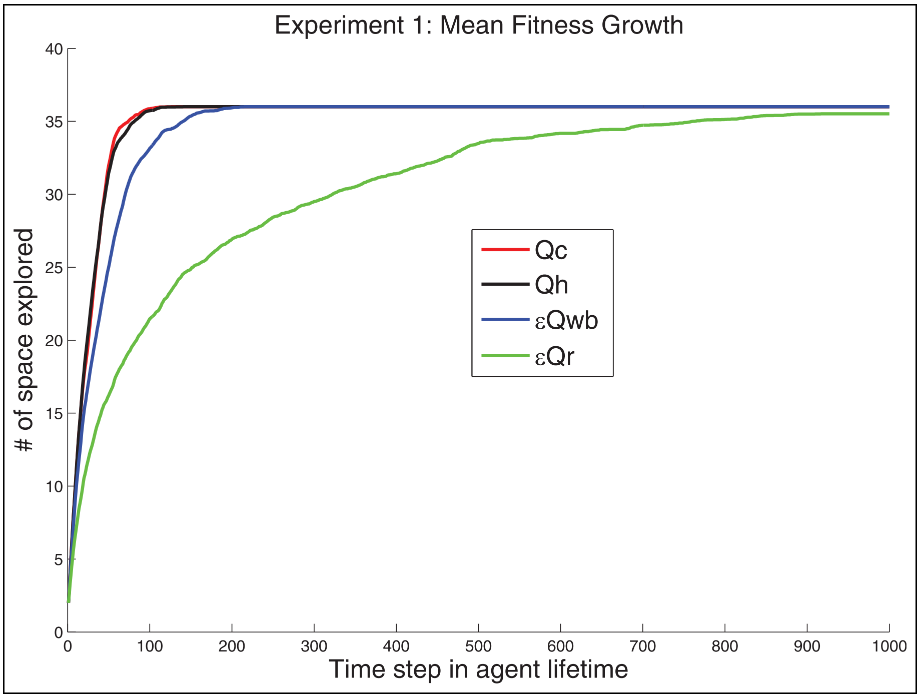

Figure 2. Experiment 1: mean cumulative fitness (number of squares explored) versus time, for four types of agents: Qc (*Q*-learning with combined well-being reward); Qh (*Q*-learning with hedonic reward); ϵQwb (ϵ-greedy *Q*-learning with well-being reward); and ϵQr (ϵ-greedy *Q*-learning with random reward); see section 'The different types of agents' for details. Agents of type Qc were the fastest in attaining the maximum fitness. The maximum possible value of cumulative fitness in this task was 36. Each curve is an average of 100 runs.

calculating their rewards at a disadvantage, compared to agents that are motivated also by less immediate factors (Singh et al., 2010b).

4. Experiment 1

Our first goal was to test the newly introduced types of intrinsic reward function in a pure exploration task, in the same setting as investigated in recent IMRL studies, notably Singh et al. (2010b).

4.1 Environment, fitness, and reward

In Experiment 1, we used a simulated physical space of size 6 × 6, illustrated in Figure 1, left (this is the similar map used by Singh et al. (2010b)). In addition to the external boundaries, this environment contains two internal barriers; the actions it affords are movements in the four cardinal directions, where available. An action fails if the agent attempts to move against an outer boundary of the grid or a barrier. Agents are initially placed at random somewhere in the right half of the map. The agent's goal in this experiment is exploration, not foraging, and the environment contains no food.

With the goal being exploration, the agent's (inverse) fitness is defined as the time it takes to visit all the 36 squares on the map. The fitness objective, then, is to minimize the number of action steps to complete the exploration.

4.2 Results

Figure 2 plots the mean cumulative fitness (number of squares explored) as a function of time for four types of agents: Qc (*Q*-learning with combined well-being reward); Qh (*Q*-learning with hedonic reward); ϵQwb (ϵ-greedy *Q*-learning with well-being reward); and ϵQr (ϵ-greedy *Q*-learning with random reward); see section 'The different types of agents' for details. Agents of type Qc were the fastest in attaining the maximum fitness of 36, indicating that combining *Q*-learning with a reward function based on balanced hedonic and eudaimonic well-being is the most effective approach in this case.

The performance of agents of types Qh, Qe, and Qc, along with the optimal reward function parameter values ($\theta_{he}, \theta_s, \theta_a, \Delta t, \theta_p, \theta_n, \alpha$), are listed in Table 1, in the rows labeled "exp 1". Agents of type Qc, which use the combined well-being reward function, perform best with the hedonic/eudaimonic balance parameter $\theta_{he} = 0.3$, which corresponds to a slight preference for eudaimonic well-being. Table 2 shows that agent type Qwb performs the best.

