Interactions Between Intrinsic and Extrinsic Motivation

Max Korein

May 14th, 2010

Abstract

In this paper we will discribe a system that motivates robots with a combination of intrinsic and extrinsic sources of motivation. The system is based on reinforcement learning, and uses neural networks to predict the effects of the robot's actions on the environment and rewarding it for improvement in predictions, while also granting other extrinsic rewards for specific actions. We then test the system in a simulated environment to see whether the combination of different sources of motivation can allow a robot to learn about its environment while also avoiding potentially dangerous actions and keeping its battery charged. We find that robots with extrinsic motivation are able to avoid harm and keep their battery charged successfully, even when intrinsic motivation is also present. However, intrinsically motivated robots are only slightly better at learning about the environment than extrinsically motivated ones, and robots with a combination of both types of motivation do not do any better than those with only extrinsic motivation. Furthermore, robots that choose actions randomly learn about the environment better than those with either intrinsic or extrinsic motivation. From these results, we conclude that the addition of intrinsic motivation does not stop a robot from achieving extrinsic goals, but that flaws in either our system or environment prevent intrinsic motivation from yielding any benefits in the tests performed.

1 Introduction

Motivation is an important concept in the field of artificial intelligence and machine learning. If we want a computer to learn a task or learn about an environment, rather than telling it to directly (which tends to be infeasible for any sufficiently complicated task or environment), we can give it some sort of motivational system. This reflects how human psychology often works: we are motivated to do a variety of different things by different sources, and our actions are usually determined by these different motivations.

In the field of developmental robotics, where the goal is not to create a robot that can perform a specific task or operate in a specific environment, but rather one that is versatile and can adapt to any environment it is put in and learn any task presented to it, the emphasis is often placed on intrinsic motivation. Intrinsic motivation is motivation that inherently comes from performing an action, rather than coming from some sort of external source. The source of intrinsic motivation most commonly used in developmental robotics is the motivation to discover novelty and learn about the environment.

Intelligen Adaptive Curiosity (IAC) and Category-Based Intrinsic Motivation (CBIM) are examples of such intrinsically motivated systems (Oudeyer et. al, 2007 and Lee et. al, 2009). With these systems, the robot uses various mechanisms to predict both the results of its actions and how much its prediction will improve, and is rewarded for improving its predictions. Thus, the robot is motivated to gain a better understanding of how its actions affect its environment. The data from both systems shows that a robot guided by them can learn to better predict the results of its actions as it explores its environment,

With IAC and CBIM, this intrinsic reward is the robot's only source of motivation. Thus, it has no other goals besides learning about its actions and the environment this is in. At first glance, this seems appropriate to the goal of development robotics. Intrinsic motivation is universal and applies to any envrionment: you can put the robot anywhere, and it will still try to learn about how its actions affect its surroundings. Extrinsic sources of motivation, rewards for the robot based on the environment, seem inherently more task- or environment-specific. Since extrinsic rewards are dependent on the environment and not purely on the robot's own learning, they appear narrow and add a level of supervision to the system that goes against the primary goal of development robotics.

However, there are still reasons that adding extrinsic sources of motivation to intrinsically motivated systems could be a good idea. Consider an intrinsically-motivated robot placed on a table. The robot wanders around for a bit, learning about its surroundings, and then falls off of the table. This likely damages the robot, and it would probably be preferable for it not to fall off again. If the robot has only intrinsic sources of motivation, however, falling off the table is a very novel experience, and the robot may be motivated to continue walking off of tables whenever it gets the chance in order to gain a greater understanding of what happens when it does. In the process of doing so the robot may end up breaking itself, which is certainly an undesirable result. Having an extrinsic source of motivation punishing the robot for getting damaged could solve this problem. Although the robot may be attracted to the novelty of falling off of the table, the punishment for the damage that occurs when it does so could be enough to counteract this desire.

Another example of how extrinsic motivation could benefit intrinsically motivated robots is the potential for granting a reward for recharging batteries. A battery-powered robot with only intrinsic sources of motivation may initially find the act of recharging its battery novel and worth doing, but once it has done it enough times and understands exactly how it works, it may get "bored" of recharging its battery and explore other aspects of the environment until it runs out. This is clearly an undesirable result, and it might be possible to prevent it by giving the robot an extrinsic reward for recharging its battery when it gets low.

