# Two Perspectives on Learning Rich Representations from Robot Experience

**Joseph Modayil**
Reinforcement Learning and Artificial Intelligence Laboratory
Department of Computing Science, University of Alberta

## Abstract

This position paper describes two approaches towards the representations that a robot can learn from its experience. In the first approach, the robot learns models for reasoning about human-interpretable aspects of the environment, for example models of space and objects. In the second approach, the robot incrementally learns predictions for the consequences of performing policies, where a policy is any experimental procedure that the robot can perform. These two approaches correspond closely to the ideas of a scientific model and an experimental prediction, and ideally the benefits of both can be accessible to a robot.

The different prerequisites needed to support reasoning and learning have led to different forms for representations learned from robot experience. The desire to enable a robot to reason about its environment in terms familiar to people has led to procedures for learning representations from a robot's low-level experience that support inference for models of space (Pierce and Kuipers 1997), objects (Modayil and Kuipers 2007), and even communication (Steels 1998). The desire to accurately predict the temporally extended consequences of a robot behaviour has led to statistically-sound fully-incremental learning algorithms (Sutton 1988; Maei and Sutton 2010). These two perspectives correspond to the complementary capabilities of scientific models and scientific experiments.

A scientific model describes some aspect of a system and ignores other aspects as being outside the scope of the model. Scientific models can be descriptive (e.g. a classification of living organisms into genus and species), or they can simulate system dynamics based on a particular description (e.g. celestial mechanics). For many computational purposes, the most useful capability of a model is to simulate some aspect of the dynamics. As an example of such models, the SLAM approach to robot mapping relies on observation models (assigning likelihoods to observations given the robot's pose and the map) and motion models (generating samples for the robot's next pose based on the motor commands). Any particular model of the physical world will be imperfect, this is true for scientific models and for the computational models available to a robot. Having access to multiple models can mitigate the limitations of any particular model, for example a robot with causal, topological, and metrical maps can use any of these models to support navigation (Kuipers 2000). However, building large systems with interacting models is challenging because of differing semantics at the interfaces between models. Human intervention is often required to orchestrate and debug the interactions between models that lack a robot-interpretable semantics.

The scientific experiment complements the scientific model, by measuring a single aspect of the world through a carefully specified experimental procedure. A robot with the ability to predict the result of performing a procedure has a limited but concrete piece of knowledge. Moreover, because the result can be observed, this knowledge can be verified by a robot, and it can be the target of a learning algorithm. Even when a model is used to provide predictions about an experiment, measurements of a precise scientific procedure can exhibit variations that the model fails to predict. For example, a robot might move at different speeds on asphalt or on ice, but a standard motion model would not predict this variation. A disagreement between two models can be used to suggest a new experiment. Conversely, generalizing from particular experiments might suggest a useful form for a model.

The interaction between models and experiments provides the foundation of scientific knowledge. This interaction has been explored as a foundation for learning in children (Gopnik, Meltzoff, and Kuhl 1999), and it could provide similar insights for structuring a robot's knowledge. An initial concern is whether such an approach is tractable on a robot, namely whether models or predictions might be learned by a robot. The remainder of this paper surveys some results that suggest that both models and predictions can be learned by a robot from its sensory-motor experience.

## Models

Computationally, models on robots are primarily used for inference via simulation. As any computational model available to a robot will be imperfect representation of the physical world, the utility of a model is of primary importance in robotics. However, the development of large simulation models in robotics has also traditionally relied on concise descriptive models that are amenable to direct human interpretation, where the accuracy of the description can be externally verified by the system designer by comparing device specifications with calibrated measurements.

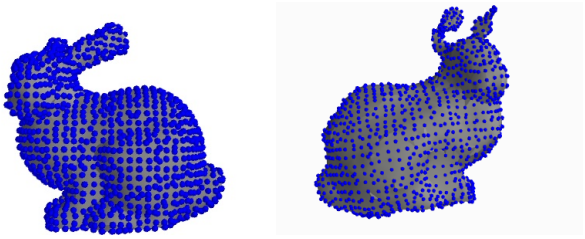

Figure 1: Given sensors distributed on a simulated complex surface (left), a learning algorithm uses observed sensor readings to reconstruct a spatial embedding of the sensors (right). (Modayil 2010)

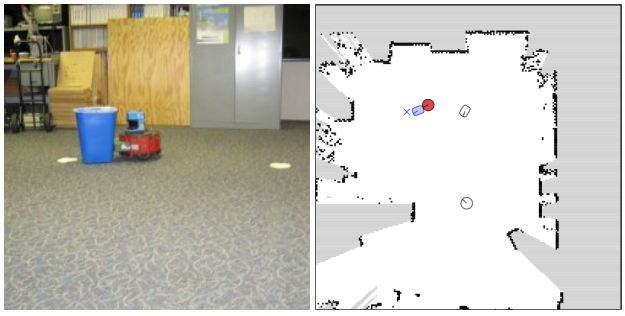

Figure 2: The robot constructs perceptual representations for physical objects with features for location, shape, and local dynamics models. Using its learned models, the robot constructs a plan within the model, and follows the plan to move the object from the starting location (hollow) to the goal location (marked x). (Modayil and Kuipers 2007)