Table 1. The optimal values of parameters of happiness-based intrinsic reward functions for the four experiments. The right column shows the mean cumulative fitness over 100 runs, along with the 95% confidence intervals.

Exp	Type	θ^* =	[θ_{he},	θ_s,	θ_a,	Δt,	θ_p,	θ_n,	α]	Fitness
exp 1	Qh	θ^* =	[1,	0.5,	0.4,	100,	0,	0,	0.6]	71.15 ± 4
	Qe	θ^* =	[0,	0.2,	0.4,	125,	0,	0.2,	0.5]	67.66 ± 4
	Qc	θ^* =	[0.3,	0.6,	0.4,	50,	0.8,	0,	0.3]	67.25 ± 3
exp 2	Qh	θ^* =	[1	0.3,	0.1,	200,	0,	0,	0.7]	786.67 ± 5
	Qe	θ^* =	[0,	0.3,	0.5,	500,	0.5,	0,	0.9]	712.75 ± 7
	Qc	θ^* =	[0.9,	0.2,	0.2,	200,	0.6,	0,	0.9]	794.99 ± 8
exp 3	Qh	θ^* =	[1	0.1,	0.5,	200,	0,	0,	0.4]	90.66 ± 1
	Qe	θ^* =	[0,	0.3,	0,	100,	0.6,	0.2,	0.3]	103.50 ± 2
	Qc	θ^* =	[0.1,	0.3,	0.3,	100,	0.6,	0.3,	0.9]	130.93 ± 3
exp 4	Qh	θ^* =	[1,	0.4,	0.2,	500,	0,	0,	0.6]	212.93 ± 53
	Qe	θ^* =	[0,	0.5,	0.3,	800,	0.3,	0.1,	0.8]	515.52 ± 81
	Qc	θ^* =	[0.3,	0.7,	0.2,	1000,	0.3,	0.3,	0.6]	818.54 ± 52

Table 2 Fitness outcomes for different experiments and agent types (see section 'The different types of agents' for the abbreviations). The cumulative fitness values means and 95% confidence intervals over 100 runs.

Scenario	Qwb	ϵQwb	Qf	Qr
exp 1	67.25 ± 3	109.37 ± 7	~	504.73 ± 49
exp 2	794.99 ± 8	667.63 ± 10	718.19 ± 8	28.25 ± 1
exp 3	130.93 ± 3	74.53 ± 2	90.26 ± 2	16.12 ± 1
exp 4	818.54 ± 52	700.82 ± 93	222.90 ± 46	159.65 ± 32

5 Experiment 2

Next, we tested the performance of the different reward function (agent) types in a foraging task, which involved both the exploration of the environment and the search for food.

5.1 Environment, fitness, and reward

The environment in Experiment 2, illustrated in the middle panel of Figure 1, is similar to that used by Singh et al. (2010b). In addition to movements in the four cardinal directions (unless obstructed by a wall), two box-related actions can be taken when the agent is next to a box: Open and Eat. In each environment, there are two boxes located at either the diagonal corners or at the top left and bottom left corners. The box locations cannot be consecutive corners to each other unless they are at the top left and bottom left corners. When the agent opens a box, it becomes half-open, at which point, the agent can consume the food inside. The box then becomes open. When the box is open, it has a 0.1 probability of becoming closed. If the box is half open and the agent does not eat the food immediately, the food disappears and the box becomes an open box containing no food. The agent is initially located at random in one of the two empty corners.

When the agent consumes a food item, its fitness is incremented by one. The fitness objective is therefore to maximize the cumulative amount of food. The fitness-based reward function reflects this objective. The intrinsic reward functions based on hedonic and eudaimonic well-being are related to food indirectly (see equation (7)).

5.2 Results

Figure 3 plots the mean cumulative fitness as a function of time, averaged over 100 runs. In this environment, the probability of an open box becoming closed and replenished is 0.1. Therefore, the expected fitness after 10,000 time steps is 1000. The plot shows the performance of agents of types Qc, Qh, and Qf, as well as ϵQwb and ϵQr. In the long run, agents of type Qc perform the best, followed closely by type Qh.

Figure 4 shows how much time each of three types of agents, Qh, Qc, and Qf, spent in each location on the map. The food boxes in this environment are in the top and bottom left corners. When agents of type Qh or Qc find a box, they stay in the vicinity of that box for the remainder of their lifetime. Agents of type Qf, which are rewarded directly by fitness, take much longer to find a box; once they do, they also stay in the same region and keep opening the box over and over.

