In Proceedings of the Sixth International Conference on Cognitive Modeling, 285-289.
Mahwah, NJ: Lawrence Earlbaum.

Simple Agents Learning to Add Useful Structures to the World

Terry Stewart (tcstewar@connect.carleton.ca)
Sanjay Chandrasekharan (schandra@sce.carleton.ca)
Institute of Cognitive Science, Carleton University
Ottawa, Canada, K1S 5B6

Abstract

We provide a computationally tractable model of how organisms can learn to add structures to the world to reduce the complexity of their tasks. This model is then implemented using two techniques: first using a genetic algorithm, and then using the Q-learning algorithm. The results clearly show that organisms with only reactive behavior can learn to systematically add structures to the world to reduce their cognitive load. We show that such learning can happen in both evolutionary time and within an agent's lifetime. An extension of this model (currently being implemented) is then illustrated, where organisms with just reactive behavior learn to systematically generate and use internal structures akin to representations.

Many organisms generate stable structures in the world to reduce their cognitive load. Wood mice distribute small objects, such as leaves or twigs, as points of reference while foraging. Under laboratory conditions, they will spontaneously make use of plastic discs for this purpose. Stopka & MacDonald (2003) show that this 'way-marking' diminishes the likelihood of losing interesting locations during foraging. Red foxes use urine to mark food caches they have emptied. This marking acts as a memory aid and helps them avoid unnecessary search (Henry, 1977, reported in Stopka & MacDonald, 2003). The male bower bird builds colorful bowers (nest-like structures), which are used by females to make mating decisions (Zahavi & Zahavi, 1997). Ants drop pheromones to trace a path to a food source. Many mammals mark up their territories.

At the most basic level, cells in the immune system use antibodies that bind to attacking microbes, thereby 'marking' them. Macrophages use this 'marking' to identify and destroy invading microbes. Bacterial colonies use a strategy called 'quorum sensing' to know that they have reached critical mass (to attack, to emit light, etc.). This strategy involves individual bacteria secreting molecules known as auto-inducers into the environment. These accumulate in the environment, and when it reaches a threshold, the colony moves into action (Silberman, 2003).

Given that this 'doping' of the world is so common in these simpler creatures, it is somewhat surprising that there has been relatively little investigation into the use of this technique by *homo sapiens*. More than any other species, humans generate these external structures to reduce the amount of physical and cognitive effort required to perform tasks in their daily lives. Examples include markers, color-codes, page numbers, credit-ratings, badges, shelf-talkers, speed bugs, road signs, post-it notes – an almost endless list.

Epistemic Structures

The pervasiveness of such structures across species indicates that adding structure to the world is a fundamental cognitive strategy (Kirsh, 1996). Note that these structures serve to make tasks easier for agents. Some of these structures have referential properties, but they do not exist for the purpose of reference. We use the term *epistemic structures* to refer to these, in deference to Kirsh's (1994) distinction between epistemic and pragmatic action.

Kirsh's (1996) model of "changing the world instead of oneself", postulates that such generation of structures involve task-external actions, and these structures work by deforming the state space, so that paths in a task environment are shortened. Such structures also allow new paths to be formed in the task environment. However, Kirsh only explicitly addresses the generation of *tools*, rather than the direct modification of the world to reduce cognitive load.

Extending his idea to develop a full computational model of how organisms generate such structures, we make two reasonable assumptions. One, organisms sometimes randomly generate structures in the environment (pheromones, urine, leaf piles) as part of their everyday activity. Two, organisms can track their physical or cognitive effort (i.e. they get 'tired'), and they have a bias to reduce tiredness.

Given these assumptions, some of the randomly generated structures are encountered while executing tasks like foraging and cache retrieval. In some random cases, these structures make the task easier for the organisms (following pheromones reduces travel time, avoiding urine makes cache retrieval faster, avoiding leaf-piles reduce foraging effort). In other words, they shorten paths in the task environment. Given the postulated bias to avoid tiredness, these paths get preference, and they are reinforced. Since more structure generation leads to more of these paths, structure generation behavior is also reinforced.

The Simulation

To test and investigate the above model of epistemic structure generation, we have developed a computational model, where simple agents in a simple world, given only feedback in terms of their 'tiredness' (i.e. the effort required to perform their task), learn to systematically add structures to their environment.

The task we have chosen is analogous to foraging behavior, i.e. navigating from a home location to a distant target location and back again. Our environment consists of a 30x30 toroidal grid-world, with one 3x3 square patch representing the agent's home, and another representing the target. This 'target' can be thought of as a food source, to fit with our analogy to foraging behavior.

