

CMPSCI 691BB Project Report:
A Framework for Intrinsically Motivated Agents

Siddharth Srivastava Shiraj Sen

Department of Computer Science,
University of Massachusetts, Amherst

13 May, 2005

1 Introduction

We propose a framework for creating intrinsically motivated autonomous agents that is based on the principle of homeostasis. The fundamental idea here is that organisms tend to pursue curious activity only when the basic needs such as food, drink, regulation of temperature or chemical balance etc have been taken care of, and the animal has achieved a state of “limbo”. McFarland [4] actually postulates drives that lead the animal into pursuing reproductive or survivalist goals, once the state of limbo is reached. From this point of view, intrinsically motivated activity can be grouped broadly into two categories: one satisfying the immediate homeostatic needs listed above, and the other, curiosity like activity which is performed once the homeostatic needs have been met. We present a framework that addresses both of these points. The approach involves a generic framework for homeostasis, and a module for curiosity which comes into play when being curious is fruitful, depending upon the internal state of the agent.

The rest of the report is organized as follows: The next section provides a short background on approaches most relevant to ours, dealing with intrinsic motivation. Section 3 describes the proposed framework for modeling a homeostatic agent. This is followed by section 4 which discusses two ways of incorporating curiosity within the framework. Section 5 notes some of the salient features of the model and brings out some observations not highlighted in earlier discussion. This is followed by a section on Future Work, and finally by our conclusions.

2 Background

Intrinsic motivation allows an agent to develop broad competence over its environment by allowing it to engage in behavior having no explicit external rewards. Such a motivational system would have to be governed by the agent’s internal state, which would also be responsible for the agent to improve upon itself.

Ideas based on intrinsic motivation have inspired approaches aimed at allowing agents to construct and extend their hierarchy of reusable skills [2] and to gain accurate descriptions of their environment [5], [6]. The approach presented by Barto, Chentanez and Singh [2] builds on the theory of *options*, by learning an *option policies* which direct the agent’s behavior for a subset of environmental states. Schmidhuber [5] tries to address the problem of understanding the external environment by generating curiosity rewards for the RL controller in response to predictor improvements.

3 The Homeostatic Framework for Intrinsic Motivation

Consider an MDP model \mathcal{M} , with a set of states \mathcal{S} , and associated action sets $\mathbf{A}(s)$. Each state of the environment can serve more than one purpose to the agent - and instead of

having separate utility-based incarnations of the states, we associate a k -dimension reward vector, $\mathbf{R}(s)$ each state. A state could thus be a good source of power while simultaneously being adverse in terms of temperature and only a mediocre source for light.

The agent maintains its own version of the state space, consisting of a k -dimension vector value function $\mathbf{V}(s)$ and a k -dimension internal *need* vector, \mathbf{D} . Components of \mathbf{D} represent the agent’s need state. The value for each component represents the percentage need for addressing that component. For instance, a typical \mathbf{D} could be

$$\langle SU, TempAdj, PowerAdj, LightAdj, GreaseAdj, HumanGoal, Curiosity \rangle$$

Here SU is a general purpose greatest priority need - and is intended for administering the agent, as in asking it to get out of the way irrespective of what its battery levels are etc. To make the SU mechanism effective, we enforce the SU component of the reward vector for *every* state to be 1. In the rest of the discussion, this component has very low significance, and its clean implementation is left as a pointer for future work. Here, we concentrate on building the framework on homeostasis and curiosity.

While the next few variables of \mathbf{D} are mechanism or domain specific needs and stand here for the percentage need for regulating temperature, power, light, and lubricant respectively, the variables $HumanGoal$ and $Curiosity$ have special meanings. The component $HumanGoal$ is 0 or 1, depending on whether the agent has been set a goal by the human or not. The $Curiosity$ component determines the fruitfulness of being curious; this is dependent upon the module used for curiosity, and is discussed later.