The optimal parameter values for the different types of agents are presented in Table 1 in the rows labeled "exp 2". For type Qc (combined well-being reward), the optimal value of the hedonic/eudaimonic balance parameter is $\theta_{he} = 0.9$, indicating a dominance of the hedonic component. In addition, $\theta_s = 0.2$ and $\theta_a = 0.2$

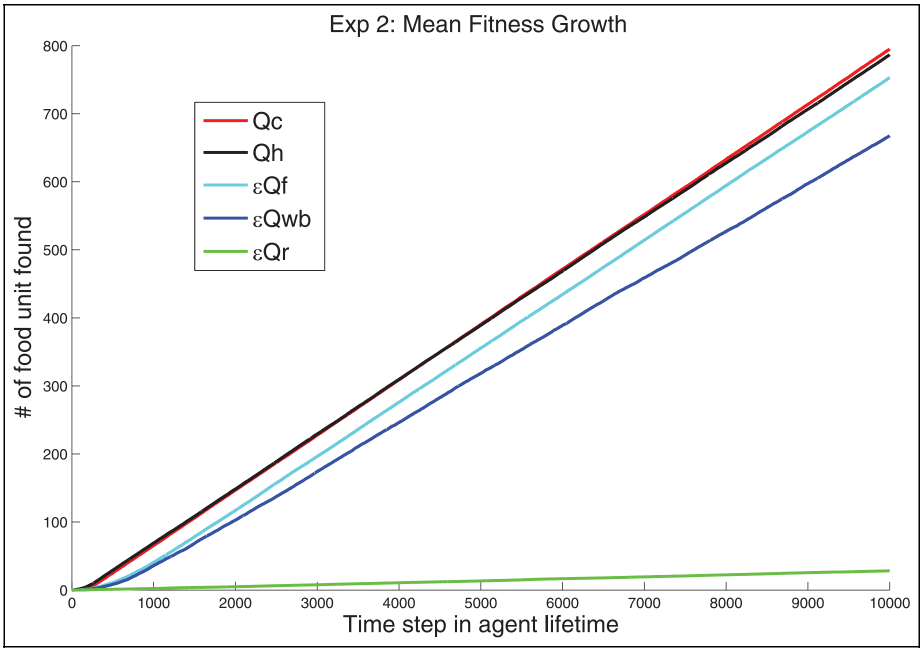

Figure 3. Experiment 2: mean cumulative fitness versus time, for five types of agents: Qc, Qh, Qf, ϵQwb, and ϵQr (see section 'The different types of agents' for details). Agents of type Qc (*Q*-learning with combined well-being reward) performed the best. In addition, in the first 1000 time steps, Qc and Qh have a higher learning rate than the other agent-types. Each curve is an average of 100 runs.

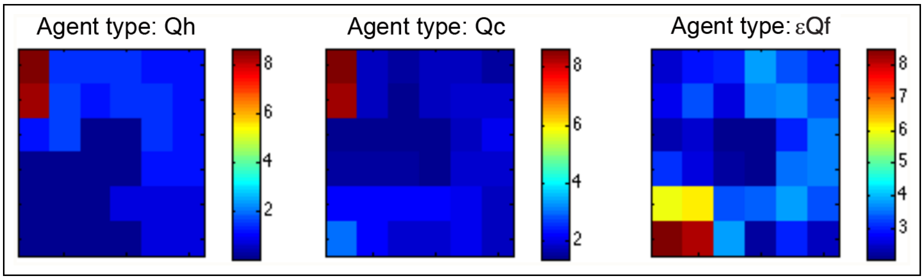

Figure 4. Experiment 2: a log frequency plot that shows the number of times each location has been visited by the agent in its lifetime. The food boxes in this environment are located in the upper and lower left corners. The three types of agents, from left to right, are: Qh, Qc, and Qf. See section 'Results' for discussion.

suggest that the best-performing agents are those that value food more than exploration. Table 2 offers more comparisons, showing that agents of type Qwb outperform others. The difference in cumulative fitness between Qwb and Qf is, however, not significant. We believe this is due to the environment being both small and relatively predictable in terms of the probability of food encounters. In the next experiment, we show that decreasing the latter results in a different ranking of agent types.

6 Experiment 3

In this experiment, we aimed to demonstrate that, unlike fitness-based reward functions that typically require hand-tuning, the happiness-based intrinsic reward formulation is flexible enough to perform well in more challenging environments than the previous one.

6.1 Environment, fitness, and reward

The same environment was used as in Experiment 2 (Figure 1, middle). The only difference was in the probability of an open box becoming closed and replenished, which was reduced from 0.1 to 0.01. With this much lower probability, the strategy of staying close to the same box is no longer optimal, the expected reward for doing this being 100 over 10,000 time steps. A better strategy is to go back and forth from one box to another. As we shall see, this change is reflected in the relative performance of the different agent types.