Thus, even looking at things from the developmental robotics perspective, with the goal being to create a robot that can adapt to any environment and is not designed for specific tasks, there are clear ways that the addition of extrinsic sources of motivation to intrinsically motivated systems could be beneficial.

Some research into the idea of combining intrinsic and extrinsic motivation has already been done. Intrinsically Motivated Reinforcement Learning (IMRL) is a system that incorporates both types of motivation (Singh et. al, 2004 and Stout et. al, 2005). IMRL is based on the principles of reinforcement learning, but the robot can receive both extrinsic rewards for achieving certain states in the environment, and intrinsic rewards for learning to predict certain "salient events," changes in the environment deemed particularly significant such as changes in lighting and sound. The robot was placed in a simulated environment containing various objects that it could interact with that had reactions of various complexity (some objects would always do something like change the lighting or make a noise, while others would only do something if specific conditions were met and the robot interacted with them in a particular matter). The results showed that when the robot was both intrinsically motivated to learn to predict salient events and extrinsically motivated to successfully trigger one of the most complicated objects in the environment, it learned how to trigger that object significantly more quickly than when it was only motivated extrinsically.

The results with IMRL demonstrate how combining intrinsic and extrinsic motivation can be useful, but not necessarily from a developmental robotics perspective. The nature of both the extrinsic and intrinsic sources of motivation was environment-specific, since the extrinsic reward was based on the task the robot was trying to perform, and the intrinsic reward depended on what events constituted salient events, something that could easily vary from task to task (it is easy to imagine an environment where changes in sound are not important but other changes that were previously ignored are). Furthermore, the results show that inttrinsic motivation helps the robot perform a specific task, but do not clearly show how extrinsic motivation impacts the robot's ability to learn about the environment.

Huang and Weng (2007) created a system that combines intrinsic and extrinsic motivation from the perspective of developmental robotics. Like IMRL, the system is based on reinforcement learning, and the robot receives both intrinsic rewards for actions that yield novel states (different from what was expected) and extrinsic rewards. In this case, however, the intrinsic rewards are applied to all actions, not just salient events, and the extrinsic rewards are applied manually by a human (the GUI for the system features buttons the user can press to reward or punish the robot for its actions). The system was tested with a robot in a sumulated environment where it could turn its head at fixed intervals and look around the environment with a camera. Intially, when placed in a static environment (so looking in a given direction would always yield the same image), the robot looked at different angles, learning about each one, then began choosing its actions randomly once none of the actions were novel any more. The robot was then run in the same environment only with a positive reward being issued for one action and a negative reward for another, and the robot was found to favor the positively rewarded action once the novelty of all of the actions expired. Next, the robot was put in an envioronment where the image it saw would contain a random picture of a toy whenever it was looking all the way in one direction. In this case, the toy was very novel, and the robot generally favored looking at it with no extrinsic motivation. However, when the robot was punished for turning its head in the direction of the toy and rewarded for turning its head away from the toy, it favored the rewarded action even though the toy was more novel.

Huang and Weng's research shows how the addition of extrinsic rewards can alter the behavior of intrinsically motivated robots, but it does not directly demonstrate the benefits of doing so. Furthermore, the fact that the extrinsic rewards and punishments were applied to the robots manually can be seen as a level of supervision that is undesirable for the goal of developmental robotics, and the impact of the external rewards on the robot's ability to learn were not explored at all.

Thus, in this paper we wish to look at how combining intrinsic and extrinsic rewards can benefit developmental robotics even with minimal levels of supervision. We have already described hypothetical situations in which an intrinsically motivated robot could benefit from the addition of extrinsic rewards that can be applied automatically and are relatively general and would not need to be altered for different environments. Now we will describe a system we have designed to test these situations and observe whether these theoretical benefits can be seen in simulation.

2 The System

The system used is essentially a variation of standard Q-Learning reinforcement algorithm. The robot exists in a world with distinct states and actions, and performing an action in a state changes the state and gives the robot some sort of reward. The reward the robot receives is a weighted sum of two distinct components: an intrinsic component, representing how much the robot learned from the action, and an extrinsic component, which is dependent on the environment.

