A Cognitive Model of Learning to Navigate

Diana Gordon
Naval Research Laboratory, Code 5510
4555 Overlook Avenue, S.W.
Washington, D.C. 20375
gordon@aic.nrl.navy.mil

Devika Subramanian
Department of Computer Science
Rice University
Houston, TX 77005
devika@cs.rice.edu

Abstract

Our goal is to develop a cognitive model of how humans acquire skills on complex cognitive tasks. We are pursuing this goal by designing computational architectures for the NRL Navigation task, which requires competent sensorimotor coordination. In this paper, we analyze the NRL Navigation task in depth. We then use data from experiments with human subjects learning this task to guide us in constructing a cognitive model of skill acquisition for the task. Verbal protocol data augments the black box view provided by execution traces of inputs and outputs. Computational experiments allow us to explore a space of alternative architectures for the task, guided by the quality of fit to human performance data.

Introduction

Our goal is to develop a cognitive model of how humans acquire skills by explicit instruction and repeated practice on complex cognitive tasks. We are pursuing this goal by designing computational architectures for the NRL Navigation task, which requires sensorimotor coordination skill. Our model design is grounded in human performance data on the task (both motor output and verbalizations). In this paper, we further develop and test the model reported in Gordon and Subramanian (1996b), which is based on *action models* for actively learning visual-motor coordination. Action models predict action consequences. The agent (our cognitive model) actively interacts with its environment by gathering *execution traces*, which are time-indexed streams of visual inputs and motor outputs, and by learning a compact representation of an effective policy for action choice from such traces, guided by action models.

This paper begins with an analysis of the NRL Navigation task and the requirements of an optimal controller for this task. We then briefly describe the human experiments from which our model (different from the optimal controller) was constructed, followed by an overview of our cognitive model from Gordon and Subramanian (1996b). Our objective is to construct the simplest model that accounts for essential elements of performance common to all individuals. The following sections explore two main topics arising from the verbal protocols: shift of attention between subtasks, and the nature of sensory predictions in the action models. We conclude that human learners shift focus between two primary subtasks of the task. This conclusion is clearly grounded in supporting evidence: the verbal protocol data, results with our cognitive model, and results using an alternative (control) architecture. Results regarding the nature of human sensory predictions, on the other hand, are less definitive than those regarding focus. We show how differences in learning rates on the task can be partially accounted for by variations on the type of sensory predictions.

The NRL Navigation and Mine Avoidance Domain

The NRL navigation and mine avoidance domain, developed by Alan Schultz at the Naval Research Laboratory and hereafter abbreviated the "Navigation task," is a simulation that can be run either by humans through a graphical interface, or by an automated agent. The task involves learning to navigate through obstacles in a two-dimensional world. A single agent controls an autonomous underwater vehicle (AUV) that has to avoid mines and rendezvous with a stationary target (goal) before exhausting its fuel. The mines may be stationary, drifting, or seeking. Time is divided into episodes. An episode begins with the agent on one side of the mine field, and random target and mine locations; it ends with one of three possible outcomes: the agent reaches the goal (success), hits a mine (failure), or exhausts its fuel (failure). Reinforcement, in the form of a binary outcome, is received at the end of each episode. An episode is further subdivided into decision cycles corresponding to actions (decisions) taken by the agent.

The agent has a limited capacity to observe the world it is in; in particular, it obtains information about its proximal environs through a set of seven consecutive sonar segments that give it a 90 degree forward field of view for a short distance. Obstacles in the field of view cause a reduction in sonar segment length; one mine may appear in multiple segments. The agent also has a range sensor that provides the current distance to the target, a bearing sensor that indicates the direction in which the target lies, and a time sensor that measures the remaining fuel. A human subject performing this task sees visual gauges corresponding to each of these sensors. The turn and speed actions are controlled by joystick motions. The turn and speed chosen on the *previous* decision cycle are additionally available to the agent.

Evidence of Task Complexity: Building an Optimal Controller

Given its delayed reward structure and the fact that the world is presented to the agent via sensors that are inadequate to guarantee correct identification of the current state, the Navigation world is a partially observable Markov decision process (POMDP). The state space defined by the sensors for the NRL Navigation task is about 10^{24}; optimal controllers for POMDPs have been constructed *tabula rasa* only for state spaces on the order of 100 states (Cassandra, Littman, & Kaelbling, 1994) because of the time (and therefore, sample) complexity.[1]

Our motivation for building an optimal controller for the task is two fold: first, it gives us an upper baseline for comparison with human performance; second, it allows us to independently analyze the complexity of learning the task without considering constraints imposed by human learning. The task analysis allows us to ask: (1) what is hard about learning the task? (2) what is an appropriate decomposition of the task to learn an optimal controller (3) what is an appropriate discretization for the task to learn an optimal controller with a bounded amount of training? (4) what is the role of action models in the learning process?[2] (5) is an optimal controller stochastic or deterministic? Answers to these questions help us understand the task better and indirectly guide the design of a suitable space of alternative architectures for modeling human learning.