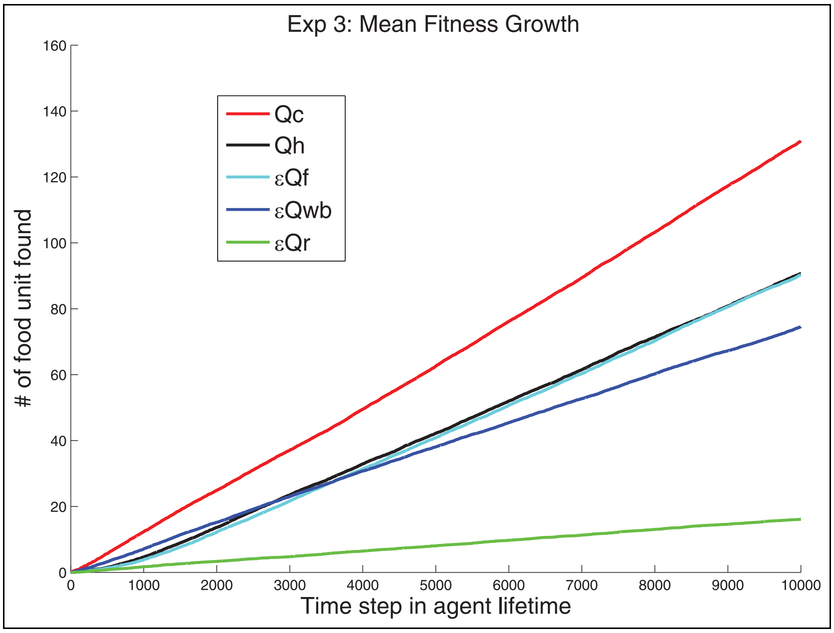

Figure 5. Experiment 3: mean cumulative fitness versus time, for five types of agents: Qc, Qh, Qf, ϵQwb, and ϵQr (see text for details). Agents of type Qc performed the best.

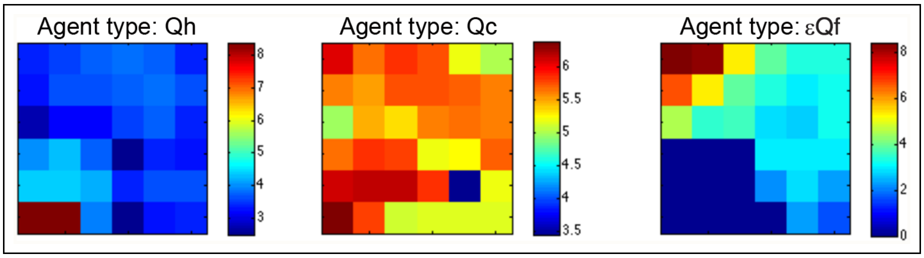

Figure 6. Experiment 3: a log frequency plot that shows the number of times each location has been visited by the agent in it's lifetime. The food boxes in this environment are located in the upper and lower left corners. The three types of agents, from left to right, are: Qh, Qc, and Qf. See text for further details.

6.2 Results

Figure 5 shows the plot of the mean cumulative fitness as a function of time. The five agent types that are compared are: Qc, Qh, Qf, ϵQwb, and ϵQr. Agents of type Qc outperform those with all other reward functions, by quite a large margin, as can be seen in Table 2. The visit frequency map in Figure 6 suggests that Qc agents, whose intrinsic reward function is based on combined well-being, outperform other agent types. This includes fitness-driven Qf, in exploration, which in this environment is the better policy, because the cost of traveling between two boxes is much smaller than waiting near the same box until it closes and replenishes. Relying on immediate (hedonic) well-being is suboptimal: Qh agents find the bottom left box and stay near it for the remainder of their lifetimes.

The optimal parameter values for the hedonic, eudaimonic, and combined well-being intrinsic reward functions are presented in Table 1, row "exp 3". The mean cumulative fitness values and their 95% confidence intervals are computed over 100 runs. The optimal value of the hedonic/eudaimonic balance parameter θ_{he} is 0.1, which indicates heavy reliance on eudaimonic well-being.