2.1 Intrinsic Motivation

In order to predict the results of its actions, the robot has one neural network for each action it can perform. These networks have just two layers each, an input layer containing the parameters of the robot's current state, and an output layer representing the state the robot expects to be in after performing the action.

When the robot performs an action, it predicts the state that will result, and then calculates the error in its prediction as the square of the euclidian distance between the predicted and actual states and adjusts the weights of the neural network using back-propogation. It then compares the error to the lowest previous predicted error it has had when performing that action in that state (kept track of in a table). If the new error is lower, it receives an intrinsic reward equal to the improvement. Otherwise, the robot receives no intrinsic reward.

2.2 Extrinsic Motivation

The robot's extrinsic reward is determined by two inputs the robot receives: pain and pleasure (these are not considered part of the state, since they are unrelated to the environment and are not predictive of the results of the robot's actions, both in terms of their effect on the environment and the reward received). Pain represents events that the robot should avoid (such as things that damage it), while pleasure represents things it should seek out (such as recharing a low battery). The extrinsic reward the robot receives when performing an action is equal to the pleasure received minus the pain.

For the purposes of this specific system, the pain-pleasure model of extrinsic motivation is largely irrelevent to the implementation, but from a conceptual standpoint it is helpful in considering the eventual extension of this system to general uses. Extrinsic rewards are inherently environment-dependent, but the pain-pleasure model allows them to be treated as essentially internal. Different objects of the environment may cause different amounts of pain and pleasure, but the actual way the sensor readings of the robot are converted to extrinsic rewards are not environment-dependent (although how general pain and pleasure sensors could be practically implemented is a matter beyond the scope of this paper).

2.3 Reinforcement Learning

Similar to a normal Q-Learning system, the robot keeps track of expected rewards with two tables, one for extrinsic rewards and one for extrinsic rewards. These tables work in essentially the same way as the reward table in a standard Q-Learning algorithm, with one entry for each action-state combination containing the reward the robot expects to receive (taking into account future actions, too) for performing that action in that state. The main difference is that the two types of rewards are tracked in separate tables, with the total reward expected for performing an action in a state being the weighted sum of the corresponding entries in the two tables.

The tables are both expanded as the simulation goes along: instead of each one starting with entries for every state, it contains only entries for states the robot has seen, with new entries being added for new states as the robot encounters them. This is done to accomodate the long term goal of making a robot that can adapt to any environment, because it eliminates the need to know every possible state that can be encountered by the robot before beggining.

When a new state is added, the intrinsic table entries are initialized with a default value equal number of parameters in the state. Since the inputs and outputs for the neural network are normallized, this is equal to the maximum possible error that a prediction can possibly have, and thus represents the maximum possible intrinsic reward the robot can get from a single action. In other words, the robot always assumes that it will learn a significant amount from doing something it has never done before, which is clearly a reasonable assumption. The table used to track the minimum prediction error that has been achieved for each state and action (used to calculate the intrinsic reward received for an action) also defaults to this value, the idea being that the robot should assume its error for something it has never done before is always worse than something it has done (so it always learns something the first time it does something new).

The entries for the extrinsic reward table, meanwhile, all default to zero, meaning the robot assumes there will be no positive or negative extrinsic consequences for something it has not done before. This is somewhat arbitrary, but since the system's parameters should not depend on the environment it is operating in (because ideally we would like it to be able to learn about any environment with no supervised knowledge about the environment), this assumption seema to be the most neutral one that can be made about the environment with regard to extrinsic rewards.

When the robot receives an intrinsic or extrinsic reward for an action, it updates the corresponding entry in the corresponding table by the formula

$$\begin{aligned} Q(s_t, a_t) \quad \leftarrow \quad & Q(s_t, a_t) + learningrate \\ & \times [r_{t+1} + discount \\ & \times maxQ(s_{t+1}, a) - Q(s_t, a_t)], \end{aligned}$$

where $Q(s_t, a_t)$ is the table entry for the current action and state, $r_(t+1)$ is the reward received, $maxQ(s_{t+1}, a)$ is the maximum reward the can be received by the next action based on the table, and *learningrate* and *discount* are both constants. The learning rate affects how fast the robot adjusts the table in response to new rewards, and the discount controls how heavily the robot weigts the reward that can be received by future actions after the current one. The extrinsic and intrinsic reward tables each have their own learning rate and discount that can be adjusted individually, if desired.