Since the focus of our present paper is the cognitive modeling of human performance on the task, we provide a summary of the answers to the above questions that are relevant to our present goal. The theoretical and experimental details of our investigation of the design of optimal controllers for this task are in Subramanian and Gordon (1997).

An optimal controller for this task achieves a performance score of 100% for the task configuration of 25 mines, small mine drift and no sensor noise. This is also the task configuration for our human experiments on this task. The optimal controller was created by reinforcement learning (Gordon & Subramanian, 1996a). The learner was initialized with a controller with a specific task decomposition, a specific abstraction of the state space that significantly reduced the complexity of learning, and with a correct but incomplete action choice policy. These three aspects of the initial controller are described in detail below. It should be noted that *tabula rasa* reinforcement learning failed to achieve over a 3% success rate even with training runs in excess of 10,000 episodes.

The structure of the optimal controller reflects the decomposition of the task into two subtasks: avoiding mines and heading toward the target. As we shall show later, this decomposition is also the one adopted by humans. The partial action policy states that when the sonars indicate proximity to mines, the optimal action is chosen to achieve the avoidance subgoal; when far from mines, the optimal action is based on the bearing sensor and the target achievement subgoal. This controller is tuned by reinforcement learning to acquire the appropriate cutoffs on sonar values to switch between the avoidance and target achievement subgoals. For this task configuration, we show that a uniform discretization of all sensor values into three qualitative ranges is sufficient to represent the optimal controller; this causes a reduction in the state space from 10^{24} to 729! Since the learning is very rapid, the results on the utility of learning action models in this domain are not very clear-cut. The fairly coarse discretization in both the sensor and the action space forces the optimal controller to be stochastic.

Our experiments with the construction of an optimal controller highlight what is difficult about this task: it is computationally infeasible to learn the task without an appropriate task decomposition. The *tabular rasa* reinforcement learner shows that acquiring the optimal strategy for this task based purely on experience in interacting with the simulation is nearly impossible. This is because each episode is up to 200 steps long and has a single binary reward at the end, which makes credit assignment extremely difficult. Human learners bring their experience in navigation to bear on this task and are already equipped with the right task decomposition. The optimal controller experiments also show the need for building an appropriate discretization of the sensor values.[3] The action choice policy (mapping from sensor state space to action) needs a compact representation, and our experiments show that a fairly coarse discretization suffices to represent it. How humans discretize the task will be an important component of our cognitive model of learning performance on the task. The optimal controller handles partial observability by maintaining sensor history. Knowledge of action in the previous time step is all that is needed for this task configuration involving 25 mines. Finally, the key strategic aspect in this task appears to be learning when and how to shift attention between the two subtasks.

Data from Human Subjects

In the experiments with humans, seven subjects were used, and each ran for two or three 45-minute sessions with the simulations. We instrumented[4] the simulation (Gordon et. al., 1994) to gather execution traces for subsequent analysis. We also obtained verbal protocols by recording subject utterances during play and by collecting answers to questions posed at the end of the individual sessions.

Two striking results we got from our data with the human subjects were (1) the fundamental similarities in task decomposition (avoid mines; navigate to target) employed by subjects and (2) the remarkable differences in individual learning and performance on this task. For example, see Figure 1, with the best and worst learning curves of the subjects. The verbal protocols, combined with the learning curves, suggest the need for a core model that captures similarities in the conceptualization of the task, and parametric variations on the core model that account for performance differences.

[1]Our navigation problem has different dynamics than the ones faced by animals like rats and ants (Gallistel, 1990), which have a richer sensor base and can use higher level features like landmarks.

[2]We conjecture that it accelerates the rate of learning.

[3]The relationship between state space discretization and value function approximation is in Moore and Atkeson (1995) values, while methods of state aggregation are detailed in Singh, Jaakola, and Jordan (1995). Our own current work (Subramanian & Gordon, 1997) explores this connection as well as algorithms for state aggregation for very high dimensional discrete state spaces. This paper only focuses on cognitive modeling and not on the automatic generation of the optimal controller.