In addition, the agent possesses an inclination function, \mathbf{I} . The inclination is a mapping from $\{1, 2, \dots, k\}$ to \mathcal{R} , the set of real numbers. In practice, the inclination could be as simple as an ordering - it’s purpose is to appropriately amplify the need values depending on the agent’s priorities. In the rest of this paper, we assume that the inclination effectively provides an ordering on the values similar to the one outlined above. First comes the SU component, followed by the homeostasis components, which in turn are followed by the $HumanGoal$ component, and lastly by the $Curiosity$ component. In addition, we always place $Curiosity$ as the last component of the need vector and the reward vector. Of course, the most ethical ordering would be to make the human goal component of greater priority than homeostasis, but this is an agent with a limited action domain, and the availability of the SU functionality makes the current ordering safe enough while ensuring that the agent maintains functionality. However, changing this particular ordering does not significantly change the approach.

The components of the reward vector from the external environment correspond exactly with the components of the internal need vector. Since the reward function is k dimensional, the Value function \mathbf{V} is also k dimensional, and is initially set to random values for all the states.

3.1 External Goal

Human designated goals are searched in the usual RL fashion: the human sets a reward on the desired state. The only difference here is that now the reward has to be set in the appropriate (viz. *HumanGoal*) dimension of the external state. Once the *HumanGoal* component of \mathbf{D} has been set, the agent pursues this goal, subject to maintaining homeostasis, which in this case translates to maintaining sound working condition.

3.2 Value Update

For any external state s , the agent computes a *personal value* $V_p(s)$ as follows:

$$V_p(s) = \langle \mathbf{I}(1)\mathbf{D}_1, \dots, \mathbf{I}(k)\mathbf{D}_k \rangle \cdot \mathbf{V}(s) \tag{1}$$

where \cdot denotes the vector dot product. Because of the special nature of curiosity, we will sometimes compute the last component of these vectors (which we have set to curiosity) differently. This is indicated during the description of the appropriate Curiosity formulation; in general, the equation above serves as a good way of understanding the framework.

The agent performs a dualistic search following the RL paradigm in the state space. At each step, it performs the action that provides the greatest expected *personal value*. On the other hand, whenever it visits a state, it updates the state's Value function objectively, using TD with the reward vector it actually finds in the state. Each component of the Value vector \mathbf{V} is updated independently.

Interestingly, with this formulation the agent remembers states of homeostatic significance that it sees during unrelated exploration. So a state that provides power, if encountered during a search for cooler places will be remembered due to the objective updates to the value function V . Any neighboring states will eventually be updated with the appropriate discounted values, and thus a policy for a different, significant goal can be formed while the agent is searching for a more immediate need.

This could be seen as a way of developing *options*, and is discussed further in the Observations section (Section 5).

3.3 Agent's Activity

In this framework, the agent always tries to take steps that take it towards states which address its greatest need(s). This occurs in a nontrivial manner because during any period, the agent could be following a complicated pattern of addressing a weighted combination of its power, temperature and lighting needs while all the time just following the positive gradient of V_p .

At this stage, the agent already has a good working model of the homeostatic form of intrinsic motivation. In absence of external goals set by the human operator, it will explore the environment with a focus on addressing its homeostatic needs. This behavior is a function of the agent’s *inclination*. The next component we need is that of non-homeostatic curiosity, where the agent is interested in exploring the environment not for the immediate goal of achieving homeostasis, or external goal. However, as we will see, the approaches outlined below do bias the agent’s curiosity towards states that address the conjunction of its inclination and the need vector.

4 Curiosity

As already mentioned, curious behavior has its own priority or preference in this framework. As such, any good measure of the fruitfulness of being curious can be placed in the *Curiosity* component of \mathbf{D} . Correspondingly, the *Curiosity* component of a state’s reward function will reflect the extent to which the agent’s curiosity will be assuaged in getting to that state. We will call this component of the reward function the *curiosity reward*, and the curiosity component of the need vector the *curiosity drive*. Modularity in the framework allows the use of any good model of curiosity here.