Agent Actions

At any given time, an agent can do one of five possible actions. The first and most basic of these is 'moving randomly'. This consists of going straight forward, or turning to the left or right by 45 degrees and then going forward. The agent does not pick which of these three possibilities occurs (there is a 1/3 chance of each).

In deciding the actions available to the agent, we needed to postulate some basic facilities within each agent. In our case, we felt it was reasonable to assume that the agents could distinguish between their home and their target. To do this, we added two more actions to the agents' repertoire. These are exactly like the first action, but instead of moving randomly, the agent would move to whichever square is sensed to be the most 'home-like' (or the most 'target-like'). Initially, the only things in the environment that are 'home-like' or 'target-like' are the home and the target themselves.

One way to think about these actions is to consider the pheromone-following ability of ants. Common models of ant foraging (e.g. Bonabeau et al, 1999) consist of the automatic release of two pheromones: a 'home' pheromone and a 'food' pheromone. The ants go towards the 'home' pheromone when they are searching for their home, and they go towards the 'food' pheromone when foraging for food. This exactly matches these two actions in our agents. The 'home' pheromone would be an example of a 'home-like' structure in the ant environment.

The fourth and fifth possible actions provide for the ability to generate these 'home-like' and 'target-like' structures. In the standard ant models, this could be thought of as the releasing of pheromones. However, our simulation has an important and very key distinction. Here, this ability to modify the environment is something the agents can do *instead* of moving around. That is, this generation process requires time and effort. The best way to envisage this is to think of an action that a creature might do which inadvertently modifies its environment in some way. Examples include standing in one spot and perspiring, or urinating, or rubbing up against a tree. These are all actions which modify the environment in ways that might have some future effect, but do not provide any sort of immediate reward for the agent. Kirsh (1996) terms these 'task-external actions'.

It must be stressed here that we are not presuming any sort of long-term planning on the part of the agents. We are simply specifying a collection of actions available to them, and they will choose these actions in a purely reactive manner (i.e. based entirely on their current sensory state). It may also be noted that our 'actions' are considered at a slightly higher level than is common in agent models. Our agents are not reacting by 'turning left' or 'going forward'; they are reacting by 'following target-like things' or 'moving randomly'. Furthermore, they do not initially have any sort of association between the action of making 'home-like' structures and the action of moving towards 'home-like' things. Any such association must be learned (either via evolution, or via some other learning rule).

Also, our agents are not designed to form structures automatically as they wander around (as is the case in standard ant models). In our simulation, a creature must expend extra effort to systematically generate these structures in the world. An agent that does this will be efficient only if the effort spent in generating these structures is more than compensated for by the effort saved in having them. Moreover, these are not permanent structures. The agents' world is dynamic and the structures do not persist forever. The 'home-likeness' or 'target-likeness' of the grid squares decrease exponentially over time. Furthermore, these structures also spread out over time. A 'home-like' square will make its neighboring squares slightly more 'home-like'. This can be considered similar to ant pheromones dispersing and evaporating, or leaf/twig piles being knocked over and blown around by wind or other passing creatures.

Table 1: The five actions available to the agent.

Agent Actions
Move randomly
Move toward 'home-like' structures
Move toward 'target-like' structures
Make 'home-like' structure
Make 'target-like' structure

Agent Sensing

Since our agents are reactive creatures and thus do no long-term planning, they require a reasonably rich set of sensors. We have given them four sensors, two external and two internal, to detect their current situation. The two external sensors sense how 'home-like' and how 'target-like' the current location is (digitized to 4 different levels). The internal sensors are two simple bits of memory. One indicates whether the agent has been to the target yet, and the other indicates how long it has been since the agent generated a structure in its environment (up to a maximum of 5 time units). This is all that the agents can use to determine which action to perform. This configuration gives each agent 192 (4 x 4 x 6 x 2) different sensory states.

Table 2: The four sensory values available to the agent.

Agent Sensory State
Similarity to 'home' of current location (0-3)
Similarity to 'target' of current location (0-3)
Amount of time since the agent last made a change to the environment (0-5)
Currently carrying anything (0-1)

Alternative State/Action Representations

The particular choices we made for the set of actions and sensory states the agent has available to it (as shown in tables 1 and 2) may be a bit non-intuitive. They are not the first representations that we tried. Initially, instead of the ‘move randomly’, ‘follow home-like’, and ‘follow target-like’ actions, we used the more traditional actions of ‘move forward’, ‘turn left’, and ‘turn right’. Using these actions requires a more complex sensory state; we would have had to add sensor values to detect how home-like and target-like the squares ahead of, to the left of, and to the right of the agent.