[4]Note that although human subjects use a joystick for actions, we do not model the joystick but instead model actions at the level of discrete turns and speeds (e.g., turn 32 degrees to the left at speed 20). Human joystick motions are ultimately translated to these turn and speed values before being passed to the simulated task. Likewise, the learning agents we construct do not "see" gauges but instead get the numeric sensor values directly from the simulation (e.g., range is 500).

A Cognitive Model

Our goal is to build the simplest model that accounts for human subject data in learning performance. In particular, some subjects become proficient at this task (no sensor noise, 25 mines) after only a few episodes. Modeling such an extremely rapid learning rate presents a challenge. In developing our learning methods, we have drawn from both the machine learning and cognitive science literature. In this section, we briefly describe our basic cognitive model, M_{focus}, previously reported in Gordon and Subramanian (1996b).

One of the more striking aspects of the verbal protocols we collected was that subjects exhibited a tendency to build internal models of actions and their consequences, i.e., *forward models* of the world. These expectations produced surprise, disappointment, or positive reinforcement, depending on whether or not the predictions matched the actual results of performing the action. For example, one subject had an expectation of the results of a certain joystick motion: "Why am I turning to the left when I don't feel like I am moving the joystick much to the left?" Another expressed surprise: "It feels strange to hit the target when the bearing is not directly ahead." Yet a third subject developed a specific model of the consequences of his movements: "One small movement right or left seems to jump you over one box to the right or left," where each box refers to a visual depiction of a single sonar segment in the graphical interface. Therefore, our cognitive model uses action models to predict the consequences of actions. We believe that even though the evidence for the use of action models in the optimal controller is unclear, it is an essential component for modeling human performance on this time-critical task – i.e., humans compensate for their limited processing speeds and memory on this task by anticipating events at least one step into the future. Jordan and Rumelhart (1992) emphasize the critical role of a forward, projective element in cognitive models.

Our cognitive model M_{focus} has four components:

A_{sonars} : sensors $\times$ actions $\rightarrow$ sonars
$A_{bearing}$: sensors $\times$ actions $\rightarrow$ bearing
P_{sonars} : sonars $\rightarrow \Re$
$P_{bearing}$: bearing $\rightarrow \Re$

The A mappings (action models) predict the next sonar and bearing readings given all current sensor readings and the currently chosen action. The P mappings rate the desirability of the sonar and bearing configurations. For sonars, high utilities are associated with large values (no or distant mines), and for the bearing sensor high utilities are associated with values closer to the target being straight ahead. Our cognitive model factors the prediction of sonar and bearing values into A_{sonars} and $A_{bearing}$ and the assessment of the desirabilities of sonar and bearing configurations into P_{sonars} and $P_{bearing}$. This factorization reflects the task decomposition used by our subjects that is revealed consistently in the verbal protocols: mine avoidance depends on sonar readings, and target achievement relies on bearing readings. Currently, P_{sonars} and $P_{bearing}$, which reflect background relevance knowledge about the task, are supplied by us, while A_{sonars} and $A_{bearing}$ are learned by direct interaction with the simulation.

The bearing predictions are discretized into 12 values in clock notation; the sonar predictions (with 220 numeric possibilities) are discretized into five equi-spaced qualitative categories (no-mines, mine-far, mine-mid, mine-close, mine-very-close) for the group of seven segments. The action set consists of three turns: turn-right, turn-left, or go-straight, at a fixed speed (20/40). At each time step, a *focus heuristic* is used to pick one of the pairs (A_{sonars}, P_{sonars}) or $(A_{bearing}, P_{bearing})$ to select an action. The focus heuristic states that if all of the sonar values are below a certain empirically determined threshold (150/220), the pair (A_{sonars}, P_{sonars}) picks the next turn; else $(A_{bearing}, P_{bearing})$ is chosen for picking the next turn. Actions are selected by performing a one-step lookahead of the current state using the appropriate A mapping, and by picking the action that maximizes the corresponding P value of the projected state.

We next investigate two key architectural questions. First, what impact does our task decomposition have on the learning rate? Second, what is the nature of the sensory predictions: are they sufficiently consistent to be a part of the core model, or should they be a parameter that can vary? If the latter, what are the performance tradeoffs between variations?

