Intrinsically Motivated Learning of Hierarchical Collections of Skills

Andrew G. Barto
Department of Computer Science
University of Massachusetts
Amherst MA
barto@cs.umass.edu

Satinder Singh
Computer Science and Engineering
University of Michigan
Ann Arbor MI
baveja@umich.edu

Nuttapong Chentanez
Electrical Engineering and Computer Science
University of Michigan
Ann Arbor MI
nchentan@umich.edu

Abstract

Humans and other animals often engage in activities for their own sakes rather than as steps toward solving practical problems. Psychologists call these intrinsically motivated behaviors. What we learn during intrinsically motivated behavior is essential for our development as competent autonomous entities able to efficiently solve a wide range of practical problems as they arise. In this paper we present initial results from a computational study of intrinsically motivated learning aimed at allowing artificial agents to construct and extend hierarchies of reusable skills that are needed for competent autonomy. At the core of the model are recent theoretical and algorithmic advances in computational reinforcement learning, specifically, new concepts related to skills and new learning algorithms for learning with skill hierarchies.

1. Introduction

Despite impressive power and utility, today's machine learning algorithms fall far short of the possibilities for machine learning. They are typically applied to single, isolated problems for each of which they have to be hand-tuned and for which training data sets have to be carefully prepared. They do not have the generative capacity required to significantly extend their abilities beyond initially built-in representations. They do not address many of the reasons that learning is so useful in allowing animals to cope flexibly with new problems as they arise over extended periods of time. Numerous researchers have persuasively argued that a developmental approach is necessary to address these shortcomings (e.g., [31]), drawing from cognitive science, neuroscience, artificial intelligence, and philosophy. According to this approach, an agent undergoes an extended developmental period during which collections of reusable skills are autonomously learned that will be useful for a wide range of later challenges.

Although these arguments are compelling, developmental approaches to artificial agent design have been slow to penetrate the mainstream of the machine learning community. Implementations remain largely exploratory, and they have not yet led to the kind of mathematical formulation required to engage the largest part of the machine learning community. This paper presents preliminary work from a long-term project that seeks to address these shortcomings by elaborating the well-developed computational reinforcement learning (RL) framework [28] to encompass the autonomous development of skill hierarchies through *intrinsically motivated learning*. An agent's activity is said to be intrinsically motivated if the agent engages in it for its own sake rather than as a step toward solving a specific problem.

Our approach builds on existing research in machine learning, with input from recent advances in the neuroscience of brain reward systems as well as classical and contemporary psychological theories of motivation. Not all of our ideas are new, having antecedents in many different areas, including some in machine learning and RL as we outline below. However, we argue that *recent theoretical and computational advances in RL provide important components for making these ideas work efficiently in artificial agents.*

2. Background

Psychologists distinguish between *extrinsic motivation*, which means being moved to do something because of some specific rewarding outcome, and *intrinsic motivation*, which refers to being moved to do something because it is inherently enjoyable. Intrinsic motivation leads organisms to engage in exploration, play, and other behavior driven by curiosity in the absence of explicit reward. In a classic paper, White [32] argued that intrinsically motivated behavior is essential for an organism to gain the competence necessary for autonomy. A system that is competent in this sense has *a broad set of reusable skills* for controlling its environment. The activity through which these broad skills are learned is motivated by an intrinsic reward system that favors the development of broad competence rather than being directed to more specific externally-directed goals. But these skills act as the "building blocks" out of which an agent can form solutions to specific problems that arise over its lifetime. Instead of facing each new challenge by trying to create a solution out of low-level primitives, it can focus on combining and adjusting higher-level skills, greatly increasing the efficiency of learning to solve new problems.

Psychology—A large collection of psychological literature inspires our approach. In 1959 White [32] influentially reviewed the evidence that the (even then) classical Hullian view of motivation in terms of reducing drives related to the biologically primary needs for food, water, sex, and escape was not sufficient to account for an animal's exploratory behavior. Ample evidence existed—and has been greatly augmented since then—that the opportunity to explore a novel environment can itself act as reward. Moreover, not only exploration incited by novelty, but also manipulation, or just activity itself, can be rewarding. This is supported by experimental evidence showing that these activities are not always secondary reinforcers: their motivational significance is built-in rather than being acquired through association with a standard primary reinforcer. The modern expression of these views is most clearly seen in developmental and educational psychology, where a distinction is drawn between intrinsic and extrinsic motivation [5].