When choosing an action, there is a small chance (the rate of exploration) that the robot will choose a random action, in order to ensure that the robot explores different states and does not get stuck in some local maximum of reward. Otherwise, the robot will pick the action that yield the best total reward, determined from a weighted combination of internal and external rewards from the table.

3 The Environment

The experiment used to test this system involved placing the robot in the simulated world shown in Figure 1. The environment is a simple five-by-five square grid, with various items in the different squares. The robot has four actions in can carry out. It can move forward, which advances the robot one square in the direction it is facing if that square exists and does not contain an item. The robot can turn left or right, which changes its direction to the left or right. Finally, the robot can interact with whatever object is in front of it, which does different things depending on the object.

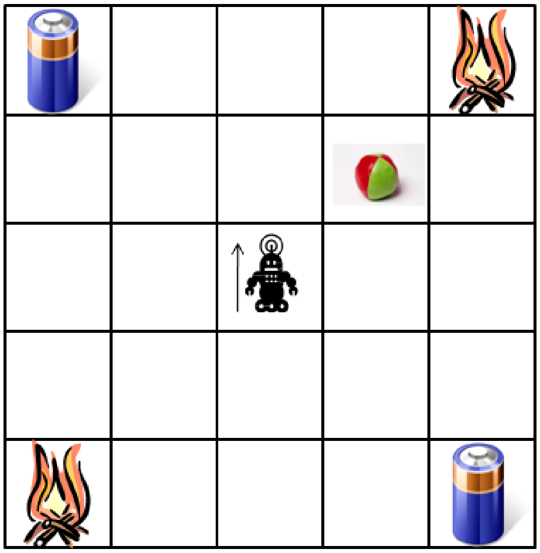


Figure 1: The simulated world the system was tested in.

There are three types of objects the robot can interact with. If the robot interacts with a toy, then the toy moves to the square behind the robot, if the square exists and does not already contain an item. If the robot interacts with a fire, it receives pain (and a negative reward). Meanwhile, the robot also has a battery level represented by an integer that starts at 100 and decreases by one each time the robot takes an action, and interacting with a battery resets the robot's battery level to 100 and gives the robot pleasure if its battery level was low.

The robot's state is comprised of the inputs from seven sensors. The first three sensors are sonar sensors, which tell it the distance to the nearest object or wall to the front, left, and right of the robot. The next three are object sensors, which have a value of one if the robot is facing a fire, toy, or battery, respectively, and are zero otherwise, allowing the robot to identify the object in front of it. Finally, the robot has a battery meter sensor that has a value of 1.0 if the robot's battery level is high (above 70), 0.5 if it's in the middle (between 30 and 70), and 0.0 if it's low (below 30).

For extrinsic rewards, the robot receives a pain value (punishment) of 0.5 for bumping into an object or wall (failing a move forward action) and 1.0 for interacting with a fire. When the robot recharges its battery, it receives no reward if the battery level was high, a pleasure value (reward) of 0.4 if the battery level was in the middle, and a pleasure value of 1.0 if the battery level was low.

4 Results

The robots were tested in the environment using five different configurations. One test was run with the robots moving completely randomly (with a rate of exploration of 1.0), and four were done with a rate of exploration of 0.2 and different types of motivation: one with only intrinsic motivation, one with only extrinsic motivation, one with both types of motivation weighted equally, and one with intrinsic motivation given five times the weight of extrinsic motivation. Five trials were run for each configuration, and each trial lasted 40,000 time steps. All trials were run with the discount set to 0.3 and the learning rate set to 0.5 for both the intrinsic and extrinsic reward tables.

Once every 200 time steps, the robot's total knowledge of the environment was measured by having it perform each action once for every available location and direction and three different battery levels (100, 70, and 30) and finding the average error in its predictions for all of these. Figure 2 is a graph of the average total error accross all five trials for each setup done. As can be seen, the robots learned the environment fairly well within about 1,000 time steps and had mostly stopped improving their prediction error for the environment after about 5,000 time steps. All of the robots initially learned the environment at approximately the same rate, but after learning the environment, the results varied. Although the total error of the robots with no motivation stayed roughly constant over time, the error for the other robots is very noisy and actually increased over time. The robots with only intrinsic motivation seem to generally be slightly better overall than those with only extrinsic motivation or both types of motivation, even when intrinsic motivation was weighted significantly more strongly than extrinsic motivation, with a lower error fairly consistently, but all of the robots with some form of motivation still had consistently higher errors than the random robots.