A Study of Focus of Attention

The verbal protocol data provides abundant evidence that subjects shift their focus of attention between avoiding mines and navigating to the target. As stated earlier, avoiding mines involves reliance on the sonar gauge, whereas navigation generally employs the bearing gauge. All of our subjects ranked the sonar gauge as the most important and bearing as the second most important.[5] Subjects appeared to use the strategy: "When mines are close, avoid the mines. When they are not, navigate towards the goal." Evidence in the protocols for the focus heuristic includes statements such as "I allow the bearing to vary anywhere within view until there are no more mines in front of me – then I pay attention to the bearing of the goal."

Arbib and Liaw (1995) note analogous arbitration between approach and avoidance behaviors in frogs. The default perceptual schema recognizes "all moving objects" and activates the accompanying motor schema of snapping. However, when the pretectum detects a "large moving object," this perceptual schema is activated, which then activates the accompanying "avoid" motor schema, thereby overriding and suppressing the default snapping schema.

To test the impact of our task decomposition (focus heuristic) upon the learning rate, we have ablated this aspect of our cognitive model by lumping the prediction of the next sensors into a single map A, and the evaluation of the sonar and bearing readings into a single utility assessment P. This version of our model, $M_{nofocus}$, projects the composite next set of sensors and chooses actions that optimize the composite P value of the projected sensor set.

We empirically test the following hypothesis:

- *Hypothesis 1:* The slope of M_{focus}'s learning curve is closer than $M_{nofocus}$'s to the slope of subject 1's learning curve, for the Navigation task.

[5] Many of the subjects of Drs. Ron Sun and Edward Merrill at University of Alabama also gave this gauges ranking on this task.

The justification for Hypothesis 1 is that subjects are using this task decomposition (focus heuristic) because it improves their learning and performance on the task. We choose to compare here, as well as throughout the experiments, with subject 1 because out of all seven subjects, subject 1's verbal protocols best reflect the decomposition and prediction issues studied in this paper (e.g., subject 2 struggled a lot with speed selection problems).

The experimental tests of our hypotheses are divided into a training (learning) phase and a testing phase.[6] Training phase length is varied at 25, 50, 75, and 100 episodes. For each training length, all variants of the model see the same training data. The testing phase remains fixed at 400 episodes.[7] Each episode can last a maximum of 200 time steps, i.e., decision cycles. In all experiments, the number of mines is fixed at 25, there is a small amount of mine drift, and no sensor noise. These task settings match exactly those used in the human subjects experiments. Performance is averaged over 10 experiments because the algorithms are stochastic during training, and testing results depend upon the data seen during training. In the graphs, curves show mean performance on the task. Standard deviation bars are at each data point.

We compare the variants of the model with subject 1's learning curve. Note that we cannot divide the human learning into a training phase and a testing phase during which the human stops learning. The curve of the human has performance averaged over a sliding window of 10 previous episodes.

Figure 2 shows the results of testing Hypothesis 1. M_{focus}, which has the stated task decomposition and the focus heuristic, better models the subject's learning curve and statistically significantly outperforms $M_{nofocus}$. Thus, our hypothesis is confirmed and we see the value of dividing the task into two subtasks and modeling the shift of focus between subtasks.

Because there is indication that this task decomposition (focus heuristic) is widely employed and can yield large benefits in performance, we further test its value on an alternative (reinforcement learning) architecture. We use a standard q-learner (Watkins, 1989), that we modified for this task to allow for fair comparisons with variants of M. The details of the q-learning architectures, called Q_{focus} and $Q_{nofocus}$ for with and without the focus heuristic, are irrelevant here. Details are in Gordon and Subramanian (1996a).

We empirically test the following hypothesis:

- *Hypothesis 2:* The slope of Q_{focus}'s learning curve is closer than $Q_{nofocus}$'s to the slope of subject 1's learning curve, for the Navigation task.

The justification for Hypothesis 2 is that the task decomposition seems to be a good model for the task, independent of the architectural choice.

Hypothesis 2 is also confirmed (see Figure 3). We conclude there is significant value in using our task decomposition.

We note that both hypothesis 1 and 2 were tested using a paired, two-tailed t-test with α = 0.05 (compensating for unequal variances whenever indicated by the F-ratio). All paired differences between learning curves of variants of the model described in this section are statistically significant.

[6] We used C4.5 (Quinlan, 1986), which learns the action models in batch and has high noise tolerance - an advantage for a POMDP.

[7] We experimented with the number of episodes and chose a setting where performance improvement leveled off for all algorithms.