However, when we used this approach of having a more complex sensory state and a less complex set of actions, the agents were unable to learn to create structures. For this reason, we used the state/action representations shown in tables 1 and 2 for the results given in this paper. We also found that the ‘time since a change was made’ sensor was also needed for the agents make use of their abilities to change the world.

The Learning Rules

Given our representations, we needed some way of determining which action the agent will perform in each of these 192 states. Note that by having this sort of mapping, we are implying a purely reactive agent. We investigated two different methods for matching sensory states to actions: a Genetic Algorithm, and Q-Learning.

Stage 1: The Genetic Algorithm

For our first model, we used a genetic algorithm to determine which action to take in each situation. The genome consisted of a simple list of actions, one to perform in each state. To evaluate a particular genome, we started 10 agents in the home location and ran the simulation for 1000 time steps. The evolutionary fitness was the agents’ average tiredness (i.e. how long it took each agent to make it back home from the target).

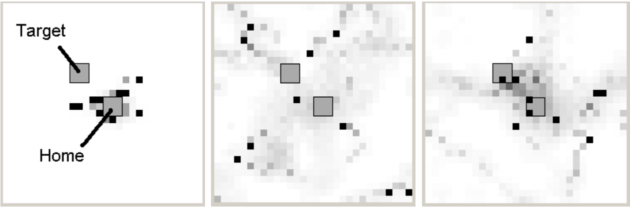

Figure 1: The computer model at 10, 100, and 300 time steps. Black dots are the agents. The shading is darker the more ‘home-like’ or ‘target-like’ a particular square is. This run shows typical agent behavior after 300 generations.

Result: Initially, the agents behaved randomly. Starting at the ‘home’, they would wander about and might, by chance, find the target and then, if they were very lucky, their home. Indeed, most agents did not find the target and make it back within the 1000 time steps. On average, we found that each agent was completing 0.07 foraging trips every 100 time steps. After a few hundred generations, the agents were soon completing an average of 1.9 trips in that same period of time. In other words, the agents were able to, on an evolutionary time scale, learn to make use of their ability to sense and generate structures in the world. Furthermore, this ability provided a very large advantage over completely random behavior.

This result confirmed that it is possible for agents to learn to systematically generate and use structures in the world in an evolutionary time scale. It also showed that we had not chosen an impossible task for the agents to learn. However, for our purposes, we were much more interested in an individual agent learning to generate epistemic structures within that agent’s lifetime. To investigate this, we turned to the Q-Learning algorithm.

Stage 2: Q-Learning

The heart of our investigation was to determine whether a simple, general learning algorithm would allow our agents to discover and make use of the strategy of systematically adding structures to the world. In keeping with our ‘tiredness’ theory, the only feedback the learning mechanism had was an indication of the exertion or effort. The delayed-reinforcement learning rule known as Q-Learning (Watkins, 1989) seemed best suited for this task. (Sarsa and other TD-Learning algorithms would also be suitable). The Q-Learning algorithm[1] develops an estimate of the eventual outcome of performing a given action in a given situation. The agent then performs the action with the highest expected payoff.

Using the Q-Learning algorithm, we again ran 10 agents for 1000 time steps. To indicate ‘tiredness’, we gave them a reinforcement value of -1 all the time (indicating a constant ‘punishment’ for expending any effort). When they returned home after finding the target, they were given a reinforcement of 0, and they were then sent back out again for another trip. Each agent independently used the Q-Learning algorithm, and there was no communication between the agents.

Result: The dark line in figure 2 shows the results averaged over 100 separate trials. We can clearly see that the agents are improving over time (i.e. they are spending less time to perform their foraging task).

Stage 3: Confirmation

Although we have observed improvement over time, we still need to show that it is the agents’ ability to systematically

[1] The estimated reward for performing action a in state s is $Q(s,a)$. This is increased by $\alpha(r+\gamma\max(Q(s',b))-Q(s,a))$, where r is the immediate reward/punishment, s' is the resulting state, γ is the future discounting rate (set to 0.5), and α in the learning rate (0.2). We used ε-greedy action selection with ε set to 0.1, so the agents choose the action with the highest expected reward 90% of the time, and the other 10% they perform an action at random.

add structures to the world that is causing this effect. To prove this, we re-ran the experiment, this time removing the agents' ability to generate structures in the world. No other changes were made.

Result: We found that when the agents were unable to generate structures in the world, Q-Learning did not provide as much improvement[2]. This result is shown in the lighter line in Figure 2. There is still a small improvement given by Q-Learning, but we can conclude that the significant improvement seen in the previous experiment (the dark line in Figure 2) is due to the agents' ability to modify their environment.