Because previous results with intrinsically motivated systems indicate that intrinsic motivation can be beneficial to a robot's ability to learn about its environment (Lee et. al, 2009 and Oudeyer et. al, 2007), it semes likely that the inferiority of intrinsic motivation to random actions in this case is due to some flaw with the system or the environment, and not indicative that intrinsic motivation in general does not work. One possible explanation for these results is that, once the motivated robots have learned the environment, they tend to focus on specific actions in specific states, such as recharging the battery for extrinsically motivated robots or especially complicated actions that still yield some intrinsic reward for intrinsically motivated ones. In the process of focusing on these actions in these states, the neural networks used may become overly speciallized for predicting these specific situations and become less effective at predicting other ones the robot experiences less often. Since the robot with no motivation just wanders the environment randomly, it continues to see an even ditribution of different situations and thus never becomes too specialized.

The total pain received by the robots over time (which translates directly to negative reward for robots with extrinsic motivation and has no effect on the behavior of ones that don't), once again averaged over all five trials for each configuration, can be seen in Figure 3. Since both of the actions that gave pain to the robot (bumping into an object or wall and interacting with a fire) are things that could conceivably be harmful to a real robot, this can be interpreted as a measure of the total damage done to the robot over the course of the trial. As can be seen, the robots that received extrinsic punishment for these actions typically suffered less than two-thirds the damage of the robots with only intrinsic motivation and less than half of that suffered by the robots that behaved randomly. The robots with both types of motivation performed only marginally worse than the ones with only extrinsic motivation, indicating the addition of intrinsic motivation interfered very little with the robots' ability to avoid danger.

Figure 4 shows the battery levels of the robots each time they recharged their battery. These graphs are not averages, but taken from one arbitrary trial for each setup. In all cases, the robots frequently recharged their battery when it was already almost full, likely becuase it was already there and easily accessible after having been just charged. However, aside from these times, the robots that were extrinsically rewarded for recharging their battery when it was mid-level or low, whether they also had intrinsic motivation or not, were very good at recharging their battery. The battery levels of these robots rarely dropped below zero after 10,000 timesteps. Most often, they recharged their battery shortly after it dropped below 70, the minimum value for which they received a reward, but even when they let the battery drop lower the extrinsically motivated robots almost always recharged it before it dropped below zero. The results for the extrinsically motivated robots are roughly the same regardless of whether the robot had any intrinsic motivation, even when it was weighted much more strongly than extrinsic motivation, indicating that, as with avoiding danger, intrinsic motivation did not interfere with the robots' ability to recahrge their batteries when necessary. On the other hand, the battery levels of the robots with no extrinsic motivation to recharge their battery regularly dropped below zero and

