Intrinsically Motivated Machines

Andrew G. Barto, Satinder Singh, and Richard L. Lewis

Abstract— **Intrinsic motivation is what causes us to do something "for its own sake," in contrast to doing something for an external reward. There is great interest in building intrinsic motivation into artificial systems by defining intrinsic reward signals within the reinforcement learning framework. Yet, what intrinsic reward signals are, and how it may differ from extrinsic reward signals, remains a murky and controversial subject. Here we approach this issue from an evolutionary perspective that leads to the conclusion there are no hard and fast features distinguishing intrinsic and extrinsic reward signals. Rather, there is a continuum along which reward signals range that depends on the directness and complexity of the relationship between the rewarded behavior and evolutionary success. This article contains work previously published by the authors and Jonathan Sorg in [26], [27].**

I. INTRODUCTION

Intrinsic motivation is what causes us to do something "for its own sake," in contrast to doing something for an external reward. The considerable interest in building intrinsic motivation into artificial agents is driven by its role in facilitating the acquisition of knowledge and the skills needed for an agent to operate successfully over extended periods of time in a domain, or across multiple domains, where it will be confronted with a spectrum of different tasks, the specifics of which are not known beforehand. Researchers call this "cumulative learning" or "developmental learning" during which the accumulation of knowledge and skills prepares the agent for specific tasks that are likely to be faced over its future. Prominent biological examples of intrinsically motivated behavior are exploration, play, manipulation, and behavior driven by curiosity. This paper describes an evolutionary perspective on intrinsic motivation that clarifies what it means to be motivated to do something for its own sake. The present article contains work previously reported by the authors and Jonathan Sorg in [26], [27].

The idea of giving intrinsic motivation to artificial systems is not new, having appeared, for example, in Lenat's AM system [10]. Most recent work in this direction uses the reinforcement learning (RL) framework [30], where reward signals are generated by the agent itself following a variety of ideas about how intrinsic motivation can be implemented computationally. This approach began in the early 1990s with Schmidhuber's introduction of methods for implementing a facsimile of curiosity using the RL framework [18], [19]. A basic assumption is that analogs of intrinsic motivation can be implemented by defining suitable reward functions. While this does not include all instances of intrinsically-motivated behavior, such as instinctive behavior that does not have to be learned, it does cover many cases of interest.

More recently, research on this subject has expanded, with contributions based on a variety of conceptions of how intrinsic motivation might be rendered in computational terms [7], [8], [9], [12], [13], [16], [17], [20], [32]. Our attention to this subject grew out of an interest in hierarchical RL [2] and the role that intrinsic motivation can play in constructing hierarchies of reusable skills [3], [4], [15], [21], [22], [23], [24], [25], [26], [29], [33].

Despite this recent attention, a computational account of intrinsic motivation, and how it may differ from extrinsic motivation, remains murky and controversial. Singh et al. [26] introduced an evolutionary framework for addressing these questions, along with the results of computational experiments that help to clarify some of the issues. They formulated a notion of an *optimal reward function* given a *fitness function*, where the latter is analogous to what in nature represents the degree of an animal's reproductive success. The present article describes this framework and some of those experimental results. This evolutionary perspective resolves what have been some of the most problematic issues surrounding the topic of intrinsic motivation, including the relationship of intrinsic and extrinsic motivation to primary and secondary reward signals, and the ultimate source of both forms of motivation.

Other researchers have reported interesting results of computational experiments involving evolutionary search for RL reward functions [1], [5], [11], [17], [28], but they did not directly address the motivational issues on which we focus. Uchibe and Doya [32] do address intrinsic reward in an evolutionary context, but their aim and approach differ significantly from ours. Following their earlier work [31], these authors treat extrinsic rewards as constraints on learning, while intrinsic rewards set the learning objective. The study closest to ours is that of Elfwing et al. [6] in which a genetic algorithm is used to search for shaping rewards [14] and other learning algorithm parameters that improve an RL learning system's performance.

II. BACKGROUND

A. RL Reward Functions

Tying down the entire behavior and learning processes of an RL system is a reward function: a real-valued function of the decision problem's states and actions given as part of the definition of the problem that the system is learning to solve. A measure of the amount of reward accumulated over

Andrew G. Barto is with the Department of Computer Science, University of Massachusetts, Amherst barto@cs.umass.edu

Satinder Singh is with the Division of Computer Science & Engineering, University of Michigan, Ann Arbor baveja@umich.edu

Richard L. Lewis is with the Department of Psychology, University of Michigan, Ann Arbor rickl@umich.edu

time constitutes the learning problem's objective function. In discussing intrinsic and extrinsic motivation it is useful to point out some correspondences between the RL framework and animal reward processes. Rewards in an RL system correspond to *primary rewards*, i.e., rewards that for animals exert their effects through processes hard-wired by evolution due to their relevance to reproductive success. Value functions used by some RL algorithms are the basis of *secondary* (or *conditoned* or *higher-order*) rewards, whereby learned predictions of primary reward act as reward themselves. To be more accurate, however, we should use the term *reward signal* instead of reward for what an RL system's reward function produces. What is usually meant by reward in psychology is an object of some kind that is given to an animal to encourage certain behavior. A non-zero reward signal may result from the presentation of an object, but it can be generated by other means as well.