One example of learning a descriptive model is constructing a spatial embedding of the robot's sensors (Modayil 2010). In this work, correlations between sensor readings are used to infer a three dimensional embedding of a two dimensional surface of sensors. The modelling assumption is that sensors that are spatially adjacent will lead to strongly correlated readings from the sensors. This can be used to form spatial reconstructions of the sensors as shown in Figure 1. Note that this descriptive model does not directly support simulation. but a similar model of sensor geometry was used to construct perceptual features for goal-seeking control laws (Pierce and Kuipers 1997).

Another algorithm (Modayil and Kuipers 2007) constructed perceptual features that correspond to physical objects. Using sensor readings from a planar laser range-finder, the algorithm clustered the observations that are not explained by a stationary world model to form a new tracked entity. Thus, portions of the robot's sensory experience that were poorly explained by one model (the robot's map) provided the substrate for developing a new model. More perceptual features were constructed for the tracked entities, which in turn enabled reasoning with a geometric model to provide a shape description, and simple control laws were learned for the one-timestep dynamics of the robot's interaction with the perceived entity. These internal models enabled the robot to exhibit competent behaviour with the physical objects (Figure 2). Moreover, this approach enables a system designer to more readily express their goals for the robot, because the robot's perception and the designer's perception are partially aligned.

These examples demonstrate that models can be learned from a robot's experience, and that the robots can use these models to exhibit competent behaviour in domain specific ways. It also demonstrates that structure in the environment perceptible by people can induce statistical regularities in the stream of robot experience, and that useful new models can be constructed by identifying these statistics. It is perhaps surprising that simple statistical models can at times be composed to bridge the gap between the robot's sensory-motor stream and the high-level abstractions that people use to describe the physical world.

However, inference across multiple models still requires substantial human effort. For example, a robot with a knee-high laser rangefinder (as in Figure 2) can perceive a walking person as either two separate objects (two legs), or as one object (one body). The robot's inability to recognize these distinctions can cause problems when communicating goals between a person and the robot. More significant problems arise when translating semantics across models. For example, the robot treats the object as immobile in one model to navigate around it, and as mobile when planning to push it. Having a single underlying semantics for the models would enable the robot to use multiple models without human intervention.

## Experimentally Verifiable Predictions

Experiments provide the foundation for identifying scientifically distinct models—two theoretical models that are experimentally indistinguishable are empirically a single model. The result of an experiment reveals a fact about the robot's environment, and a prediction of such a result can constitute the robot's knowledge of this fact. If a robot's knowledge is expressed as the predicted consequences of an experimental procedure, then the robot can correct errors in its knowledge by performing an experiment (Sutton 2009).

Algorithms for learning to make such predictions can take different forms. Temporal-difference algorithms (Sutton 1988) are particularly efficient for incremental online learning on robots. These algorithms work with function approximation, in particular they can learn an approximate answer as a function of a feature vector computed by the robot. A limitation of the standard algorithms in this setting is that the robot's predictions might diverge when the robot's behaviour differs from the target policy (the experimental procedure). Gradient-based algorithms (Maei and Sutton 2010) remove this limitation, enabling the robot to learn off-policy, using the behaviour from its one lifetime to learn about the consequences of different ways of behaving.

A formal representation for the predicted consequences of following an experimental procedure is provided by a general value function (Sutton et al. 2011). The experimental procedure is formally a policy $\pi$ that describes the probability of taking an action from any robot state. The end of the

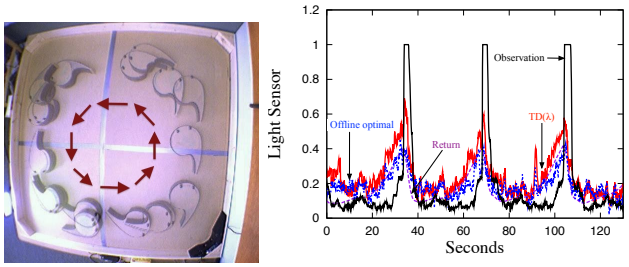


Figure 3: The robot moves around the wooden pen (left) and learns to make thousands of real-time predictions. One of these predictions (right) is for when one of its light sensors will be saturated. The ideal return (purple) rises before the observation (black), and the learned prediction (red) closely matches the ideal. (Modayil, White, and Sutton 2012)

experiment at time $T$ is determined by $\gamma$, a function that returns the continuation probability from every state (and thus terminates with probability $1-\gamma$). The result of the experiment started at time $t$ is given by the real-valued return $G_t$, which is the sum of a pseudo-reward signal $r$ accumulated at each step of the experiment with an outcome $z$, that is added at termination. A general value function $q$ is the expected return from following the experiment from any state-action pair,