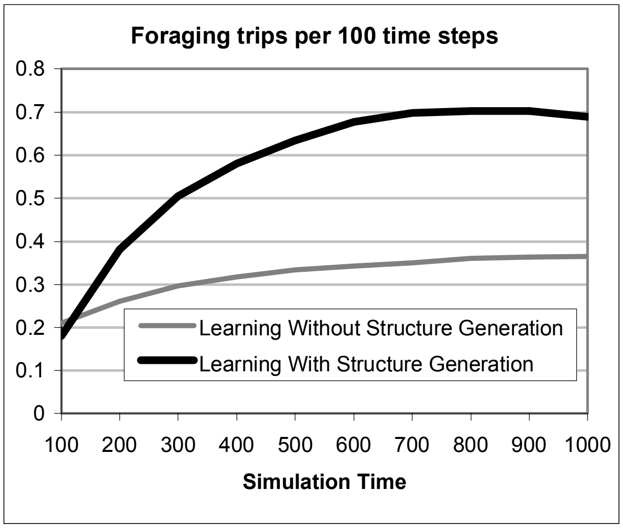

Figure 2: The effect of epistemic structure generation. The foraging rate is measured in trips per 100 time steps. A foraging rate of 0.5 means that trips require an average of 200 time steps to complete.

We can also see from Figure 2 that having these extra actions available does incur some cost in the early stages. Initially, the agents perform slightly worse. However, the advantage of being able to form epistemic structures quickly improves the agents' performance. By the end of the simulation, agents require only around 150 time steps to make a complete trip (a foraging rate of 0.66 trips in 100 time steps). This is twice as quick as agents without the structure-forming ability.

When we analyzed the actions of the agents, we found that they actually spent 58% of their time generating structures. This is striking, since time spent generating these structures means less time for wandering around trying to find the target or their home. Table 3 gives the breakdown of how time was allocated to different actions. The data indicates that epistemic structure generation allowed the agents to go from spending 300 time steps down to 150 time steps to complete their foraging task, even though over half of those 150 time steps are spent standing still. There is clearly a large efficiency advantage to making use of these structures.

Table 3: Time spent performing various actions over 1000 time steps.

Action	With Structure Generation	Without Structure Generation
Move randomly	10%	32%
Toward 'home-like'	19%	36%
Toward 'target-like'	13%	32%
Make 'home-like'	35%	
Make 'target-like'	23%	

Model Capabilities

The Q-Learning system is a concrete implementation of our model: a simple learning mechanism that allows agents with purely reactive behavior to systematically add structures to the world to lower search.

The Q-Learning model implemented in this simulation can explain the generation of structure that is used both by the agent generating the structure, and by the other agents in the environment. The agents ended up forming structures that were useful for everyone, even though they were just concerned about reducing their own tiredness. This was only possible because the agents were similar to each other. This is similar to how paths are formed in fields: one person cuts across the field to reduce his physical effort, others, sharing the same system and wanting to reduce their effort, find the route optimal. As more people follow the route, a stable path is formed.

Other Models

It is worth noting that our model presents a novel simulation of ant behavior. The closest existing models are those in (Bonabeau et al, 1999) which use the 'home-pheromone' and the 'food-pheromone'. This is in contrast to such models as (Nakamura & Kurumatani, 1996), where a land-based and an airborne pheromone are used, or any models of the Cataglyphis species of ant, which uses a complex landmark-navigation scheme which allows it to return directly to the nest (Miller & Wehner, 1988).

That said, all of these other models assume both that pheromones are continually being released while the ant forages, and that there is no learning happening during the foraging behavior. Our model does not make either of these assumptions.

We were unable to find references indicating that real ants might, in fact, learn to use pheromones, or any research that indicates that the effort required to produce these pheromones might interfere with the foraging behavior. This indicates our model may not be a good one for understanding ants. However, the fact that our agents are able to learn to reflexively generate these cognitively

[2] Q-Learning also did not provide significant improvement if the agents were only able to generate one type of structure, or if any of the agent's sensors were removed.

beneficial structures in the absence of any immediate feedback to their benefit, indicates a simpler way to model more complex creatures that exhibit such behavior.

Conclusions

The model presented here shows that a simple agent using Q-Learning can learn to modify its environment in such a way as to reduce the amount of effort required to perform a task. This ability to change the environment is one that is common in simple creatures, but has not been the focus of attention of computational modeling. This ability to change the world is known to be fundamental for a broad range of human activity. This result indicates a new domain of investigation for more complex learning agents in more dynamic and realistic environments.