The psychology literature is less helpful in specifying the concrete properties of experience that incite intrinsically motivated behavior, although there have been many hypotheses. Berlyne [2] probably had the most to say on these issues, suggesting that the factors underlying intrinsic motivational effects involve novelty, surprise, incongruity, and complexity. He also hypothesized that moderate levels of novelty have the highest hedonic value because the rewarding effect of novelty is overtaken by an aversive effect as novelty increases. This is consistent with many other views holding that situations intermediate between complete familiarity (boredom) and complete unfamiliarity (confusion) have the most hedonic value. Another hypothesis about what we find satisfying in exploration and manipulation is that we enjoy "being a cause" [9], which is a major component of Piaget's theory of child development [20]. In this paper, we use only the degree of surprise of salient stimuli as intrinsic reward, but this is merely a starting point.

Neuroscience—The neuromodulator dopamine has long been associated with reward learning and rewarded behavior, partly because of clear evidence of its key role in drugs of addiction [6]. The original observation [12, 8, 18, 26] that the activity of dopamine cells in the monkey midbrain in reward-learning tasks closely follows the form of a key training signal in RL (the temporal difference prediction error) is an important backdrop for our approach.

Recent studies [15, 3] have focused on the idea that dopamine not only plays a critical role in the extrinsic motivational control of behaviors aimed at harvesting explicit rewards, but also in the intrinsic motivational control of behaviors associated with novelty and exploration. For instance, salient, novel sensory stimuli inspire the same sort of phasic activity of dopamine cells as unpredicted rewards [25, 11]. However, this activation extinguishes more or less quickly as the stimuli become familiar. This may underlie the fact that novelty itself has rewarding characteristics [21]. Theoretical treatments [14, 15] have directly related dopamine activity with mechanisms for controlling exploration in RL such as exploration and shaping bonuses [27, 4, 19]. Although space here does not permit development of these connections, they form key components of our approach to intrinsically motivated RL.

Computational Models of Intrinsic Motivation—Although there have been previous computational studies related to intrinsic motivation, most relevant is recent work from the epigenetic robotics community, some of which discusses the important role of novelty and curiosity in intelligent behavior (e.g., [13, 16]). However, this work does not build upon the mathematical framework of RL and does not use the recently-developed RL methods that we employ. Closely related RL research is that of Schmidhuber (e.g., [23, 24]) on curiosity and exploration. While some promising initial results were demonstrated, this work was left in a very preliminary state, and it also predates the new RL methods that we use.

Interestingly, the most closely related recent computational work comes from the field of architecture and design. In a study of artificial creativity, Saunder's recent thesis [22] presents a system that includes intrinsic motivation based on novelty and surprise following Berlyne's [2] theories. We find this work inspiring, though it focuses on searching design spaces rather than the development of reusable sequential skills.

3. Intrinsic Motivation in Reinforcement Learning

RL is a very active area of machine learning, with considerable attention also being received from decision theory, operations research, and control engineering. RL algorithms address the problem of how a behaving agent can learn to approximate an optimal behavioral strategy, usually called a *policy*, while interacting directly with its environment. In the terms of control engineering, RL consists of methods for the on-line approximation of closed-loop solutions to stochastic optimal control problems, usually under conditions of incomplete knowledge of the system being controlled. One can think of a problem's optimality criterion as defining a primary reward function, and one can think of an approximate solution as the skill of expertly controlling the given system according to this optimality criterion.

In what follows, we describe the elements of the standard RL framework that our approach builds upon, and then we describe a preliminary simulation we have produced that shows how these elements can be exploited for intrinsically motivated learning.