$$q^{\pi,\gamma,r,z}(s,a) = \mathbb{E}[G_t | S_t = s, A_t = a, T \sim \gamma, A_{t+1:T-1} \sim \pi],$$

where the symbol $\sim$ denotes the variable on the left side is drawn from the (induced) distribution on the right side. The robot does not have access to the underlying state of the environment, but uses a real-valued vector $x_t \in \mathbb{R}^n$ to capture features of the environmental state, $s_t$, and action, $a_t$, at time $t$. A learning algorithm can compute a linear approximation to this general value function, with a weight vector $w \in \mathbb{R}^n$, by $\sum_i w^i x_t^i \approx q(s_t, a_t)$.

Another result has demonstrated the ability of this approach to learning at scale. Recent work (Modayil, White, and Sutton 2012) has shown that thousands of predictions can be learned with thousands of features at more than 10Hz on a laptop (Figure 3). Predictions for all of a robot's sensors are learned within a few hours and with considerable accuracy, using a shared feature vector and shared parameter settings for all the predictions.

This approach provides the robot with an abundance of predictive facts with semantics that are accessible to the robot. Multiple predictions about the consequences of following a policy can be aggregated to form an *option model* (Sutton, Precup, and Singh 1999). Algorithms for planning with such models under function approximation (Sutton et al. 2008) can be used to modify robot behaviour. These results suggest that collections of empirical predictions might be assembled into models.

## Discussion

The two approaches have demonstrated complementary strengths that might be combined. The first approach shows that high-level models that support reasoning can be learned from low-level robot experience. The second approach has shown that learning to make well-formed predictions is feasible in real-time with a single set of semantics. These results suggest that the functionality of non-predictive models forms might be recast in predictive forms. For example, instead of view-invariant representations of object's shape, a robot could learn to predict what it would observe after approaching and facing an object. A unified semantics for reasoning and learning would provide a robot with a powerful method for understanding its environment.

## Acknowledgments

Discussions with Richard Sutton and Benjamin Kuipers have influenced several ideas in this paper, and comments from reviewers have improved the presentation. This work has been financially supported by Alberta Innovates–Technology Futures, the Alberta Innovates Centre for Machine Learning, the Glenrose Hospital, and the Reinforcement Learning and Artificial Intelligence Laboratory.

## References

Gopnik, A.; Meltzoff, A.; and Kuhl, P. 1999. *The Scientist in the Crib*. HarperCollins.

Kuipers, B. J. 2000. The Spatial Semantic Hierarchy. *Artificial Intelligence* 119:191–233.

Maei, H. R., and Sutton, R. S. 2010. GQ($\lambda$): A general gradient algorithm for temporal-difference prediction learning with eligibility traces. In *Proceedings of the Third Conference on Artificial General Intelligence*.

Modayil, J., and Kuipers, B. J. 2007. Autonomous development of a grounded object ontology by a learning robot. In *Proc. 22nd National Conf. on Artificial Intelligence (AAAI-2007)*.

Modayil, J.; White, A.; and Sutton, R. S. 2012. Multi-timescale nexting in a reinforcement learning robot. In *From Animals to Animats 12*. Springer. 299–309.

Modayil, J. 2010. Discovering sensor space: Constructing spatial embeddings that explain sensor correlations. In *IEEE 9th International Conference on Development and Learning (ICDL)*, 120–125.

Pierce, D. M., and Kuipers, B. J. 1997. Map learning with uninterpreted sensors and effectors. *Artificial Intelligence* 92:169–227.

Steels, L. 1998. The origins of syntax in visually grounded robotic agents. *Artificial Intelligence* 103:133–156.

Sutton, R. S.; Szepesvari, C.; Geramifard, A.; and Bowling, M. 2008. Dyna-style planning with linear function approximation and prioritized sweeping. In *Proceedings of the 24th Conference on Uncertainty in Artificial Intelligence*, 528–536.

Sutton, R. S.; Modayil, J.; Delp, M.; Degris, T.; Pilarski, P. M.; and Precup, D. 2011. Horde: A scalable real-time architecture for learning knowledge from unsupervised sensorimotor interaction. In *Proceedings of the 10th International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*.

Sutton, R. S.; Precup, D.; and Singh, S. 1999. Between MDPs and semi-MDPs: A framework for temporal abstraction in reinforcement learning. *Artificial Intelligence* 112(1):181–211.

Sutton, R. S. 1988. Learning to predict by the methods of temporal differences. *Machine Learning* 3(1):9–44.

Sutton, R. S. 2009. The grand challenge of predictive empirical abstract knowledge. In *Working Notes of the IJCAI-09 Workshop on Grand Challenges for Reasoning from Experiences*.