Although characteristics of the the reward function effect the difficulty of the RL problem, and how well various RL algorithms perform, most RL algorithms make no assumptions about the reward function (except maybe boundedness). Reward signals can be generated by any type of process depending on a decision problem's states and the agent's actions. If the decision problem is construed to include components of the learning system itself, such as memory or prediction mechanisms, reward functions can depend on states of these components as well. For example, Schmidhuber [18], [19] proposed that by defining a reward signal as a function of changes in the errors of a prediction component, one obtains an analog of curiosity, whereby the agent will actively seek experiences that enable decreases in prediction errors. Crucially, the agent is not rewarded for making correct predictions, but rather is rewarded for *improving its predictions*.

The point is that there is great latitude in defining reward functions. RL algorithms "don't care" what generates reward signals. Some reward signals may be provided by other agents who can only observe the learning agent's external behavior, while other reward signals may be generated by mechanisms that monitor the internal workings, and the histories thereof, of the learning agent itself. This is extensively discussed in [4], [25], [26], [27].

It is tempting to define extrinsic rewards signals as those coming from outside the learning agent, and intrinsic reward signals as those generated inside the agent. However, this view is fraught with difficulties. In the first place, since it is reward *signals* that are important for learning, one must note that all reward signals are generated inside the agent. In the case of ourselves, for example, our brains contain extensive neural circuits devoted to generating reward signals. Some directly reflect external stimuli, as when we ingest tasty food, and some may be entirely internal. But the generation of most reward signals involves both internal and external information: e.g., a food reward signal depends on an internal state of satiety as well as an external object, and a curiosity reward signal—if we follow something like Schmidhuber proposed—depends on external events and our ability to predict them.

It is also tempting to define intrinsic reward signals as those reward signals learned through association with a primary reward signal. This is the process of secondary, or conditioned, reinforcement by which stimuli that predict primary reward become rewarding themselves. In psychological accounts, the view that intrinsic reward equals secondary reward is rejected due to experimental results showing that intrinsically motivated behavior is motivationally energizing and rewarding on its own and not because it predicts the satisfaction of a primary biological need. For example, children spontaneously explore very soon after birth, so there is little opportunity for them to experience extensive pairing of this behavior with primary reward.

There has never been any doubt, however, that exploration, manipulation, and other apparently intrinsically-motivated behaviors are important for an animal's survival and reproductive success if deployed in the right way. Appropriately cautious exploration, for example, clearly has implications for reproductive success because it can enable efficient foraging and successful escape when those needs arise. But an animal is not motivated to perform these behaviors because behaving this way previously *in its own lifetime* predicted biologically-primary rewards. The preponderance of evidence supports the view that the motivational forces driving these behaviors are built-in by the evolutionary process; not learned.

III. Evolutionary Perspective

It is therefore natural to investigate what an evolutionary perspective might tell us about the nature of intrinsic reward signals and how they might differ from extrinsic reward signals. We adopt the view discussed above that intrinsic reward is not the same as secondary reward. It is likely that the evolutionary process gave exploration, play, discovery, etc., positive hedonic valence because these behaviors contributed to reproductive success throughout evolution. Consequently, we regard intrinsic reward signals in the RL framework as primary reward signals, hard-wired from the start of the agent's life. Like any other primary reward signal in RL, they come to be predicted by the value-function learning system. These predictions can support secondary reinforcement so that predictors of intrinsically rewarding events can acquire rewarding qualities through learning just as predictors of extrinsically rewarding events can.

The evolutionary perspective thus leads to an approach in which adaptive agents, and *therefore their reward functions*, are evaluated according to their *expected fitness* given an explicit fitness function and some distribution of environments of interest. The fitness function maps trajectories of agent-environment interactions to scalar fitness values, and may take any form. In our approach, we search a space of primary reward functions for the one that maximizes the expected fitness of an RL agent that learns using that reward function.

Features of such an *optimal reward function*[1] and how these features relate to the environments in which agent lifetimes are evaluated provide insight into the relationship between extrinsic and intrinsic reward signals.

We turn next to a formal framework that captures the requisite abstract properties of agents, environments, and fitness functions and defines the evolutionary search for good reward functions as an optimization problem.