There are two broad classes of methods for assigning values to curiosity rewards and drives. One is to use the reduction in transition probability variances as in Michael Duff’s [8] approach, or the TD errors as in the approach presented by Barto et al [2]. The second class of methods would be using the concept of Value of Perfect Information, as presented by Dearden, Friedman and Andre in [3]. In the next few sections, we present possible formulations for each of these classes of methods.

All of these approaches however rely on a good formulation of curiosity drive to be successful, and varying the nature of this variable can cause high level changes in the agent’s behavior. In this report, we will treat the reciprocal of average returns (the personal value obtained) over the past few steps to be the curiosity drive. This way, the agent becomes curious only when it is getting very low value pertinent to its internal state. For instance, if the agent is facing a low power situation, then its power “need” will be high; this will bias its V_p to prefer states that are good in terms of power supply; the extent of its curiosity drive will be proportional to the inverse of the average reward that it receives while performing its exploration. It is likely that during this exploration the agent comes across states that provide really high rewards albeit for different, and unimportant components - but then the % need of the agent for these other components will be low, resulting in low personal value from those states, and the agent will therefore be driven to be curious, and rightly so.

This formulation of curiosity drive is quite desirable: within this framework, it presents an interesting kind of behavior. When the agent is in “limbo”, naturally the curiosity component will influence action selection, and thus the policy. However, even if the agent

has an external goal that it is chasing, but is not able to make significant progress, the agent will have phases of curiosity. This is productive because given a good curiosity reward formulation, the agent could learn vital new information through the curious behavior. Moreover, this information is likely to be relevant because of the vector valued Reward and Value structure. For instance, if the agent is “bored” searching for a particular square in the grid world, and it transitions to the curiosity phase, it is still going to notice the external goal state if it comes across it; in fact, on seeing the goal state the external reward would jump, turn down the curiosity drive, and get the agent back on track! Even if the reward is not strong enough to significantly turn off the curiosity drive, it will in any case bias the V_p so that even with curiosity the agent keeps preferring the goal state and eventually the average reward increases and the curiosity drive goes away.

We feel this is a good, “positive attitude” to have in an autonomous agent designed to work with people with limited ability. Instead of getting passively bored when the external goal set by the human is too high, it gets pro-actively bored, and starts exploring with an interest towards the actual assigned goal.

4.1 TD Curiosity

In order to use the temporal difference error (or improvement in TD error) as a curiosity reward, the agent maintains another vector per state, $\mathbf{TD}(s)$, representing the most recent TD error found on visiting the state s . This is a $k - 1$ dimensional vector, representing the TD errors on all dimensions except the last, which was set aside in the reward and need vectors for curiosity. The agent now computes a personalized TD error:

$$TD_p(s) = \langle \mathbf{I}(1)\mathbf{D}_1, \dots, \mathbf{I}(k-1)\mathbf{D}_{k-1} \rangle \cdot \mathbf{TD}(s) \quad (2)$$

TD_p is biased so that the components with greater needs and inclination have a greater say.

In the sequel, the \mathbf{TD} vector will actually contain the *improvements* in TD error. This serves as a good guard against getting sucked into unpredictable states. Computing and storing the TD error vector, or the vector with improvements in TD error is not a significant computational overhead because the Value functions are already being computed independently for each dimension using the TD method.

This error (or its improvement) has to be used as the curiosity reward. There are two ways of accomplishing this: one would be to add this error as the last component of each state’s reward vector:

$$\mathbf{R}'(s) = \langle \mathbf{R}_{\bar{k}}(s), TD_p(s) \rangle \quad (3)$$

Where \mathbf{R}' is the new reward function, and $\mathbf{R}_{\bar{k}}(s)$ represents the vector having the first

$k-1$ components of $\mathbf{R}(s)$: $\mathbf{R}_{\bar{k}}(s) = \langle R_1(s), \dots, R_{k-1}(s) \rangle$. $\mathbf{R}'(s)$ is thus the original reward vector of state s , with the curiosity reward component set to $TD_p(s)$.