Internal and External Environments—According to the "standard" view of RL (e.g., [28]) the agent-environment interaction is envisioned as the classical interaction between a controller (the agent) and the controlled system (the environment), with a specialized reward signal coming from the environment to the agent that provides at each moment of time an evaluation (usually with a scalar reward value) of the agent's ongoing behavior. The component of the environment that provides this evaluation is usually called the "critic" (Fig. 1A). The agent learns to improve its skill in controlling the environment in the sense of learning how to increase the total amount of reward it receives over time from the critic. With appropriate mathematical assumptions, the problem faced by the learning agent is that of approximating an optimal policy for a Markov Decision Process (MDP).

Sutton and Barto [28] carefully point out that the scheme in Fig. 1A is quite abstract and that one should not identify this RL agent with an entire animal or robot. An animal's reward signals are determined by processes within its brain that monitor not only external events through exteroceptive systems but also the animal's internal state, which includes information pertaining to critical system variables (e.g., blood-sugar level) as well as memories and accumulated knowledge. The critic is in an animal's head. Fig. 1B makes this more explicit by "factoring" the environment of Fig. 1A into an *external environment* and an *internal environment*, the later of which contains the critic which determines primary reward. Notice that this scheme still includes cases in which reward can be thought of as an external stimulus (e.g., a pat on the head or a word of praise). These are simply stimuli transduced by the internal environment so as to generate the appropriate level of primary reward.

Because Fig. 1B is a refinement of Fig. 1A (that is, it is the result of adding structure rather than changing it), the standard RL framework already encompasses intrinsic reward. In fact, according to this model, *all* reward is intrinsic, and what psychologists would call extrinsic reward is just intrinsic reward that is directly triggered by external events. But the point of departure for our approach is to note that the internal environment contains, among other things, the organism's motivational system, *which needs to be a sophisticated system that should not have to be redesigned for different problems*. In contrast, the usual practice in applying RL algorithms is to formulate the problem one wants the agent to learn how to solve (e.g., win at backgammon) and define a reward function specially tailored for this problem (e.g., reward = 1 on a win, reward = 0 on a loss). Sometimes considerable ingenuity is required to craft an appropriate reward function. In effect, a different special-purpose motivational system is hand-crafted for each new problem. This should be largely unnecessary.

Skills—Autonomous mental development should result in a collection of reusable skills. But what do we mean by a skill? Recent RL research provides a concrete answer to this question, together with a set of algorithms capable of improving skills with experience. To combat the complexity of learning in difficult domains, RL researchers have turned to principled ways of exploiting "temporal abstraction," where decisions are not required at each step, but rather where each decision invokes the execution of a temporally-extended activity which follows its own closed-loop policy until termination. Substantial theory exists on how to plan and learn when temporally-extended skills are added to the set of actions available to an agent. Since a skill can invoke other skills as components, hierarchical control architectures and learning algorithms naturally emerge from this conception of a skill. Specifically, our approach builds on the theory of *options* [29].

Briefly, an option is something like a subroutine. It consists of 1) an *option policy* that directs the agent's behavior for a subset of the environment states, 2) an *initiation set* consisting of all the states in which the option can be initiated, and 3) a *termination condition*, which specifies the conditions under which the option terminates. It is important to note that an option is not a sequence of actions; it is a closed-loop control rule, meaning that it is responsive to on-going state changes. Theoretically, when options are added to the set of admissible agent actions, the usual MDP formulation of RL extends to semi-Markov decision processes (SMDPs), with the one-step actions now becoming the "primitive actions." All of the theory and algorithms applicable to SMDPs can be appropriated for decision making

