# Reactive Agents Learn to Add Epistemic Structures to the World

**Sanjay Chandrasekharan (schandra@sce.carleton.ca)**
**Terry Stewart (tcstewar@connect.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 cognitive complexity. 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 cognitive complexity (minimize search or inference), for themselves, for others, or both. Wood mice (*Apodemus sylvaticus*) distribute small objects, such as leaves or twigs, as points of reference while foraging. They do this even under laboratory conditions, using plastic discs. Such 'way-marking' diminishes the likelihood of losing interesting locations during foraging (Stopka & MacDonald, 2003). Red foxes (*Vulpes vulpes*) 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 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. The auto-inducers accumulate in the environment, and when it reaches a threshold, the colony moves into action (Silberman, 2003).

Such 'doping' of the world is commonly seen in lower animals. Most large animals (large body & brain size) do not exploit this strategy. Humans, however, do so to a tremendous degree. Markers, color-codes, page numbers, credit-ratings, badges, shelf-talkers, speed bugs, road signs, post-it notes, the list of epistemic structures used by humans is almost endless. Humans also add structures to the world to reduce cognitive complexity for artifacts. Examples include bar codes (makes check-out machines' decisions easier), content-based tags in web pages (makes Web agents' decisions easier), sensors on roads (helps the traffic light program's decision-making), etc.

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 predominantly serve a task-smoothening function – they make tasks easier for agents. Some of these structures have referential properties, but they do not exist for the purpose of reference. From here onwards, we will term such stable structures that provide "cognitive congeniality" (Kirsh, 1996), *epistemic structures*. The term is derived from a distinction between epistemic and pragmatic action made by Kirsh (1994).

How do organisms generate and use such structures? Can this generation of structures be captured computationally? These are the questions we address in this paper.

## A Taxonomy and a Property

Most of the literature on epistemic structures is by David Kirsh, and from the field of Distributed Cognition in general. Kirsh's work explores the structural and computational properties of such structures, and how they function. We are interested in the other half of the problem, i.e., how such structures are generated and used. We use Kirsh's model to develop a situated cognition model of how such structures are generated. We then outline two simulations we implemented to test this model. An extension of this model (currently in progress) is then described.

Epistemic structures can be classified into three types, based on whom they are generated for. (examples of each in brackets).

1. Structures generated for oneself (Cache marking, bookmarks)
2. Structures generated for oneself and others (Pheromones, color codes)
3. Structures generated exclusively for others (Warning smells, badges)

A central feature of such structures is their task-specificity (more broadly, function/goal-orientedness). To illustrate this concept, consider the following example. Think of a major soccer match in a large city, and thousands of fans arriving in the city to watch. The organizers put up large soccer balls on the streets and junctions leading up to the venue. Fans would then simply follow the balls to the game venue. Obviously, the ball reduces the fans' cognitive load, but how? To see how, we have to examine the condition where big soccer balls don't exist to guide the fans.

Imagine a soccer fan walking from his hotel to the game venue. She makes iterated queries to the world to find out her world state (What street is this? Which direction am I going?), and then does some internal processing on the information gained through the queries. After every few set of iterated queries and internal processing, she updates her world state and mental state, and this continues until she reaches her destination.

What changes when the ball is put up? The existence of the big soccer ball cuts out the iterated queries and internal processing. These are replaced by a single query for the ball, and its confirmation. The agent just queries for the ball, and once a confirmation of its presence comes in, she updates her world state and internal state. The ball allows the agent to perform in a reactive, or almost-reactive mode, i.e., move from perception to action directly. The key advantage is that almost no (or significantly less) inference or search is required.

This happens because the ball is a task-specific structure; it exists to direct soccer fans to the game venue. Other structures, like street names and landmarks in a city, are function-neutral or task-neutral structures. The fans have to access these task-neutral structures and synthesize them to get the task-specific output they want. Once the huge ball, a task-specific structure, exists in the world, they can use this structure directly, and cut out all the synthesizing. How the soccer fans manage to discover the ball's task-specificity is a separate and relevant issue, but we will not address it here. Task-specificity is a property of all epistemic structures found in nature, including pheromones and markers.

Kirsh's model of “changing the world instead of oneself” (Kirsh, 1996), 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. Kirsh’s model tackles only physical structures generated by organisms, like tools. He does not consider structures generated for cognitive congeniality.

## The Tiredness Model

How are task-specific structures that lower cognitive complexity generated? In this paper we consider the case of non-human organisms like ants, wood mice and red foxes. We will make two reasonable assumptions here. One, organisms sometimes generate random 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 built-in tendency to reduce tiredness.

Now, 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.

This theoretical framework gives us the basis for building artificial agents who also display the ability to learn to systematically generate useful structures in their environment.

## 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 feedback only 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 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 towards 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.

### 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) possible different sensory states.

## The Learning Rules

For a purely reactive agent, we need some way of determining which action the agent will perform in each of these 192 states. 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).