A. Optimal Reward Functions

We define an optimal reward function as follows. For a given RL agent A, there is a space, R_A, of reward functions that map an agent's state to a scalar primary reward that drives reinforcement learning. The composition of the state can depend on the agent architecture and its learning algorithm. There is a distribution over Markov decision process (MDP; [**?**]) environments in some set $\mathcal{E}$ in which we want our agents to perform well (in expectation). A specific reward function $r_A \in R_A$ and a sampled environment $E \in \mathcal{E}$ produces h, the history of agent A learning in environment E using the reward function r_A. A given fitness function F produces a scalar evaluation $F(h)$ for any such history h. An optimal reward function, $r_A^* \in R_A$, is the reward function that maximizes the expected fitness over the distribution of environments.

The formulation is very general because the constraints on A, R_A, F, and $\mathcal{E}$ are minimal. A is constrained only to be an agent that uses a reward function $r_A \in R_A$ to drive its search for behavior policies. F is constrained only to be a function that maps (finite or infinite) histories of agent-environment interactions to scalar fitness values. And $\mathcal{E}$ is constrained only to be a set of MDPs, though the Markov assumption can be easily relaxed.

The above formulation essentially defines a search problem—the search for r_A^*. This search is for a primary reward function and is to be contrasted with the search problem faced by an agent during its lifetime, that of learning a good value function, and hence a good secondary reward function, specific to its environment. Thus, our concrete hypothesis is (1) the r_A^* derived from search will capture physical regularities across environments in $\mathcal{E}$ as well as complex interactions between $\mathcal{E}$ and specific structural properties of the agent A (note that the agent A is part of its environment and is constant across all environments in $\mathcal{E}$), and (2) the value functions learned by an agent during its lifetime will capture regularities present within its specific environment that are not necessarily shared across environments.

IV. Experiment: Emergent Intrinsic Reward for Play and Manipulation

We now describe a computational experiment in which we directly specify the agent A with associated space of reward functions R_A, a fitness function F, and a set of environments $\mathcal{E}$, and derive r_A^* via (approximately) exhaustive search. This experiment was designed to serve three purposes. First, it will provide a concrete and transparent illustration of the basic optimal reward framework above. Second, it will demonstrate the *emergence* of interesting reward function properties that are not direct reflections of the fitness function—including features that might be intuitively recognizable as candidates for plausible intrinsic and extrinsic rewards in natural agents. Third, it will demonstrate the *emergence* of interesting reward functions that capture regularities across environments, and similarly demonstrate that value function learning by the agent captures regularities within single environments.

This experiment was designed to illustrate how our optimal reward framework can lead to the emergence of an intrinsic reward for actions such as playing with and manipulating objects in the external environment, actions that do not directly meet any primal needs (i.e., are not fitness inducing) and thus are not extrinsically motivating.

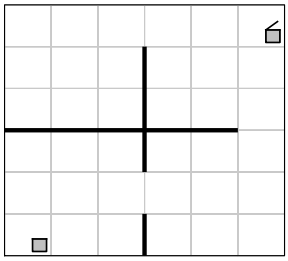

Fig. 1. Boxes environments. Each boxes environment is a 6×6 grid with two boxes that can contain food. The two boxes can be in any two of the four corners of the grid; the locations are chosen randomly for each environment. The agent has four (stochastic) movement actions in the four cardinal directions, as well actions to open closed boxes and eat food from the boxes when available. See text for further details.

A. Boxes Environments

We use a simulated physical space shown by the 6×6 grid in Fig. 1. It consists of four subspaces (of size 3×3). There are four movement actions, `North`, `South`, `East` and `West`, that if successful move the agent probabilistically in the direction implied, and if they fail leave the agent in place. Actions fail if they would move the agent into an outer bound of the grid or across a barrier, represented by one of the thick black lines in the figure. Consequently, the agent has to navigate through gaps in the barriers to move to adjacent subspaces. In each sampled environment two boxes are placed in randomly chosen special locations (from among the four corners and held fixed throughout the lifetime of the agent). This makes a uniform distribution over a space of six environments (the six possible locations of two indistinguishable boxes in the four corners). In addition to the usual movement actions, the agent has two special actions: `open`, which opens a box if it is closed and the agent is at the location of the box and has no effect otherwise (when a closed box is opened it transitions first to a half-open state for one time step and then automatically to an open state at the next time step regardless of the action by the agent), and

[1]We use this term despite the fact that none of our arguments depend on our search procedure finding true globally-optimal reward functions. We are concerned with reward functions that confer advantages over others and not with absolute optimality.

`eat`, which has no effect unless the agent is at a box location, the box at that location is half-open, and there happens to be food (prey) in that box, in which case the agent consumes that food.