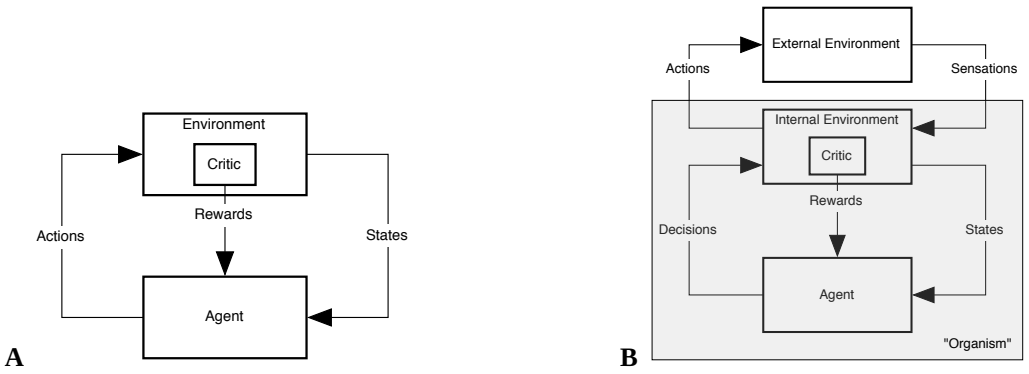

Figure 1. *Agent-Environment Interaction in Reinforcement Learning.* **A***: Reward is supplied to the agent from a "critic" in its environment.* **B***: An elaboration of Panel A in which the environment is factored into and internal and external environment, with reward coming form the former. The shaded box corresponds to what we would think of as the "organism."*

and learning with options [1, 29].

Two components of the the options framework are especially important for our approach:

1. *Option Models*: An option model is a probabilistic description of the effects of executing an option. As a function of an environment state where the option is initiated, it gives the probability with which the option will terminate at any other state, and it gives the total amount of reward expected over the option's execution. Option models can be learned from experience (usually only approximately) using standard methods. Option models allow stochastic planning methods to be extended to handle planning at higher levels of abstraction.

2. *Intra-option Learning Methods*: These methods allow the policies of many options to be updated simultaneously during an agent's interaction with the environment. If an option *could have* produced a primitive action in a given state, its policy can be updated on the basis of the observed consequences even though it was not directing the agent's behavior at the time. Intra-option methods essentially "multiplex" experience to greatly increase the efficiency of learning [29].

In most of the work with options, the set of options must be provided by the system designer. While an option's policy can be improved through learning, each option has to be predefined by providing its initiation set, termination condition, and the reward function that evaluates its performance. Many researchers have recognized the desirability of automatically creating options, and several approaches have recently been proposed (e.g., [7, 10, 17]). For the most part, these methods extract options from the learning system's attempts to solve a particular problem, whereas our approach creates options outside of the context of solving any particular problem.

Developing Hierarchical Collections of Skills—It is clear that children accumulate skills while they engage in intrinsically motivated behavior, e.g., while at play. When they notice that something they can do reliably results in an interesting consequence, they remember this in a form that will allow them to bring this consequence about if they wish to do so at a future time when they think it might contribute to a specific goal. Moreover, they improve the efficiency with which they bring about this interesting consequence with repetition, before they become bored and move on to something else. *We claim that the concepts of an option and an option model are exactly appropriate for developing analogs of this type of behavior in artificial agents*. An option model is not a passive model of environment dynamics; it is conditioned on the agent's activity. An option model basically says that "If I begin this behavior in this situation, then this is what is likely to happen." When stored appropriately, the agent will effectively know that it has the means to efficiently bring about these consequences, which is what the agent needs to know to both learn higher-level skills (that use lower-level skills as building blocks) and to learn how to solve specific tasks as they arise.

All skills acquired in this way do not have to be useful. Later learning in the context of specific tasks will assign values to skills depending on how useful they turn out to be. We already know how to do this using recently-developed hierarchical RL algorithms. The major computational challenge is to develop and cache a set of skills that is rich in skills that are likely to be widely useful. Intrinsic reward does not have to infallibly identify useful activities, but it has to do a reasonable job of identifying good candidates—and it shouldn't miss too much. If we speculate about the

evolution of intrinsic motivational systems in animals, it is plausible that they have been tuned through evolution to do exactly this, resulting in the kind of "drive for mastery" that has been discussed by psychologists for at least half a century.

What kind of intrinsic reward function do we propose to implement? While there are several sources of inspiration for this as discussed above, in this work we focus on the striking connection between computational RL algorithms and the activity of dopamine neurons. In particular, we will illustrate how we use a kind of "surprise" analogous to the so-called novelty responses of dopamine neurons to implement one form of intrinsic reward.