7 Experiment 4

Unlike most of the simulation studies of IMRL, which involve environment maps not larger than 10×10 (Sequeira et al., 2011, 2014; Singh et al., 2009, 2010b), our next experiment is situated on a 100×100 grid, allowing us to investigate the scaling properties of the IMRL formulations under consideration. The foraging problem in Experiment 4 is also more realistic in that it involves both positive and negative rewards, which is the first step towards accounting for approach versus

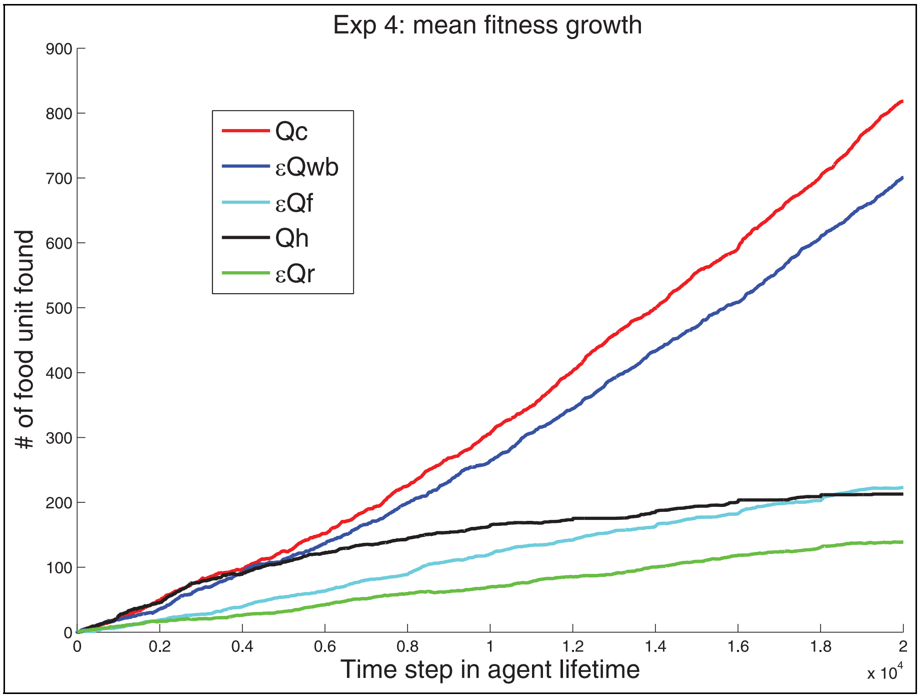

Figure 7 Experiment 4: mean cumulative fitness versus time, for five agent types: Qc, ϵQwb, Qf, Qh, and ϵQr (see text for details). Agents of type Qc outperform others in learning to accumulate food while avoiding poison.

avoidance behaviors (Eiser, Fazio, Stafford, & Prescott, 2003).

7.1 *Environment, fitness, and reward*

The simulated 100 × 100 environment used in this experiment is illustrated in Figure 1, right. It contains randomly located 3 × 3 food patches (red), along with less common 3 × 3 poison patches (blue; the proportion of poison is 30%). Agents are initially placed randomly on the map.

The lifetime of an agent is 20,000 steps; after every 2000 steps, the environment is replenished with both food and poison that has been consumed. When an agent consumes a food item, it's fitness increases by 1; a poison item decreases it by 1. In this environment, therefore, the optimal strategy is to cover as much space as possible, while learning where food and poison are found.

7.2 *Results*

Figure 7 shows the mean cumulative fitness as a function of time, expected over 100 runs. Among the five agent types it compares (Qc, ϵQwb, Qf, Qh, and ϵQr), agents of type Qc outperform others in learning to accumulate food while avoiding poison.

The learned optimal parameters for the hedonic, eudaimonic, and combined well-being intrinsic reward functions are presented in Table 1. One aspect to be noted is that for the optimal reward weight in combined well-being reward, the weights of positive and negative events are equal: $\theta_p = 0.3$, $\theta_n = 0.3$. Thus, unlike in the previous experiments, in Experiment 4 the decrease in hedonic well-being (for example, by an encounter with poison) is as important in shaping behavior as the increase (see Gao and Edelman (2016), experiment 5). A comparison of performance in terms of mean cumulative fitness appears in Table 2. We note that agents with fitness-based reward (Qf) did not achieve high cumulative fitness, presumably due to the difficulty in finding the balance between exploration and exploitation.

8 Discussion

This paper addressed the problem of designing an intrinsic reward function for IMRL agents by searching for an optimal formulation of reward in a space that contained both fitness-based and, critically, emotion-based functions. We found that the latter–specifically, functions based on intrinsic reward features inspired by the concepts of hedonic and eudaimonic well-being–led to more effective learning. Such reward features can serve to provide flexible intrinsic guidance for decision-making in a range of environments.

A comparison of the mean cumulative fitness attained by agents equipped with the different reward functions, across all four of our experiments (Table 2), indicates a statistically significant advantage of reward functions based on the components of well-being versus those based on fitness. This advantage and the observed agent behaviors are consistent with the findings of Singh et al. (2010b) and Sequeira et al. (2014). Our work differs from those previous approaches in that our agents rely on their intrinsic estimates of well-being to find the balance between exploration and exploitation.