An open box closes with probability 0.1 at every time step. A closed box always contains food. The prey always escapes when the box is open. Thus to consume food, the agent has to find a closed box, open it, and eat immediately in the next time step when the box is half-open. When the agent consumes food it feels *satiated* for one time step. The agent is *hungry* at all other time steps. The agent-environment interaction is not divided into trials or episodes. The agent's observation is 6 dimensional: the x and y coordinates of its location, its hunger-status, the open/half-open/closed status of both boxes, as well the presence/absence of food in the square where the agent is located. These environments are Markovian because the agent senses the status of both boxes regardless of location and because closed boxes always contain food; hence each immediate observation is a state.

B. Fitness

Each time the agent eats food its fitness is incremented by one. This is a surrogate for what in biology would be reproductive success (we could just as well have replaced the consumption of food event with a procreation event in our abstract problem description). The fitness objective, then, is to maximize the amount of food eaten over the agent's lifetime. Recall that when the agent eats it becomes satiated for one time step, and thus a direct translation of fitness into reward would assign a reward of $c > 0$ to all states in which the agent is satiated and a reward of $d < c$ to all other states. Thus, there is a space of *fitness-based reward functions*. We will refer to fitness-based reward functions in which d is constrained to be exactly 0 as *simple fitness-based reward functions*. Note that our definition of fitness is incremental or cumulative and thus we can talk about the cumulative fitness of even a partial (less than lifetime) history.

C. Agent

Our agent (A) uses the lookup-table ϵ-greedy Q-learning [34] algorithm with the following parameters: 1) Q_0, the initial Q-function (we use small values chosen uniformly randomly for each state-action pair from the range $[-0.001, 0.001]$) that maps state-action pairs to their expected discounted sum of future rewards, 2) α, the step-size, or learning-rate parameter, and 3) ϵ, the exploration parameter (at each time step the agent executes a random action with probability ϵ and the greedy action with respect to the current Q-function with probability $(1-\epsilon)$).

For each time step t, the current state is denoted s_t, the current Q-function is denoted Q_t, the agent executes an action a_t, and the Q-learning update is as follows:

$$Q_{t+1}(s_t, a_t) = (1-\alpha)Q_t(s_t, a_t) + \alpha[r_t + \gamma \max_b (Q_t(s_{t+1}, b)],$$

where r_t is the reward specified by reward function r_A for the state s_t, and γ is a discount factor that makes immediate reward more valuable than later reward (we use $\gamma = 0.99$ throughout).

We emphasize that the discount factor is an agent parameter that does not enter into the fitness calculation. That is, the fitness measure of a history remains the total amount of food eaten in that history for any value of γ the agent uses in its learning algorithm. It is well known that the form of Q-learning used above will converge asymptotically to the optimal Q-function[2] and hence an optimal policy. Thus, our agent uses its experience to continually adapt its action selection policy to improve the discounted sum of rewards, as specified by r_A, that it will obtain over its future (remaining in its lifetime). Note that the reward function is distinct from the fitness function F.

D. Space of Possible Rewards Functions

To make the search for an optimal reward function tractable, each reward function in the search space maps abstract features of each immediate observation to a scalar value. Specifically, we considered reward functions that ignore agent location and map each possible combination of the status of the two boxes and the agent's hunger-status to values chosen in the range $[-1.0, 1.0]$. This range does not unduly restrict generality because one can always add a constant to any reward function without changing optimal behavior. Including the box-status features allows the reward function to potentially encourage "playing with" boxes while the hunger-status feature is required to express the fitness-based reward functions that differentiate only between states in which the agent is satiated from all other states (disregarding box-status and agent location).

E. Finding a Good Reward Function

The psuedo-code below describes how we use simulation to estimate the *mean cumulative fitness* for a reward function r_A given a particular setting of agent (Q-learning) parameters (α, ϵ).

set (α, ϵ)
for $i = 1$ **to** N **do**
  Sample an environment E_i from $\mathcal{E}$
  In A, intialize Q-function
  Generate a history h_i over lifetime for A in E_i
  Compute fitness $F(h_i)$
end for
return average of $\{F(h_1), \ldots, F(h_N)\}$

For the results we report below, we estimate the mean cumulative fitness of r_A as the maximum estimate obtained (using the pseudo-code above) over a coarse discretization of the space of feasible (α, ϵ) pairs. Finding good reward functions for a given fitness function thus amounts to a large search problem. We discretized the range $[-1.0, 1.0]$ for each feasible setting of the three reward features such that we evaluated $54,000$ reward functions in the reward function space. We chose the discretized values based on experimental

[2]Strictly speaking, convergence with probability one requires the step-size parameter α to decrease appropriately over time, but for our purposes it suffices to keep it fixed at a small value.

experience with the boxes environments with various reward functions.

Note that our focus is on demonstrating the generality of our framework and the nature of the reward functions found rather than on developing efficient algorithms for finding good reward functions. Thus, we attempt to find a good reward function $\hat{r}_A^*$ instead of attempting the usually intractable task of finding the optimal reward function r_A^*, and we are not concerned with the efficiency of the search process. Neikum et al. [15] use genetic programming to perform this search in a more sophisticated manner.