Future Work

Interestingly, this same model could explain generation and tracking of *internal* structures in organisms. The actions which generated structure in our simulation were actions that affected the environment. But this does not have to be the case. Just as we had both internal and external sensors, we can have actions which affect either the state of the world *or* the state of the agent itself. In other words, we can use this model to investigate the generation of *internal* structure (i.e. representations).

As an example, consider foraging bees. Suppose that, just as our agents left traces in the world of their activity via their structure-generating actions, we have the bees leave a sequence of internal memory traces corresponding to landmarks (say a tall tree, a lake, a garden) as a result of their everyday foraging activity. In some foraging trips of some bees, the trace sequences match to some degree the external structures they perceive. Such trips involve less search, because they lead to food more directly, i.e. they form shorter paths in the task environment. Over time, using the exact same learning mechanisms that apply in the external case, the bias against tiredness leads to such paths being used more, and so they are reinforced. This could lead to landmark-based navigation, which does, in fact, exist in bees (Gould, 1990). As in the case of external structures, the generation of such memory traces is reinforced because more traces lead to more such shorter paths in the task environment. We are currently working on a computational model of this example. Interestingly, recent research shows a similar use of landmarks by homing pigeons, which follow highways, railways and rivers to reach their destination with less cognitive effort (Guilford, 2004).

This idea presents a situated cognition model of how memory structures come to be used as task-specific structures, and why such internal structures are systematically generated. If such task-specific memory structures are considered to be representations (that is, they stand for something specific in the world), then the model explains, in a computationally tractable manner, how organisms with just reactive behavior can learn to generate and use representations.

The model also explains what such 'primitive' representations are: they are internal traces of the world that allow the agent to shorten paths in a task environment. Roughly, they are computation-reducing structures (and equivalently, energy-saving structures). They are internal 'stepping stones' that allow organisms to efficiently negotiate the ocean of stimuli they encounter. This means the traditional cognitive science view, that thinking is computations happening over representations, presents a secondary process. In the stepping stone view, representations are crucial for organisms, but they are just useful, incidental entities, not fundamental entities by themselves.

All source code for the simulations can be found at: http://www.carleton.ca/iis/TechReports/code/2004-01/

References

Alcock, J. (1998*). Animal Behavior: An evolutionary approach*, Massachusetts, Sinauer Associates.

Bogen, J. (1995). Teleological explanation. In Honderich (Ed.), *The Oxford Companion to Philosophy*. Oxford University Press, New York.

Bonabeau E., Dorigo M. and Theraulaz G. (1999) *Swarm intelligence: From natural to artificial systems.* Santa Fe Institute studies in the sciences of complexity. Oxford University Press, New York.

Clark, A. (1997). *Being There: putting brain, body, and world together again*, Cambridge, Mass., MIT Press.

Gould, J.L. (1990). Honey bee cognition. *Cognition*, *37*, 83-103.

Guilford, T., Roberts, S. & Biro, D. (2004). Positional entropy during pigeon homing II: navigational interpretation of Bayesian latent state models. *Journal of Theoretical Biology*.

Henry, J.D. (1977). The use of urine marking in the scavenging behaviour of the red fox (Vulpes vulpes). *Behaviour*, *62*, 82-105.

Kirsh, D., & Maglio, P. (1994). On distinguishing epistemic from pragmatic action. *Cognitive Science*, *18*, 513-549.

Kirsh, D. (1996). Adapting the environment instead of oneself. *Adaptive Behavior*, *4:3/4*, 415-452.

Miller, M., & R. Wehner (1988). Path integration in desert ants, Cataglyphis fortis. *Proceedings of the National Academy of Sciences, 85*: 5287-5290.

Nakamura, M., & Kurumatani, K. (1996). Formation mechanism of pheromone pattern and control of foraging behavior in an ant colony model. *Proceedings of the Fifth International Conference on Artificial Life*, 67-74.

Silberman, S. (2003), The Bacteria Whisperer. *Wired*, Issue 11.04, April 2003.

Stopka, P. & Macdonald, D. W. (2003) Way-marking behavior: an aid to spatial navigation in the wood mouse (Apodemus sylvaticus). *BMC Ecology*, *3:3.*

Watkins, C. (1989). *Learning From Delayed Rewards*, Doctoral dissertation, Department of Psychology, University of Cambridge, Cambridge, UK.

Zahavi, A., & Zahavi, A. (1997). *The Handicap Principle: A missing piece of Darwin's puzzle*. Oxford: University Press, Oxford.