Whatever the details of how intrinsic reward is defined, it should diminish with continued repetition of the activity that generates it. For example, continued exercise of causal influence on the environment should effectively lose its rewarding quality after becoming sufficiently "routine" (i.e., the agent gets bored). As a result, the agent moves on to learn another skill based on its discovery of another mode of controlling its environment, and so on. Similarly, exploration of regions about which the agent is not yet ready to learn should be aversive to the agent. Skills formed through earlier experience are available as action choices in this RL process. Policies for new skills have the potential of invoking existing skills. This will allow the construction of hierarchically organized collections of skills that become more sophisticated as the agent continues to accumulate experience. This process will naturally produce what Utgoff and Stracuzzi [30] called "many-layered" learning in which the agent learns what is easy to learn first, then uses this knowledge to learn harder things. This results in a generative power that is absent from current machine learning systems.

4. An Example

To make our discussion above more concrete, we briefly describe an example implementation of some of these ideas in a simple artificial "playroom" domain shown in Fig. 2A. In the playroom are a number of objects: a light switch, a ball, a bell, two movable blocks that are also buttons for turning music on and off, as well as a toy monkey that can make sounds. The agent has an eye, a hand, and a visual marker (seen as a cross hair in the figure). At any time step, the agent has the following actions available to it: 1) move eye to hand, 2) move eye to marker, 3) move eye one step north, south, east or west, 4) move eye to random object, 5) move hand to eye, 6) move hand to marker, 7) move marker to eye, and 8) move marker to hand. In addition, if both the eye and and hand are on some object, then natural operations suggested by the object become available, e.g., if both the hand and the eye are on the light switch then the action of flicking the light switch becomes available, and if both the hand and eye are on the ball, then the action of pushing the ball become available (the ball when pushed moves in a straight line to the marker), etc. Finally, there is a visual-search action that moves the eye to a random object in the room.

The objects in the playroom all have potentially interesting characteristics. The bell rings once and moves to a random adjacent square if the ball is kicked into it. The light switch controls the lighting in the room. The color of any of the blocks in the room is only visible if the light is on, otherwise they appear similarly gray. The blue block if pressed turns music on, while the red block if pressed turns music off. Either block can be pushed and as a result it moves to a random adjacent square. The toy monkey makes frightened sounds if simultaneously the room is dark and the music is on and the bell is rung. These objects were designed to have varying degrees of difficulty to engage. For example, to get the monkey to cry out requires the agent to do the following sequence of actions: 1) get its eye to the light switch, 2) move hand to eye, 3) push the light switch to turn the light on, 4) find the blue block with its eye, 5) move the hand to the eye, 6) press the blue block to turn music on, 7) find the light switch with its eye, 8) move hand to eye, 9) press light switch to turn light off, 10) find the bell with its eye, 11) move the marker to the eye, 12) find the ball with its eye, 13) move its hand to the ball, and 14) kick the ball to make the bell ring. Notice that if the agent has already learned how to turn the light on and off, how to turn music on, and how to make the bell ring, then those learned skills would be of obvious use in simplifying this process of engaging the toy monkey.

For this simple example, the agent has built-in notions of salience of stimuli. In particular, changes in light and sound intensity are considered salient by the playroom agent. The agent behaves by choosing actions according to an ϵ-greedy policy with respect to its value function [28]. Because the initial value function is uninformative, the agent starts by exploring its environment randomly. Each first encounter with a salient event initiates the learning of an option and an option-model for that salient event. For example, the first time the agent happens to turn the light on, it initiates the data-structures necessary for learning and storing the light-on option, including the initiation set, the policy, the termination probabilities, as well as for storing the light-on option-model including the terminal-state probabilities and the expected reward until termination. As the agent moves around the world, all the options and their models are simultaneously updated using intra-option learning algorithms. Initially, of course, the light-on option and its model will be nearly empty.