F. Results

Recall the importance of regularities within and across environments to our hypotheses. In this experiment, what is unchanged across environments is the presence of two boxes and the rules governing food. What changes across environments—but held fixed within a single environment—are the locations of the boxes.

We ran this experiment under two conditions. In the first, called the *constant condition*, the food always appears in closed boxes throughout each agent's lifetime of $10,000$ steps. In the second, called the *step condition*, each agent's lifetime is $20,000$ steps, and food appears only in the *second half* of the agent's lifetime, i.e., there is never food in any of the boxes for the first half of the agent's lifetime, after which food always appears in a closed box. Thus in the step condition, it is impossible to increase fitness above zero until after the $10,000^{th}$ time step.

The step condition simulates (in extreme form) a developmental process in which the agent is allowed to "play" in its environment for a period of time in the absence of any fitness-inducing events (in this case, the fitness-inducing events are positive, but in general there could also be negative ones that risk physical harm). Thus, a reward function that confers advantage through exposure to this first phase must reward events that have only a distal relationship to fitness. Through the agent's learning processes, these rewards give rise to the agent's intrinsic motivation. Notice that this should happen in both the step and constant conditions; we simply expect it to be more striking in the step condition.

The left and middle panels of Fig. 2 show the mean (over 200 sampled environments) cumulative fitness as a function of time within an agent's lifetime under the two conditions. As expected, in the step condition, fitness remains zero under any reward function for the first $10,000$ steps. Also as expected, the best reward function outperforms the best fitness-based reward function over the agent's lifetime. The best fitness-based reward function is the best reward function in the reward function space that satisfies the definition of a fitness-based reward function for this class of environments. We note that the best fitness-based reward function assigns a negative value to states in which the agent is hungry (this makes the agent's initial Q-values optimistic and leads to efficient exploration; see Sutton and Barto [**?**] for an explanation of this effect). The best reward function outperforms the best simple fitness-based reward by a large margin (presumably because the latter cannot make the initial Q-values optimistic).

Table I shows the best reward functions and best fitness-based reward functions for the two conditions of the experiment, e.g., the best reward function for the Step condition is as follows: being satiated has a positive reward of 0.5 when both boxes are open and 0.3 when one box is open, being hungry with one box half-open has a small negative reward of -0.01, and otherwise being hungry has a reward of -0.05. Note that the agent will spend most of its time in this last situation. Of course, as expected and like the best fitness-based reward function, the best reward function has a high positive reward for states in which the agent is satiated. More interestingly, the best reward function in our reward function space rewards opening boxes (by making their half-open state rewarding relative to other states when the agent is hungry). This makes the agent "play" with the boxes and as a result learn the environment-specific policy to optimally navigate to the location of the boxes and then open them during the first half of the step condition so that when food appears in the second half, the agent is immediately ready to exploit that situation.

The policy learned under the best reward function has an interesting subtle aspect: it makes the agent run back and forth between the two boxes, eating from both boxes, because this leads to higher fitness (in most environments)[3] than staying at, and taking food from, only one box. This can be seen indirectly in the rightmost panel where the mean cumulative number of times both boxes are open is plotted as a function of time. It is clear that an agent learning with the overall best reward function keeps both boxes open far more often than one learning from the best fitness-based reward function. Indeed the behavior in the latter case is mainly to loiter near (an arbitrary) one of the boxes and repeatedly wait for it to close and then eat.

Finally, it is also noteworthy that there are other reward functions that keep both boxes open even more often than the best reward function (this is seen in the rightmost panel), but this occurs at the expense of the agent not taking the time to actually eat the food after opening a box. This suggests that there is a fine balance in the best reward function between intrinsically motivating "playing" with and manipulating the boxes and extrinsically motivating eating.

G. Summary

This experiment demonstrates that the evolutionary pressure to optimize fitness captured in the optimal reward framework can lead to the emergence of reward functions that assign positive *primary* reward to activities that are not directly associated with fitness. This was especially evident

[3]The agent could hang out at one box and repeatedly wait for it to close randomly and then open it to eat, but the probability of an open box closing was specifically (experimentally) chosen so that it is better for the agent in the distribution over environments to repeatedly move between boxes to eat from both. Specifically, an open box closes with probability 0.1 and thus on average in 10 time steps, while the average number of time steps to optimally travel between boxes across the 6 environments is less than 10 time steps.