In natural foraging situations, biological agents must maintain such a balance to survive (Kamil & Sargent, 1981). The results of Experiments 3 and 4, in which the best performance was achieved by agents that blended short-term (hedonic) and long-term (eudaimonic) intrinsic rewards, suggest a possible role for the components of happiness in playing off novelty against current achievement — a finding that is also consistent with our earlier results involving evolutionary agent-based simulations (Gao & Edelman, 2016), as well as with a broader set of philosophical and psychological considerations (Edelman, 2012).

We note that our IMRL model made use of the reward prediction error; a theoretical construct that arose from computational RL and that has been intensively studied by neuroscientists who work on the dopamine reward system in the brain (Bayer & Glimcher, 2005; Colombo, 2014; Glimcher, 2011). Thus, unlike the traditional ϵ-greedy Q-learning method used by many other IMRL models (Sequeira et al., 2014; Singh et al., 2009, 2010b), in our learning algorithm the Q-table updates depend in part on the difference between actual and predicted rewards.

In RL, the balance between exploration and exploitation is a very important problem. Usually, a random action is selected with probability ϵ to assure that the agent explores the environment. Our current work is a greedy policy (without random actions), which has the best results because exploration is directly reinforced through rewards, since the novelty of states and of actions is part of the "hedonic" and "eudamonic" reward. For this reason, an ϵ-greedy policy is not necessary and the results with and without ϵ are shown in Table 2. Furthermore, from the intrinsic reward perspective, the decision to choose explore versus exploit should come from internally. Therefore, it is natural to model the balance as part of the agent's choice rather than a parameter in the algorithm.

Possible directions for future work include extending our model to multi-agent settings and embedding it in a multi-generational evolutionary context (see Gao and Edelman (2016), experiment 3). In particular, it would be interesting to examine the social and evolutionary dynamics of learning by IMRL agents that are motivated by hedonic and eudaimonic factors. Such an investigation could focus on designing domain-independent reward features that assess the social acceptability of behaviors to achieve cooperation among learning agents. It should also be interesting to address the relative effectiveness and possible trade-off between social sharing and individualistic self-motivated exploration.

Funding

This research received no specific grant from any funding agency in the public, commercial, or not-for-profit sectors.

References

Ahn, H., & Picard, R. (2006). Affective cognitive learning and decision making: The role of emotions. In *Proceedings of the 18th European meeting on cybernetics and systems research*. Vienna, Austria: Austrian Society for Cybernetic Studies.

Akaishi, R., Kolling, N., Brown, J., & Rushworth, M. (2016). Neural mechanisms of credit assignment in a multicue environment. *The Journal of Neuroscience*, *36*(4), 1096–1112.

Baldassarre, G., & Mirolli, M. (2013). *Intrinsically motivated learning in natural and artificial systems*. Berlin: Springer.

Barto, A. (1995). Adaptive critics and the basal ganglia. In J. C. Houk, J. Davis & D. Beiser (Eds.), *Models of information processing in the basal ganglia* (pp. 215–232). Cambridge, MA: MIT Press.

Barto, A. (2013). Intrinsic motivation and reinforcement learning. In G. Baldassarre, & M. Mirolli (Eds.), *Intrinsically motivated learning in natural and artificial systems* (pp. 16–47). Berlin: Springer.

Barto, A., & Simsek, O. (2005). Intrinsic motivation for reinforcement learning systems. In *Proceedings of the thirteenth yale workshop on adaptive and learning systems* (pp. 113–118). New Haven, CT: Yale University.

Bayer, H., & Glimcher, P. (2005). Midbrain dopamine neurons encode a quantitative reward prediction error signal. *Neuron*, *47*, 129–141.

Berridge, K. (2003). Pleasures of the brain. *Brain and Cognition*, *52*, 106–128.

Bratman, J., Singh, S., Sorg, J., & Lewis, R. (2012). Strong mitigation: Nesting search for good policies within search for good reward. In *Proceedings of the 11th international conference on autonomous agents and multiagent systems-volume 1* (pp. 407–414). Valencia, Spain: International Foundation for Autonomous Agents and Multiagent Systems.

Broekens, D. (2007). *Affect and learning: A computational analysis*. Leiden: Faculty of Science, LIACS, Leiden University.

Cohen, J., & Blum, K. (2002). Reward and decision. *Neuron*, *36*, 193–198.

Colombo, M. (2014). Deep and beautiful. The reward prediction error hypothesis of dopamine. *Studies in History and Philosophy of Science Part C: Studies in History and Philosophy of Biological and Biomedical Sciences*, *45*, 57–67.