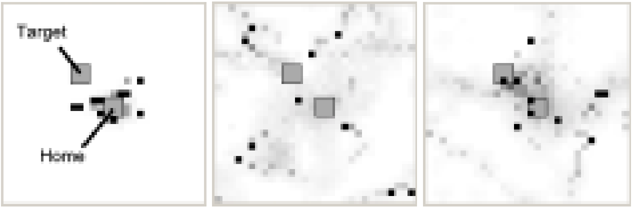

Figure 1: The computer model at 10, 100, and 300 time steps. Black dots are the agents. The shading is darker the more ‘home-like’ or ‘target-like’ a particular square is. This run shows typical agent behaviour after 300 generations.

**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 behaviour.

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. (Other similar algorithms will be investigated in future work). 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 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 are able to conclude that the significant improvement seen in the previous experiment is due to the agents' ability to modify their environment.

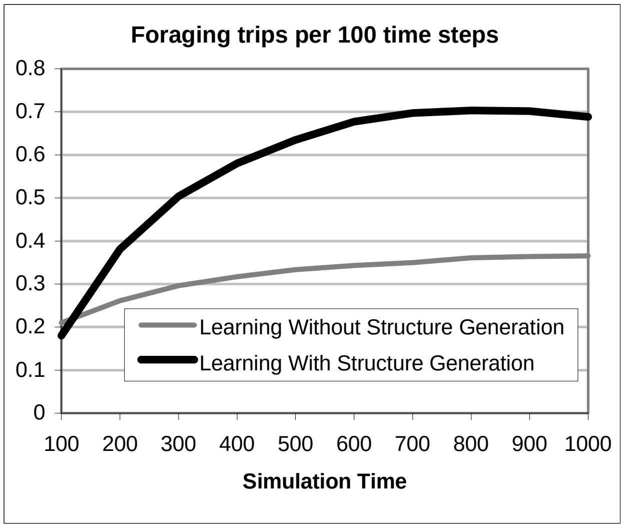

Figure 2: The effect of epistemic structure generation. The foraging rate is measured in trips per 100 time steps. A foraging rate of 0.5 means that trips require an average of 200 time steps to complete.

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.

Table 1: Time spent performing various actions.

| 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% | |

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 1 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

[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, $\gamma$ is the future discounting rate (set to 0.5), and $\alpha$ in the learning rate (0.2). We used an $\varepsilon$-choice rule with $\varepsilon$ 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.

[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.

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.

There are many Reinforcement Learning algorithms available other than Q-Learning, and any one of them could be used in this sort of model. As we investigate other, more complex situations, we will try using these alternatives to Q-Learning, such as actor-critic methods. All of these models learn in a similar way, but with rather different details, and so the resulting high-level behaviour may be different.

## Conclusions

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 'tiredness'-based learning model implemented in this simulation can explain the generation of task-specific structure in cases 1 and 2 (structures for oneself and structures for oneself & others). Case 2 (structures generated for oneself & others) is explained by appealing to the similarity of systems – if a structure provides congeniality for me, it will provide congeniality for other systems like me. In our computer model, the agents ended up forming structures that were useful for everyone, even though they were just concerned about reducing their own tiredness. This was possible only 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.

For case 3, (structures generated exclusively for others), the 'tiredness' model explains only some cases. For instance, it could explain the generation of warning smells and colors exclusively for others, because the effect of such structures could be formulated in terms of tiredness (the release of some chemical ends up cautioning predators, which reduces the number of fleeing responses the organism makes, thus reducing tiredness, which, when fed back, reinforces the initial action). However, this model, as it stands, cannot explain the generation of structures like the bower or the peacock's tail, which do not seem to provide any tiredness benefit for the generator.

## Other Models

It is worth noting that our model presents a novel simulation of ant behaviour. 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 behaviour. Our Q-Learning 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 foraging behaviour. So 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 beneficial structures in the absence of any immediate feedback to their benefit, indicates a simpler way to model more complex creatures that exhibit such behaviour.

## Future Work

Our current simulation implements a learning process based on the feedback of tiredness. It leads to organisms generating task-specific *external* structures in the world. These are structures that lower cognitive load, accessed by organisms at run-time, while they execute tasks.

Interestingly, the same model can 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 leads to landmark-based navigation, which, in fact, exists 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 homing pigeons using human-generated environment structure in a similar fashion to reduce cognitive load. They follow highways and railways systematically to reach their destination (Guilford, 2004).

The above framework 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 the 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 – it describes a privileged path in the task environment. In the stepping stone view, representations are crucial for organisms, but they are just useful, incidental entities, not fundamental entities by themselves. We are exploring the philosophical implications of this view.

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

## Acknowledgment

Some of the ideas presented in this paper were developed while the first author worked as a pre-doctoral fellow with the Adaptive Behavior and Cognition (ABC) Group of the Max Planck Institute for Human Development, Berlin. He acknowledges the group's support, particularly the strong encouragement and critical feedback from Dr. Peter Todd and Dr. John Hutchinson.

## References

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

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

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

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. Positional entropy during pigeon homing II: navigational interpretation of Bayesian latent state models. *Journal of Theoretical Biology*, published online, (2004).

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*, Vol 4, No. 3/4, 415-452.

Miller, M., & R. Wehner (1988). Path integration in desert ants, Cataglyphis fortis. *Proceedings of the National Academy of Sciences* USA 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*, published online, http://www.biomedcentral.com/1472-6785/3/3

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: Oxford University Press.