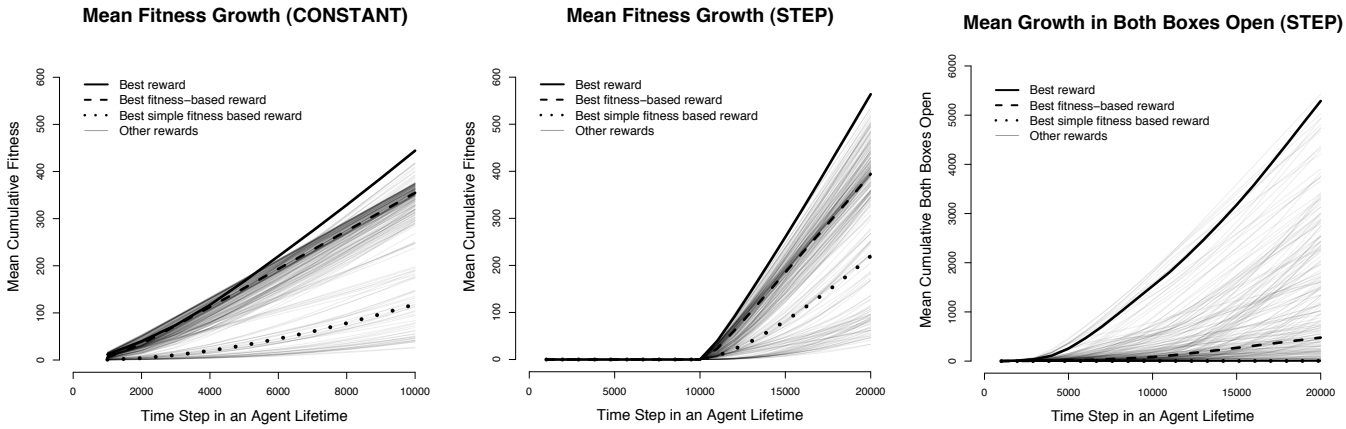

Fig. 2. Results from Boxes environments. The leftmost panel shows for the *constant* condition the mean cumulative (over agent lifetime) fitness achieved by all the reward functions sampled in our search for good reward functions. The middle panel shows the same results but for the *step* condition. The rightmost panel shows for the *step* condition the mean cumulative growth in the number of time steps both boxes were open for all the reward functions explored. In each panel, the curves for the best reward function, for the best fitness-based reward function, and for the best simple fitness-based reward functions are distinguished. See text for further details.

TABLE I

RESULTS FOR THE *step* AND *constant* CONDITIONS. EACH ROW OF PARAMETER VALUES DEFINES A REWARD FUNCTION BY SPECIFYING REWARD VALUES FOR EACH OF SEVEN FEASIBLE COMBINATIONS OF STATE FEATURES. THE COLUMN HEADINGS O, NOT-O, AND HALF-O, ARE SHORT FOR OPEN, NOT-OPEN AND HALF-OPEN RESPECTIVELY. SEE TEXT FOR FURTHER DETAILS.

CONDITION	REWARD TYPE	REWARD AS A FUNCTION OF STATE						
		Satiated		Hungry				
		o/o	o/not-o	o/o	o/not-o	o/half-o	not-o/half-o	not-o/not-o
Constant	Best	0.7	0.3	-0.01	-0.05	0.2	0.1	-0.02
	Best fitness-based	0.7	0.7	-0.005	-0.005	-0.005	-0.005	-0.005
Step	Best	0.5	0.3	-0.05	-0.05	-0.01	-0.01	-0.05
	Best fitness-based	0.5	0.5	-0.01	-0.01	-0.01	-0.01	-0.01

in the step condition of the experiment: during the first half of the agent's lifetime, no fitness-producing activities are possible, but intrinsically rewarding activities (running between boxes to keep both boxes open) are pursued that have fitness payoff later. The best (primary) reward captures the regularity of needing to open boxes to eat across all environments, while leaving the learning of the environment-specific navigation policy for the agent to accomplish within its lifetime by learning the Q-value function.

V. CONCLUSION

We believe that the optimal reward framework described here clarifies the computational role and origin of intrinsic and extrinsic motivation. More specifically, the experimental results support two claims about the implications of the framework for intrinsic and extrinsic motivation.

First, both intrinsic and extrinsic motivation can be understood as emergent properties of reward functions selected because they increase the fitness of learning agents across some distribution of environments. When coupled with learning, a primary reward function that rewards behavior that is useful across many environments can produce greater evolutionary fitness than a function exclusively rewarding behavior directly related to fitness. For example, in the experiment above, eating is necessary for evolutionary success in all environments, so we see primary rewards generated by (satiated) states resulting immediately from eating-related behavior. But optimal primary reward functions can also motivate richer kinds of behavior less directly related to basic needs, such as play and manipulation of the boxes, that can confer significantly greater evolutionary fitness to an agent. This is because what is learned as a result of being intrinsically motivated to play with and manipulate objects contributes, within the lifetime of an agent, to that agent's ability to survive and reproduce.