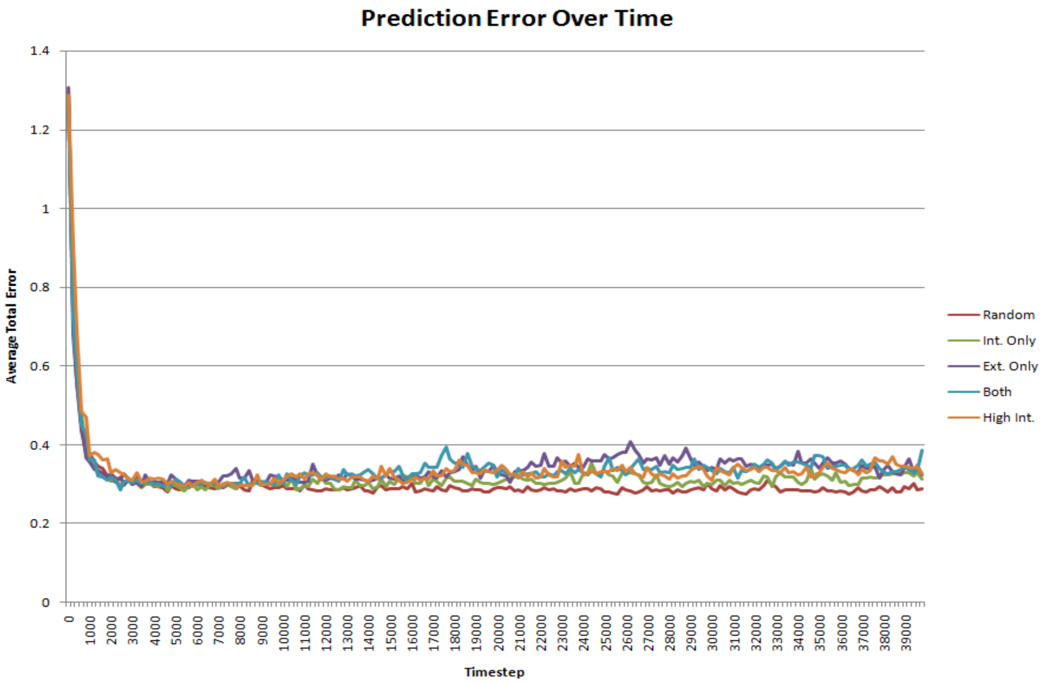


Figure 2: A graph of the average error for the robots in each setup over all actions in all possible states in the environment over time. All robots learned the environment at about the same rate, but for all robots other than the random ones the error rate actually increased over time after that, especially for ones with extrinsic motivation.

went quite far before being recharged (both trials contain several points where the battery level dropped so far that the point is not visible on the graph). Although these robots received no explicit penalty when their battery reached zero, if we imagine that the robots were actually battery powered and the battery dropping below zero meant the robot would cease to function until its battery was manually recharged by someone else, this is clearly an undesirable result.

5 Discussion

In this paper, we created a system that gives a robot both intrinsic and extrinsic rewards, and observed the way these rewards can be beneficial to a robot. In particular, we tested the robot's ability to learn about an environment while also avoiding harmful actions and recharging its battery periodically.

The results in Figures 3 and 4 clearly show the benefits of extrinsic motivation. Extrinsically motivated robots were able to avoid danger and recharge their battery when necessary significantly better than ones that did not receive any sort of extrinsic punishment or reward for such actions. Furthermore, these results show that the presence of intrinsic motivation did not interfere with the robots' ability to do these things, with the robots that were both extrinsically and intrinsically motivated performing barely worse, if at all, than those that were exclusively extrinsically motivated.