Daswani, M., & Leike, J. (2015). A definition of happiness for reinforcement learning agents. In *Artificial general intelligence* (pp. 231–240). Berlin, Germany: Springer.

Doll, B. B., Simon, D. A., & Daw, N. D. (2012). The ubiquity of model-based reinforcement learning. *Current Opinion in Neurobiology*, *22*, 1075-1081.

Edelman, S. (2012). *The happiness of pursuit*. New York, NY: Basic Books.

Eiser, J., Fazio, R., Stafford, T., & Prescott, T. (2003). Connectionist simulation of attitude learning: Asymmetries in the acquisition of positive and negative evaluations. *Personality and Social Psychology Bulletin*, *29*, 1221–1235.

Feijo, R., Cornell, J., & Garzn, M. (2006). Grasp quality measures. *Recercat Home*, *38*, 65–88.

Gao, Y., & Edelman, S. (2016). Between pleasure and contentment: Evolutionary dynamics of some possible parameters of happiness. *PLoS One*, *11*(5).

Glimcher, P. (2011). Understanding dopamine and reinforcement learning: The dopamine reward prediction error hypothesis. *Proceedings of the National Academy of Sciences*, *108*, 15647–15654.

Gottlieb, J., Oudeyer, P.-Y., Lopes, M., & Baranes, A. (2013). Information-seeking, curiosity, and attention: Computational and neural mechanisms. *Trends in Cognitive Sciences*, *17*, 585–593.

Henderson, L., & Knight, T. (2012). Integrating the hedonic and eudaimonic perspectives to more comprehensively understand wellbeing and pathways to wellbeing. *International Journal of Wellbeing*, *2*, 196–221.

Kakade, S., & Dayan, P. (2002). Dopamine: Generalization and bonuses. *Neural Networks the Official Journal of the International Neural Network Society*, *15*(4-6), 549-559.

Kamil, A. C., & Sargent, T. D. (1981). *Foraging behavior: Ecological, ethological, and psychological approaches*. New York: Garland.

Kaplan, F., & Oudeyer, P. (2004). Maximizing learning progress: An internal reward system for development. In F. Iida, R. Pfeifer, L. Steels, & Y. Kuniyoshi (Eds.), *Embodied artificial intelligence* (pp. 259–270). Berlin: Springer.

Knutson, B., & Gibbs, S. (2007). Linking nucleus accumbens dopamine and blood oxygenation. *Psychopharmacology*, *191*, 813–822.

Lee, D., Seo, H., & Jung, M. W. (2012). Neural basis of reinforcement learning and decision making. *Annual Review of Neuroscience*, *35*, 287–308.

Lindquist, K., Wager, T., Kober, H., Bliss-Moreau, E., & Barrett, L. (2012). The brain basis of emotion: A meta-analytic review. *Behavioral and Brain Sciences*, *35*, 121–143.

Marsella, S., Gratch, J., & Petta, P. (2010). Computational models of emotion. *A Blueprint for Affective Computing-A sourcebook and manual* (pp. 21–46). New York, NY: Oxford University Press.

Mery, F. (2013). Natural variation in learning and memory. *Current Opinion in Neurobiology*, *23*, 52–56.

Minsky, M. (2006). *The emotion machine: Commonsense thinking, artificial intelligence, and the future of the human mind.* New York: Simon & Schuster.

Mirolli, M., Santucci, V., & Baldassarre, G. (2013). Phasic dopamine as a prediction error of intrinsic and extrinsic reinforcements driving both action acquisition and reward maximization: A simulated robotic study. *Neural Networks the Official Journal of the International Neural Network Society, 39C*, 40–51.

Mnih, V., Kavukcuoglu, K., Silver, D., Rusu, A. A., Veness, J., Bellemare, M. G., & . . . Hassabis, D. (2015). Human-level control through deep reinforcement learning. *Nature*, *518*, 529–533.

Nelson, K. (2015). A bio-social-cultural approach to early cognitive development: entering the community of minds. In R. Scott, & S. Kosslyn (Eds.), *Emerging trends in the social and behavioral sciences*. New York: John Wiley & Sons.

Niv, Y. (2009). Reinforcement learning in the brain. *Journal of Mathematical Psychology*, *53*, 139–154.

Oudeyer, P., & Kaplan, F. (2007). What is intrinsic motivation? A typology of computational approaches. *Frontiers in Neurorobotics*, *1*, 6.

Oudeyer, P., Kaplan, F., & Hafner, V. (2007). Intrinsic motivation systems for autonomous mental development. *IEEE Transactions on Evolutionary Computation*, *11*, 265–286.