The agent's intrinsic reward is generated in a way suggested by the novelty response of dopamine neurons. The

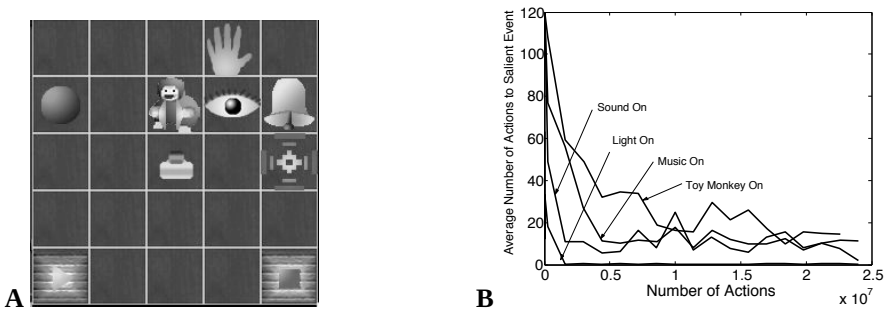

Figure 2. **A.** *Playroom domain. See text for details.* **B.** *Speed of learning of various skills. See text for details*

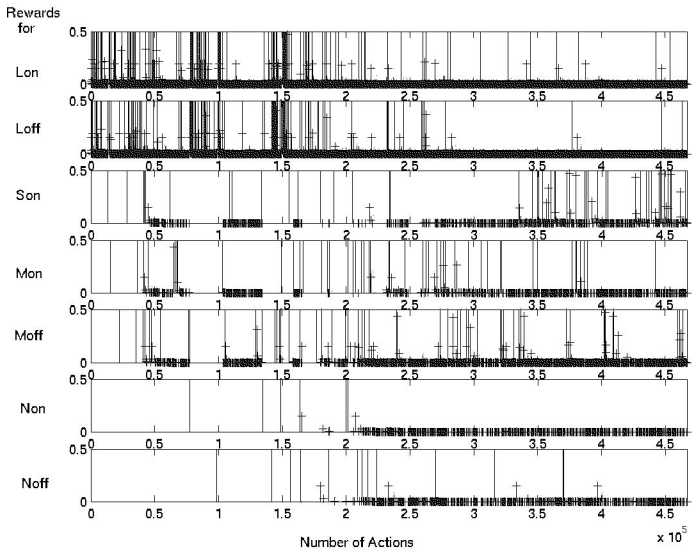

Figure 3. *Occurrence and magnitude of rewards for the salient events. See text for details.*

intrinsic reward for each salient event is proportional to the error in its prediction of that salient event according to the learned option model for that event. The intrinsic reward is used to update the value function the agent is using to determine its behavior in the playroom. As a result, when the agent encounters an unpredicted salient event a few times, its updated value function drives it to repeatedly attempt to achieve that salient event. There are two interesting side effects of this: 1) as the agent tries to repeatedly achieve the salient event, learning improves both its policy for doing so and its option-model that predicts the salient event, and 2) as its option policy and option model improve, the intrinsic reward diminishes and the agent gets "bored" with the associated salient event and moves on. Of course, the option policy and model become accurate in states the agent encounters frequently. Occasionally, the agent encounters the salient event in a state (set of sensor readings) that it has not encountered before, and it generates intrinsic reward again (it is "surprised").

A summary of results is presented in Fig. 3. Each panel of the figure is for a distinct salient event. The graph in each panel shows both the time steps at which the event occurs and the intrinsic reward associated by the agent to each occurrence. Each occurrence is denoted by a vertical bar whose height denotes the amount of associated intrinsic reward. Note that as one goes from top to bottom in this figure, the salient events become harder to achieve and, in fact, become more hierarchical. Indeed, the lowest one for turning on the monkey noise (Non) needs light on, music on, light off, sound on in sequence. A number of interesting results can be observed in this figure. First note that the salient events that are simpler to achieve occur earlier in time. For example, Lon (light turning on) and Loff (light turning off) are the simplest salient events, and the agent makes these happen quite early. The agent tries them a number of times (determined by the learning rate parameter and details of the agent's current value function) before getting bored and moving on to other salient events. The reward obtained for

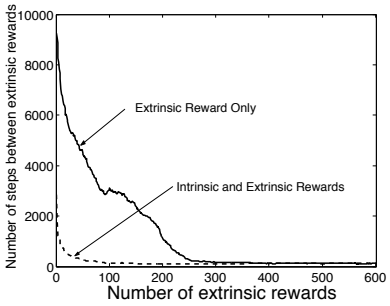

Figure 4. *The effect of intrinsically motivated learning when extrinsic reward is present. See text for details.*

each of these events diminishes after repeated exposure to the event. Thus, automatically, the skill of achieving the simpler events are learned before those for the more complex events.