Second, the difference between intrinsic and extrinsic motivation is one of degree—there are no hard and fast features that distinguish them. A stimulus or activity comes to elicit reward to the extent that it helps the agent attain evolutionary success based on whatever the agent does to translate primary reward to learned secondary reward, and through that to behavior during its lifetime. What we call intrinsically rewarding stimuli or activities are those that bear only a distal relationship to evolutionary success. Extrinsically rewarding stimuli or events, on the other hand, are those that have a more immediate and direct relationship to evolutionary success.

Our optimal reward framework and experimental results

thus explain why evolution would give exploration, manipulation, play, etc. positive hedonic valence, i.e., make them rewarding, along with stimuli and activities that are more directly related to evolutionary success. The distinction between intrinsic and extrinsic motivation is therefore a matter of degree, but their source and role is computationally clear: both intrinsic and extrinsic motivation are emergent properties of a process that adjusts reward functions in pursuit of improved evolutionary success.

VI. ACKNOWLEDGMENTS

This research was funded by AFOSR grant FA9550-08-1-0418. Any opinions, findings, conclusions or recommendations expressed here are those of the authors and do not necessarily reflect the views of the sponsors.

REFERENCES

[1] D. H. Ackley and M. Littman. Interactions between learning and evolution. In C.G. Langton, C. Taylor, C.D. Farmer, and S. Rasmussen, editors, *Artificial Life II (Proceedings Volume X in the Santa Fe Institute Studies in the Sciences of Complexity*, pages 487–509. Addison-Wesley, Reading, MA, 1991.

[2] A. G. Barto and S. Mahadevan. Recent advances in hierarchical reinforcement learning. *Discrete Event Dynamcal Systems: Theory and Applications*, 13:341–379, 2003.

[3] A. G. Barto and Ö. Şimşek. Intrinsic motivation for reinforcement learning systems. In *Proceedings of the Thirteenth Yale Workshop on Adaptive and Learning Systems*, pages pp. 113–118, Center for Systems Science, Dunham Laboratory, Yale University, New Haven CT, 2005.

[4] A. G. Barto, S. Singh, and N. Chentanez. Intrinsically motivated learning of hierarchical collections of skills. In *Proceedings of the International Conference on Developmental Learning (ICDL)*, 2004.

[5] T. Damoulas, I. Cos-Aguilera, G. M. Hayes, and T. Taylor. Valency for adaptive homeostatic agents: Relating evolution and learning. In M. S. Capcarrere, A. A. Freitas, P. J. Bentley, C. G. Johnson, and J. Timmis, editors, *Advances in Artificial Life: 8th European Conference, ECAL 2005, LNAI vol. 3636*, pages 936–945. Springer-Verlag, Berlin, 2005.

[6] S. Elfwing, E. Uchibe, K. Doya, and H. I. Christensen. Co-evolution of shaping rewards and meta-parameters in reinforcement learning. *Adaptive Behavior*, 16:400–412, 2008.

[7] X. Huang and J. Weng. Novelty and reinforcement learning in the value system of developmental robots. In C. G. Prince, Y. Demiris, Y. Marom, H. Kozima, and C. Balkenius, editors, *Proceedings of the Second International Workshop on Epigenetic Robotics : Modeling Cognitive Development in Robotic Systems*, pages 47–55, Edinburgh, Scotland, 2002. Lund University Cognitive Studies.

[8] F. Kaplan and P.-Y. Oudeyer. Maximizing learning progress: An internal reward system for development. In F. Iida, R. Pfeifer, L. Steels, and Y. Kuniyoshi, editors, *Embodied Artificial Intelligence*, pages 259–270. Springer-Verlag, 2004.

[9] A. Kulakov and G. Stojanov. Structures, inner values, hierarchies and stages: Essentials for developmental robot architectures. In C. G. Prince, Y. Demiris, Y. Marom, H. Kozima, and C. Balkenius, editors, *Proceedings of the Second International Workshop on Epigenetic Robotics : Modeling Cognitive Development in Robotic Systems*, pages 63–69, Edinburgh, Scotland, 2002. Lund University Cognitive Studies.

[10] D. B. Lenat. *AM: An Artificial Intelligence Approach to Discovery in Mathematics*. PhD thesis, Stanford University, 1976.

[11] M. L. Littman and D. H. Ackley. Adaptation in constant utility nonstationary environments. In *Proceedings of the Fourth International Conference on Genetic Algorithms*, pages 136–142. 1991.

[12] S. Marsland, U. Nehmzow, and J. Shapiro. Novelty detection for robot neotaxis. In *Proceedings of the Second International ICSC Symposium on Neural Computation*, Berlin, Germany, 2000.

[13] K. E. Merrick and M. L. Maher. *Motivated Reinforcement Learing*. Springer-Verlag, Berlin, 2009.

[14] A. Ng, D. Harada, and S. Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Proceedings of the Sixteenth International Conference on Machine Learning*. Morgan Kaufmann, 1999.