Pessiglione, M., Seymour, B., Flandin, G., Dolan, R., & Frith, C. (2006). Dopaminedependent prediction errors underpin reward-seeking behaviour in humans. *Nature*, *442*, 1042–1045.

Redgrave, P., Vautrelle, N., & Reynolds, J. N. J. (2011). Functional properties of the basal ganglia's re-entrant loop architecture: Selection and reinforcement. *Neuroscience*, *198*, 138-151.

Rogers, C. R. (1963). Actualizing tendency in relation to m̈otivesänd to consciousness. In M. R. Jones (Ed.), *Nebraska symposium on motivation* (pp. 1–24). Oxford: University of Nebraska Press.

Rumbell, T., Barnden, J., Denham, S., & Wennekers, T. (2012). Emotions in autonomous agents: Comparative analysis of mechanisms and functions. *Autonomous Agents and Multi-Agent Systems*, *25*, 1–45.

Rutledge, R., Skandali, N., Dayan, R., & Dolan, R. (2014). A computational and neural model of momentary subjective well-being. *Proceedings of the National Academy of Sciences*, *111*, 12252–12257.

Salichs, M. A., & Malfaz, M. (2006). Using emotions on autonomous agents. The role of happiness, sadness and fear. *Integrative Approaches to Machine Consciousness, Part of AISB*, *6*, 157–164.

Salichs, M. A., & Malfaz, M. (2012). A new approach to modeling emotions and their use on a decision-making system for artificial agents. *IEEE Transactions on Affective Computing*, *3*, 56–68.

Schembri, M., Mirolli, M., & Baldassarre, G. (2007). Evolving internal reinforcers for an intrinsically motivated reinforcement-learning robot. In *IEEE international conference on development and learning* (pp. 282–287). Imperial College, London: IEEE.

Scherer, K. R., Schorr, A., & Johnstone, T. (2001). *Appraisal processes in emotion*. New York: Oxford University Press.

Schmidhuber, J. (2015). Deep learning in neural networks: An overview. *Neural Networks*, *61*, 85–117.

Sequeira, P., Melo, F., & Paiva, A. (2011). Emotion-based intrinsic motivation for reinforcement learning agents. In *Affective computing and intelligent interaction* (pp. 326–336). Berlin: Springer.

Sequeira, P., Pedro, M., Francisco, S., & Paiva, A. (2014). Learning by appraising: An emotion-based approach to intrinsic reward design. *Adaptive Behavior*, *22*, 330–349.

Silver, D., Huang, A., Maddison, C. J., Guez, A., Sifre, L., van den Driessche, G., & . . . Hassabis, D. (2016). Mastering the game of Go with deep neural networks and tree search. *Nature*, *529*, 484–503.

Singh, S., Lewis, R., & Barto, A. (2009). Where do rewards come from. In *Proceedings of the annual conference of the cognitive science society* (pp. 2601–2606). Amsterdam, NL: Cognitive Science Society.

Singh, S., Lewis, R., Barto, A., & Sorg, J. (2010b). Intrinsically motivated reinforcement learning: An evolutionary perspective. *IEEE Transactions on Autonomous Mental Development*, *2*, 70–82.

Singh, S., Lewis, R. L., Barto, A. G., & Sorg, J. (2010a). Intrinsically motivated reinforcement learning: An evolutionary perspective. *IEEE Transactions on Autonomous Mental Development*, *2*, 70-82.

Sorg, J., Singh, S., & Lewis, R. (2010). Internal rewards mitigate agent boundedness. In *Proceedings of the 27th international conference on machine learning* (pp. 1007–1014). Haifa, Israel: Omnipress.

Sutton, R., & Barto, A. (1998). *Reinforcement learning: An introduction*. Cambridge, MA: MIT press.

Tapia, D. H., Silva, A. X., Ballesteros, G. I., Figueroa, C. C., Niemeyer, H. M., & Ramírez, C. C. (2015). Differences in learning and memory of host plant features between specialist and generalist phytophagous insects. *Animal Behaviour*, *106*, 1–10.

About the Authors

Shimon Edelman BSEE MSc PhD, holds degrees in electrical engineering and in computer science and is interested in all aspects of mind, brain, and behavior. He is presently Professor of Psychology at Cornell University in Ithaca, NY, where he works on behavioral, neural, evolutionary, and computational aspects of vision, language, consciousness, and happiness.

Yue Gao has recently earned her PhD in computer science from Cornell University. Her research interests include affective computing, cognitive science, and machine learning. Earlier, she worked as software developer at Epic Systems. She also holds an MSc in computer science from Cornell and a BS in mathematics and computer science from University of Wisconsin at Madison.