Of course, the events keep happening despite their diminished capacity to reward because they are needed to achieve the more complex events. Consequently, the agent continues to turn the light on and off even after it has learned this skill because this is a step along the way toward turning on the music, as well along the way toward turning on the monkey noise. Finally note that the more complex skills are learned relatively quickly once the required sub-skills are in place, as one can see by the few rewards the agent receives for them. The agent is able to bootstrap and build upon the options it has already learned for the simpler events. The fact that all the options are learned is also seen in Fig. 2B, which shows how the time it takes the agent to bring about each option's target event changes with the agent's experience (there is an upper cutoff of 120 steps). This figure also shows that the simpler skills are learned earlier than the more complex ones.

An agent having a collection of skills learned through intrinsic reward can learn a wide variety of extrinsically rewarded tasks more easily than an agent lacking these skills. To illustrate, we looked at a playroom task in which extrinsic reward was available only if the agent succeeded in making the monkey cry out. This requires the 14 steps described above. This is difficult for an agent to learn if only the extrinsic reward is available, but much easier if the agent can use intrinsic reward to learn a collection of skills, some of which are relevant to the overall task. Fig. 4 compares the performance of two agents in this task. Each starts out with no knowledge of task, but one employs the intrinsic reward mechanism we have discussed above. The extrinsic reward is always available, but only when the monkey cries out. The figure, which shows the average of 100 repetitions of the experiment, clearly shows the advantage of learning with intrinsic reward.

5. Discussion

While the experiment and results described above serve as a concrete illustration of our basic ideas, they are merely a starting point in our study of intrinsically motivated learning. One of the key aspects of the Playroom example is that intrinsic reward is generated only by unexpected salient events. But this is only one of the simplest possibilities and has many limitations. It cannot account for what makes many forms of exploration and manipulation "interesting." In the future, we intend to implement computational analogs of other forms of intrinsic motivation as suggested by the psychological and neuroscience literatures and guided by the statistical

Despite the "toy" nature of this domain, these results are among the most sophisticated we have seen involving intrinsically motivated learning. Moreover, they were achieved quite directly by combining a collection of existing RL algorithms for learning options and option-models with a simple notion of intrinsic reward. The idea of intrinsic motivation for artificial agents is certainly not new, but we hope to have shown that the elaboration of the formal RL framework in the direction we have suggested, together with the use of recently-developed hierarchical RL algorithms, provides a fruitful basis for developing competently autonomous agents.

Acknowledgement Andrew Barto's research was funded by NSF grant CCF 0432143 and by a grant from DARPA's IPTO program. Satinder Singh and Nuttapong Chentanez were funded by NSF grant CCF 0432027 and by a grant from DARPA's IPTO program. The authors thank Peter Dayan for his essential input regarding the neuroscience of intrinsic reward systems.

References

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

[2] D. E. Berlyne. *Conflict, Arousal. and Curiosity*. McGraw-Hill, N.Y., 1960.

[3] P. Dayan and B. W. Balleine. Reward, motivation and reinforcement learning. *Neuron*, 36:285–298, 2002.

[4] P. Dayan and T. J. Sejnowski. Exploration bonuses and dual control. *Machine Learning*, 25:5–22, 1996.

[5] E. L. Deci and R. M. Ryan. *Intrinsic Motivation and Self-Determination in Human Behavior*. Plenum Press, N.Y., 1985.

[6] G. Di Chiara. Drug addiction as dopamine-dependent associative learning disorder. *European Journal of Pharmacology*, 375(1-3):13–30, 1999.

[7] B. Digney. Learning hierarchical control structure from multiple tasks and changing environments. In *From Animals to Animats 5: The Fifth Conference on Simulation of Adaptive Behavior*, Cambridge, MA, 1998. MIT Press.

[8] K. J. Friston, G. Tononi, G. N. Reeke, O.Sporns, and G. M. Edelman. Value-dependent selection in the brain: Simulation in a synthetic neural model. *Neuroscience*, 59:229–243, 1994.