With this formulation of the curiosity reward, the agent calculates personalized value of a state s as before:

$$V_p(s) = \langle \mathbf{I}(1)\mathbf{D}_1, \dots, \mathbf{I}(k)\mathbf{D}_k \rangle \cdot \mathbf{V}(s) \quad (4)$$

Note that this formulation does not lead to a circularity in curiosity: that is, the agent does not end up being curious about curiosity because we left it out of the calculation for curiosity reward, TD_p .

However, there are two notable side effects of this formulation:

- The reward structure of the original MDP is now dynamic. This can lead to complications in offline methods of solving MDPs.
- The positive side effect is that now, even when not being curious, the agent will be developing a path, or a policy to states with curiosity reward because the single step reward vector update and the value function calculation occurs on all dimensions, including the one for curiosity.

Another formulation of a curiosity reward could be to add it in the Value function instead of the reward function. In this case, the calculation of V_p is simpler:

$$V_p(s) = \langle \mathbf{I}(1)\mathbf{D}_1, \dots, \mathbf{I}(k)\mathbf{D}_k \rangle \cdot \langle \mathbf{V}_{\bar{k}}(s), TD_p(s) \rangle \quad (5)$$

Again, $\mathbf{V}_{\bar{k}}(s)$ represents the vector having the first $k - 1$ components of $\mathbf{V}(s)$. In the dot product, for the value function of the state we use the first $k - 1$ components of the original value function and TD_p as the curiosity reward.

With either of these curiosity reward formulations, the agent has a desirable, directed form of curiosity. To add to the comments in section 4 on the nature of its curiosity, the agent now has a “boredom with success” attitude. Without this formulation of the curiosity reward, the agent could get stuck in situations where there is an external reward that it knows how to obtain. The curiosity *drive* would be really low, because of a high average reward. However, with the improvement in TD error incorporated in the curiosity *reward*, the personalized value function will also bring those states into consideration which have better TD improvement rewards, but are low on the explicit external, “easy” reward.

The framework thus incorporates a version of the Optimal Level Theory [1] and the Autotelic Principle [7]. It tends to turn away from tasks that are either too challenging or tasks that it knows well how to accomplish. There is a middle level of tasks that keep it engaged, which evolves with respect to its internal state, the external goal provided by the human and the spare time that it gets when the human has not assigned it any goals. Interestingly, during these phases of turning away, rather than being inactive or waiting for

hints or doing something completely different, it engages itself in behavior aimed towards understanding parts of environment that are relevant to its internal state and external goal.

4.2 Curiosity for Perfect Information

Curiosity based on a search for perfect information is a very rational approach - the agent is curious to the extent that it can gain as accurate information about its environment as possible, and consequently follows actions that have a possibility of improving this information state. From the point of view of an intrinsically motivated agent, a shortcoming of this approach as presented by Dearden et. al. is that its curiosity is undirected. An agent following this approach would be curious about the environment, irrespective of its internal state. This is a serious problem because we would like the agent to be autonomous enough to take care of its functional fitness - and certainly not drain its power supply being curious even if it has no other explicit, external goal.

We propose to address this problem by using the values V_p in the computation of the VPI - the “gains” used in the VPI calculation now become subject to internal state, and we have a good model of directed, rational curiosity.

Another problem with using the VPI approach directly in our framework is that it’s reward and the VPI are functions of state and action pairs. In our approach, we would like to associate the VPI and rewards to just states. This problem is easily solved by a simple transformation (that leads to an increase in state space, that is linear in $|\mathcal{S}|$ and $|\mathcal{A}|$). For each state and action pair (s, a) that rewards are associated with, we create new *commitment* states $\langle s, a \rangle$. A new MDP with states of the form s and $\langle s, a \rangle$, reward function \mathbf{R}' , transition function T' and action sets \mathbf{A}' is created as follows:

$$\begin{array}{ll}
 \textit{NewMDP} & \textit{OldMDP} \\
 \mathbf{R}'(\langle s, a \rangle) & \leftarrow \langle \mathbf{0}_{\bar{k}}, VPI(s, a) \rangle \\
 \mathbf{R}'(s) & \leftarrow \langle \mathbf{R}_{\bar{k}}(s), 0 \rangle \\
 \mathbf{A}'(s) & \leftarrow \mathbf{A}(s) \\
 \mathbf{A}'(\langle s, a \rangle) & \leftarrow \{a\} \\
 T'(\langle s, a \rangle, a) & \leftarrow T(s, a) \\
 T'(s, a) & \leftarrow \{\langle s, a \rangle\}
 \end{array}$$

($\mathbf{0}_{\bar{k}}$ represents a k-1 dimension zero vector)

To summarize what is happening here, the reward function for the new MDP, \mathbf{R}' has a curiosity reward (viz. VPI) associated only with the commitment states. And for these states, the reward along all other components is zero. For all the original states s , \mathbf{R}' is the

same as the old reward function, and gives zero curiosity reward. The set of actions for the original states remains the same; the commitment states have only the committed action available. Correspondingly, the transition probabilities now reflect the fact that it is only possible to go to a commitment state $\langle s, a \rangle$ once action a is chosen in state s . From $\langle s, a \rangle$ however, the transition probabilities are exactly equal to the corresponding probabilities in the old MDP, if action a was chosen at state s .

Without curiosity, this transformation preserves optimal policies because all the real next states corresponding to all possible actions get the same extra discounting.

In this formulation of the state space with VPI and curiosity, the policy of the agent is again determined by equation (1). Only now, we also have *Curiosity* need as formulated above, and the VPI as the curiosity reward at the commitment states. The agent thus treats *Curiosity* as any other need; this need has its inclination value just like the other needs (we placed it last in priority..) and is rewarded by visiting states with high values of VPI.

4.2.1 Not So Perfect Complexity

Unfortunately, the method outlined above in its current state has to pay for its extremely rational philosophy through its complexity. The *VPI* method relies on a prior distribution on all the state transition probabilities and the reward probabilities. To obtain the VPI estimates for each state action pair, the method has to perform sampling - and in our case, this is more complicated because of the variable nature of V_p 's, which are determined by the agent's needs. Although this approach is promising and seems to provide a good theoretical basis, an elegant formulation of the trade-offs that the agent should make, and meshes quite well with the framework, it needs further work in order to address the complexity issue.

5 Observations

In this section, we list some of the key points about the behavior of an agent following this approach that might not have been stressed earlier. We also discuss the relevance of this approach with the particular situation presented in the question for this project.

- The approach presents a method for making external reward a function of the internal state, without extra iterations due to changes in the internal state
- Value estimates are updated on all dimensions. This means that the agent would, for instance, notice a fan (a temperature reduction state) while focusing on looking for sources of power.
- Policies developed along each dimension can be seen as **Options**. They represent sub-goals and sub-policies different from the one for the assigned external goal. For instance, the initiation set for the option of searching for a light source would be

the set of internal states when the light need, in conjunction with its inclination, overshoots the other needs. To this extent, the framework provides a method for learning options pertaining to the pre-assigned needs in the agent's need vector.