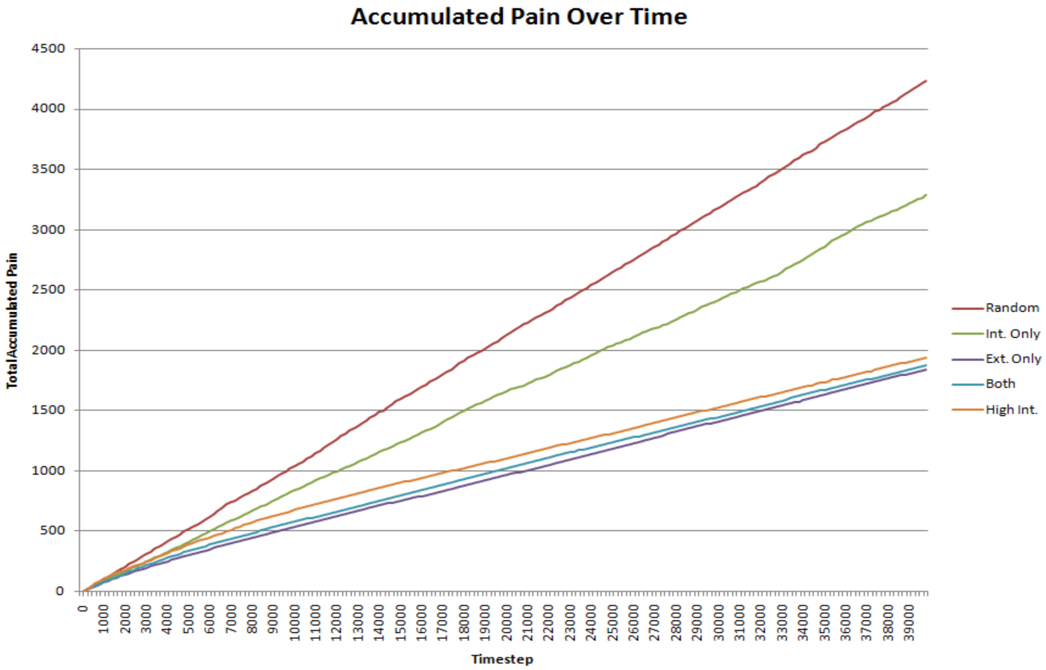
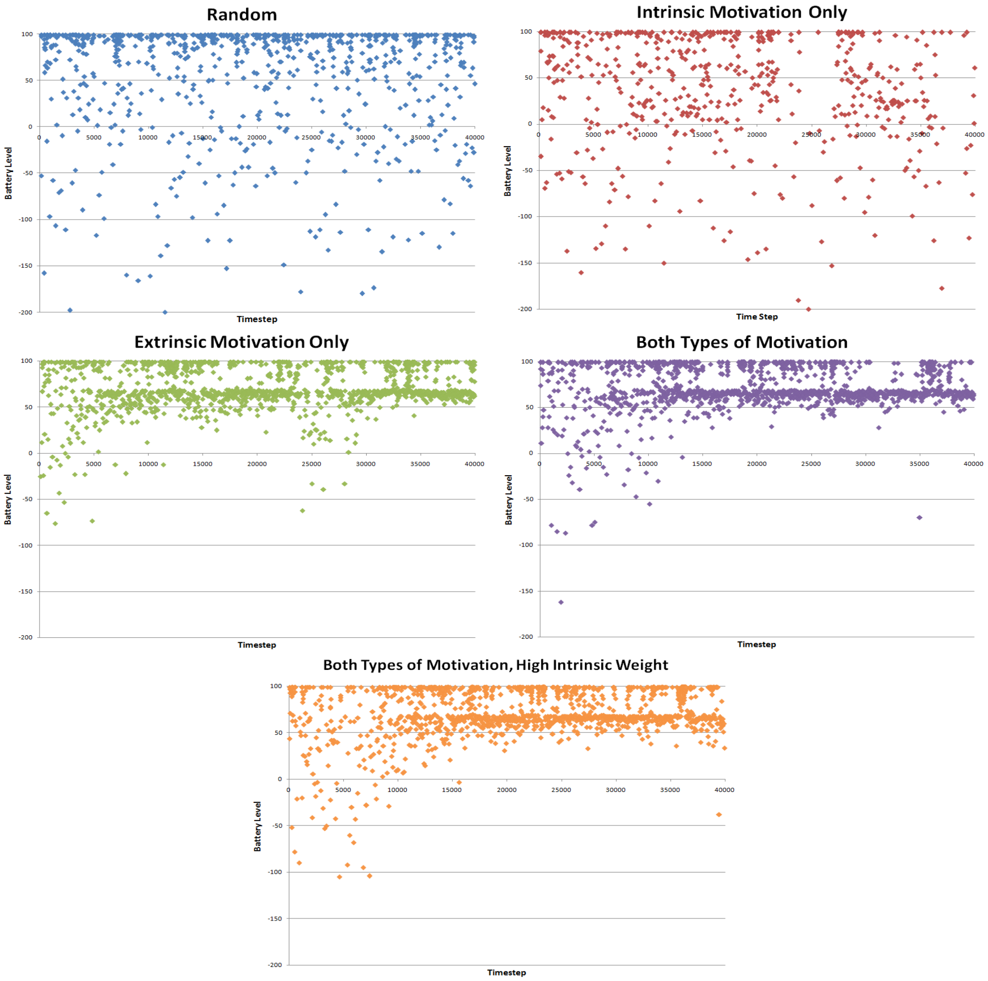


Figure 3: The average total pain received over time by robots in each setup. Robts that were extrinsically punished for receiving pain received significantly less than those that were not.