[15] S. Niekum, A. G. Barto, and L. Spector. Genetic programming for reward function search. *IEEE Transactions on Autonomous Mental Development*, 2(2):132143, 2010. Special issue on Active Learning and Intrinsically Motivated Exploration in Robots: Advances and Challenges.

[16] P.-Y. Oudeyer, F. Kaplan, and V. Hafner. Intrinsic motivation systems for autonomous mental development. *IEEE Transactions on Evolutionary Computation*, 11, 2007.

[17] M. Schembri, M. Mirolli, and G. Baldassarre. Evolving internal reinforcers for an intrinsically motivated reinforcement-learning robot. In *Proceedings of the 6th International Conference on Development and Learning (ICDL2007)*, 2007.

[18] J. Schmidhuber. Adaptive confidence and adaptive curiosity. Technical Report FKI-149-91, Technische Universitat Munchen, 1991.

[19] J. Schmidhuber. A possibility for implementing curiosity and boredom in model-building neural controllers. In *From Animals to Animats: Proceedings of the First International Conference on Simulation of Adaptive Behavior*, pages 222–227, Cambridge, MA, 1991. MIT Press.

[20] J. Schmidhuber. Driven by compression progress: A simple principle explains essential aspects of subjective beauty, novelty, surprise, interestingness, attention, curiosity, creativity, art, science, music, jokes. In G. Pezzulo, M. V. Butz, O. Sigaud, and G. Baldassarre, editors, *Anticipatory Behavior in Adaptive Learning Systems. From Psychological Theories to Artificial Cognitive Systems*, pages 48–76. Springer, Berlin, 2009.

[21] Ö. Şimşek and A. G. Barto. Using relative novelty to identify useful temporal abstractions in reinforcement learning. In *Proceedings of the Twenty-First International Conference on Machine Learning (ICML 04)*, 2004.

[22] Ö. Şimşek and A. G. Barto. An intrinsic reward mechanism for efficient exploration. In *Proceedings of the Twenty-Third International Conference on Machine Learning (ICML-06)*, pages 833–840, New York, 2006. ACM Press.

[23] Ö. Şimşek and A. G. Barto. Betweenness centrality as a basis for forming skills. Technical Report TR-2007-26, University of Massachusetts, Department of Computer Science, Amherst, MA, 2007.

[24] Ö. Şimşek, A. P. Wolf, and A. G. Barto. Identifying useful subgoals in reinforcement learning by local graph partitioning. In *Proceedings of the Twenty-Second International Conference on Machine Learning (ICML 05)*, 2005.

[25] S. Singh, A. G. Barto, and N. Chentanez. Intrinsically motivated reinforcement learning. In *Advances in Neural Information Processing Systems 17: Proceedings of the 2004 Conference*, Cambridge MA, 2005. MIT Press.

[26] S. Singh, R. L. Lewis, and A. G. Barto. Where do rewards come from? In N.A. Taatgen and H. van Rijn, editors, *Proceedings of the 31st Annual Conference of the Cognitive Science Society*, pages 2601–2606. Cognitive Science Society, 2009.

[27] S. Singh, R. L. Lewis, A. G. Barto, and J. Sorg. Intrinsically motivated reinforcement learning: An evolutionary perspective. *IEEE Transactions on Autonomous Mental Development*, 2(2):7082, 2010. Special issue on Active Learning and Intrinsically Motivated Exploration in Robots: Advances and Challenges.

[28] M. Snel and G. M. Hayes. Evolution of valence systems in an unstable environment. In *Proceedings of the 10th International Conference on Simulation of Adaptive Behavior: From Animals to Animats*, pages 12–21, Osaka, Japan, 2008.

[29] J. Sorg, S. Singh, and R. L. Lewis. Internal Rewards Mitigate Agent Boundedness. In *Proceedings of the 27th International Conference on Machine Learning*, Haifa, Israel, 2010. To appear.

[30] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA, 1998.

[31] E. Uchibe and K. Doya. Constrained reinforcement learning from intrinsic and extrinsic rewards. In *Proceedings of the IEEE International Conference on Developmental Learning*. London, UK, 2007.

[32] E. Uchibe and K. Doya. Finding intrinsic rewards by embodied evolution and constrained reinforcement learning. *Neural Networks*, 21(10):1447–1455, 2008.

[33] C. Vigorito and A. G. Barto. Intrinsically motivated hierarchical skill learning in structured environments. *IEEE Transactions on Autonomous Mental Development*, 2(2):8390, 2010. Special issue on Active Learning and Intrinsically Motivated Exploration in Robots: Advances and Challenges.

[34] C. J. C. H. Watkins. *Learning from Delayed Rewards*. PhD thesis, Cambridge University, Cambridge, England, 1989.