- The approach is interruption safe: if suddenly the temperature goes up due to working near the radiator for some time, the agent's need for addressing the temperature goes up. To address it the agent will prefer states that provide better rewards in that vector, or in other words provide cooling. Once this is done, the requirement drops again. Since the other internal parameters don't change during this period, at the end of this phase the agent starts chasing the goal it was chasing before the interruption. The independent value functions ensure that the history of previous exploration on the other dimensions is not lost.
- The approach also presents a nontrivial manner of employing gradients instead of thresholds as proposed by McFarland [4]. It is argued there that a threshold based behavior for addressing needs such as power supply is too short sighted, and does not take into account the gradient of power loss during an activity. Our approach generalizes their solution. In looking for optimal value, the agent implicitly chases the most needed goal according to the homeostatic ordering. The desirability of a state that provides power will increase as the need component corresponding to power increases. That is, states that provide good power will be considered more and more rewarding as the need for power increases. States that provide power and another desired attribute, temperature regulation for instance, in addition to power supply, will be preferred over states that provide rewards on some other two attributes. This will lead to a complicated pattern, while ensuring that the next action always optimizes on at least the most needed reward.
- This approach presents a generalized solution to the problem suggested for the project. In particular, the initial, exploration phase for this approach is just the situation where the external reward is not set by the human operator. The approach takes care of homeostasis and curiosity, enabling the agent to learn both in the beginning during such a phase, and during any other phase when either the external goal is absent, or the external goal is beyond reach (leading to a high curiosity drive).

6 Future Work

The homeostatic framework and the curiosity model which has been presented above for intrinsically motivated agents is at present a theoretical model which seems to exhibit several of the desirable properties found in intrinsically motivated organisms, while balancing the learning among all the goals. However, experiments are needed to be carried out to see how tractable it is to simultaneously learn these k -dimensional value vectors. Moreover, future work should involve making the model flexible enough to break a goal into a number

of sub-tasks and learn the value functions for these primitive tasks, so that if it is provided with a new goal by a human which consists of piecing together a number of previously learned sub-tasks, it can re-use its knowledge bank without re-learning a completely new policy for this task. The agent should be learning the policies for only the primitive actions which make up a majority of the tasks. The next stage of learning should involve learning the relations between these policies i.e., the transition probabilities among these policies. This will be helpful, for once a task is provided, the agent can identify the sub-tasks involved, and then chart out a new policy which is a sequence or linear combination of the policies learned so far, in order to accomplish the task.

In particular, it would be interesting to iron out the details in the Curiosity approaches outlined above - for instance, what should be the optimal number past steps that the curiosity drive should consider? Which of the TD error methods will work better in practice? Are there better representations of the curiosity drive, or fruitfulness of being curious? We hope to address these questions better with a better understanding of RL and associated literature.

7 Conclusions

Any approach that includes intrinsic motivation would have to include an internal state dependence of the “reward”, and a prioritization of the needs that motivate it. In this report we presented a framework for doing this, and some ideas for incorporating “relevant” curiosity within the framework. The approach appears promising and robust because of its similarity to evolutionarily selected behavior observed in intrinsically motivated organisms. We look forward to refining this approach, and studying these ideas more carefully.

References

- [1] H. R. Arkes, J. P. Garske, “Optimal Level Theories”, *Psychological Theories of Motivation*.
- [2] A. Barto, S. Singh, N. Chentanez, “Intrinsically Motivated Learning of Hierarchical Collection of Skills”, *ICDL*, 2004.
- [3] R. Dearden, N. Friedman, D. Andre, “Model based Bayesian Exploration”, *Proceedings of Uncertainty in Artificial Intelligence*, 1999.
- [4] D. McFarland, T. Bosser “Intelligent Behavior in Animals and Robots”.
- [5] J. Schmidhuber, “Self-Motivated Development Through Rewards for Predictor Errors / Improvements”, *Developmental Robotics - AAAI Spring Symposium*, 2005

- [6] J. Schmidhuber, “ Exploring the Predictable”, *In Ghosh, S. Tsutsui, eds., Advances in Evolutionary Computing, p. 579-612*, Springer, 2002.
- [7] L. Steels, “The Autotelic Principle”, In Fumiya, I. and Pfeifer, R. and Steels, L. and Kunyoshi, K., editor, *Embodied Artificial Intelligence, Lecture Notes in AI (vol. 3139), pages 231-242*, Springer Verlag. 2004.
- [8] M. Duff “Design for an Optimal Probe”, *Proceedings of ICML*, 2003
- [9] R. Sutton, A. Barto, “Reinforcement Learning : an introduction”, MIT Press, Cambridge, MA, 1998.