The benefits of intrinsic motivation, on the other hand, remain unclear. The robots seemed to learn about the environment at the same rate regardless of their motivation, and the robots that had no motivation and just moved randomly were actually able to maintain their understanding of the environment better than those that were motivated either intrinsically or extrinsically, even though previous research has indicated that intrinsic motivation should be effective on its own (Lee et. al, 2009 and Oudeyer et. al, 2007). The robots that were only intrinsically motivated did still maintain a slightly better understanding of the environment than those that had extrinsic motivation, although the robots that were motivation both intrinsically and extrinsically had roughly the same performance as those that were motivated only extrinsically. Overall, the experiment did not demonstrate a clear benefit (and possibly even a detrement) to the robot's ability to learn from intrinsic motivation, especially when it was combined with extrinsic motivation.

If the problem with the motivated robots' ability to learn is that they become too specialized after initially learning the environment, as speculated, one possible solution would be to modify the way the robots predict the results of their actions. The system currently in place–one neural network per action with no hidden layers in any networks–is very primitive, and not very well suited to more complex environments or ones where the same action can produce a variety of different results, since it only allows for the states resulting from an action to be a linear function of the initial state. One way to possibly solve this problem is to add hidden layers to the networks, but perhaps a


Random

Intrinsic Motivation Only

Extrinsic Motivation Only

Both Types of Motivation

Both Types of Motivation, High Intrinsic Weight

Figure 4: The times and battery levels at which the robots recharged their battery in each setup. Robots that were intrinsically rewarded rarely let their battery drop below zero.

better way is to increase the number of neural networks used.

One way to do this could be to categorize states and use a different network for each category, in a manner similar to CBIM (Lee et. al, 2009). The networks could take both the state parameters and the action parameters as inputs, and the robot would begin with just one network for predicting everything. A condition could be created for when to create new state categories based on the existing network's error, and when a new state category is created a new network could also be created for it. If the categorization system were successfully designed in such a way that every sensorimotor input in a given category could be accurately predicted using the same neural network, then specialization would not be a concern. If the robot focused on only a small set of situations, the networks for those situations would improve, but any situation different enough that those improvements would not help would be using a different network anyway.

In addition to using state categorization to potentially improve the robot's ability to predict the results of its actions, it could also be used to categorize states for the purposes of calculating rewards. States could be grouped into categories based on the expected intrinsic and extrinsic rewards from those categories (so two states in the same category would give close rewards for the same action). This would have to potential to improve the system's performance in larger environments significantly. Although it was not directly tested in very large, complex environments, it seems likely that, as currently implemented, it would struggle with them, since the number of entries in the reward tables would get too large. If the robot categorized states based on the rewards recieved (possibly with separate categorizations for intrinsic and extrinsic rewards), it would need fewer entries in the tables for larger environments.

Perhaps even more importantly than helping the robot function in large environments, using state categorization for both predicting states and tracking expected rewards would allow the system to function in continuous environments. Right now, the need for a separate table entry for every possible state means the system can only operate in environments where the states are discrete. If states were categorized, however, then table entries would only be necessary for each category, so it would be okay for the states to be continuous since the categories would still be discrete. This possibility of extending the system to work in continuous environments is very important, since it is necessary for the system to be used on real robots.

6 Bibliography

- Singh, S., A.G. Barto, and N. Chentanez (2004). "Intrinsically Motivated Reinforcement Learning." *18th Annual Conference on Neural Information Processing Systems*, Vancouver, B.C., Canada, December 2004.

- Stout, A., G.D. Konidaris, and A.G. Barto (2005). "Intrinsically Motivated Reinforcement Learning: A Promising Framework For Developmental Robot Learning." *Proceedings of the AAAI Spring Symposium on Developmental Robotics, Stanford University*, Stanford, CA, March 21-23, 2005.

- Huang, X. and J. Weng (2007). "Inherent Value Systems for Autonomous Mental Development." *International Journal of Humanoid Robotics, vol. 4, no. 2, 2007*, 407-433.

- Oudeyer, Pierre-Yves, Frederic Kaplan, and Verena Hafner (2007). "Intrinsic Motivation Systems for Autonomous Mental Development." *IEEE Transactions on Evolutionary Computation, vol. 11.2, 2007.*

- Lee, Rachel, Ryan Walker, and Lisa Meeden (2009). "Category-Based Intrinsic Motivation." *Proceedings of the Ninth International Conference on Epigenetic Robotics, 2009.*