[9] K. Groos. *The Play of Man*. D. Appleton, N.Y., 1901.

[10] B. Hengst. Discovering hierarchy in reinforcement learning with HEXQ. In *Maching Learning: Proceedings of the Nineteenth International Conference on Machine Learning*, pages 243–250, San Francisco, CA, 2002. Morgan Kaufmann.

[11] J. C. Horvitz, T. Stewart, and B. Jacobs. Burst activity of ventral tegmental dopamine neurons is elicited by sensory stimuli in the awake cat. *Brain Research*, 759:251–258, 1997.

[12] J. C. Houk, J. L. Adams, and A. G. Barto. A model of how the basal ganglia generates and uses neural signals that predict reinforcement. In J. C. Houk, J. L. Davis, and D. G. Beiser, editors, *Models of Information Processing in the Basal Ganglia*, pages 249–270. MIT Press, Cambridge, MA, 1995.

[13] 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.

[14] S. Kakade and P. Dayan. Dopamine bonuses. In T. K. Leen, T. G. Dietterich, and V. Tresp, editors, *Advances in Neural Information Processing Systems 13*, pages 131–137. MIT Press, 2001.

[15] S. Kakade and P. Dayan. Dopamine: Generalization and bonuses. *Neural Networks*, 15:549–559, 2002.

[16] F. Kaplan and P.-Y. Oudeyer. Motivational principles for visual know-how development. In C. G. Prince, L. Berthouze, H. Kozima, D. Bullock, G. Stojanov, and C. Balkenius, editors, *Proceedings of the Third International Workshop on Epigenetic Robotics : Modeling Cognitive Development in Robotic Systems*, pages 73–80, Edinburgh, Scotland, 2003. Lund University Cognitive Studies.

[17] A. McGovern. *Autonomous Discovery of Temporal Abstractions from Interaction with An Environment*. PhD thesis, University of Massachusetts, 2002.

[18] P. R. Montague, P. Dayan, and T. J. Sejnowski. A framework for mesencephalic dopamine systems based on predictive hebbian learning. *Journal of Neuroscience*, 16:1936–1947, 1996.

[19] 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.

[20] J. Piaget. *The Origins of Intelligence in Children*. Norton, N.Y., 1952.

[21] P. Reed, C. Mitchell, and T. Nokes. Intrinsic reinforcing properties of putatively neutral stimuli in an instrumental two-lever discrimination task. *Animal Learning and Behavior*, 24:38–45, 1996.

[22] R. Saunders. *Curious Design Agents and Artificial Creativity: A Synthetic Approach to the Study of Creative Behaviour*. PhD thesis, University of Sydney, 2002.

[23] 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.

[24] J. Schmidhuber and J. Storck. Reinforcement driven information acquisition in nondeterministic environments, 1993. Technical report, Fakultat fur Informatik, Technische Universit at Munchen.

[25] W. Schultz. Predictive reward signal of dopamine neurons. *Journal of Neurophysiology*, 80:1–27, 1998.

[26] W. Schultz, P. Dayan, and P. R. Montague. A neural substrate of prediction and reward. *Science*, 275:1593–1598, March 1997.

[27] R. S. Sutton. Integrated modeling and control based on reinforcement learning and dynamic programming. In R. P. Lippmann, J. E. Moody, and D. S. Touretzky, editors, *Advances in Neural Information Processing Systems: Proceedings of the 1990 Conference*, pages 471–478, San Mateo, CA, 1991. Morgan Kaufmann.

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

[29] R. S. Sutton, D. Precup, and S. Singh. Between mdps and semi-mdps: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence*, 112:181–211, 1999.

[30] P. E. Utgoff and D. J. Stracuzzi. Many-layered learning. *Neural Computation*, 14:2497–2539, 2002.

[31] J. Wang, J. McClelland, A. Pentland, O. Sporns, I. Stockman, M. Sur, and E. Thelen. Autonomous mental develoopment by robots and animals. *Science*, 291:599–600, 2001.

[32] R. W. White. Motivation reconsidered: The concept of competence. *Psychological Review*, 66:297–333, 1959.