A Study of Sensor Predictions

Our cognitive model M_{focus} has two action models: A_{sonars}, which predicts the qualitative magnitudes of the sonar segments, and $A_{bearing}$, which predicts the magnitude of the bearing, on the next time step. When using these action models, our cognitive model chooses the turn that would yield the "best" next prediction, as evaluated by P_{sonars} or $P_{bearing}$. Evidence in the cognitive literature (Kent, 1981) suggests people learn specific values, but over time these specifics are chunked into relevant categories. For example, although people might memorize every size, color, and shape of birds they have seen, over time they generalize to a prototypical bird.

Rarely did the verbal protocols refer to such specific expectations as "bearing will be slightly to the left." More often the subjects were using coarse categories in their expectations, such as "left," "close," "further," or "larger." Nevertheless, few subjects verbalized their expectations, and the evidence on this topic is less clear than the evidence for the focus heuristic. Some verbal statements indicated magnitude (value) predictions, though the granularity of these predictions varied. Other statements reflected predictions of *change* (derivative) in sensor values. Variation occurred in both inter- and intra-subject protocols. To reflect such a mixture of responses, our cognitive model will be parameterized in this respect.

We compare the learning curves of differences of various versions of our model to better understand the performance tradeoffs. We first compare two versions of our model: M_{focus}, as described earlier, and $M_{focus+deriv}$, a variant of M_{focus} that predicts and evaluates sensor derivatives rather than magnitudes. Derivative predictions are quantized into three categories: increasing, decreasing and no change.

$$Ad_{sonars} : \text{sensors} \times \text{actions} \longrightarrow \text{d(sonars)/dt}$$
$$Ad_{bearing} : \text{sensors} \times \text{actions} \longrightarrow \text{d(bearing)/dt}$$
$$Pd_{sonars} : \text{d(sonars)/dt} \longrightarrow \Re$$
$$Pd_{bearing} : \text{d(bearing)/dt} \longrightarrow \Re$$

Because there is no clear evidence for subjects preferring magnitude versus derivative predictions, we do not have a hypothesis about which will perform better. Figure 4 presents the empirical comparison of the two versions of the model. In retrospect, it is not surprising to find that $M_{focus+deriv}$ outperforms M_{focus} since the former captures the goal of improving sensor values more explicitly and in a much more succinct form. However, what is quite surprising is the *degree* to which $M_{focus+deriv}$ outperforms M_{focus}.

We further test this performance advantage without the use of the focus heuristic to be certain it is independent of this heuristic. To do this, we compare $M_{nofocus}$, which makes magnitude predictions, with $M_{nofocus+deriv}$, which makes derivative predictions.

Figure 5 shows the results of this comparison. The results are quite surprising. Not only does $M_{nofocus+deriv}$ outperform $M_{nofocus}$, but the performance of our model with the derivative predictions is nearly the same regardless of whether it does or does not use the focus heuristic (compare Figures 4 and 5)! Both $M_{focus+deriv}$ and $M_{nofocus+deriv}$ are excellent performers and appear to closely approximate the curve of subject 1. We conjecture that although derivative predictions are more effective than the task decomposition for this particular Navigation task, our subjects also used the task decomposition because people have evolved to solve a wide range of tasks. Approach/avoidance is broadly applicable.

In a final experiment, we test whether $M_{focus+deriv}$ did better than M_{focus} because it had fewer categories, or if it was because they were derivatives. To answer this, we use a version of M_{focus} that makes magnitude predictions but the magnitudes are divided into fewer (nominal) categories than in M_{focus}. The categories chosen best reflect the verbal protocol data. Bearing values are "ahead," "behind," "right," and "left." Sonar values are "far," "mid," and "close." This version of the model is abbreviated $M_{focus+fewcat}$. A comparison of M_{focus}, $M_{focus+deriv}$, and $M_{focus+fewcat}$ is in Figure 6.

Figure 6 suggests that the derivative predictions yield the best performance. Apparently, there is a distinct advantage in predicting the change in sensor values, rather than sensor magnitudes on the next time step, to select a turn. Human vision is designed to notice changes, e.g., see Kent (1981), and our results confirm the value of this design.

The differences between the curves for $M_{focus+fewcat}$ and M_{focus} are not statistically significant (α = 0.05). All other paired differences between learning curves of variants of the model in Figure 6, as well as all other figures in this section, are significant (α = 0.05).

Although the verbal protocol data indicates mixed usage of prediction types, our results here show the tradeoffs between the different types. To model human learners, the most accurate model is one that can be parameterized to reflect inter- and intra-individual choices. Future experiments will determine the conditions under which each type of prediction is made so that we can parameterize our cognitive model in this respect.

Discussion and Future Work

The development of an optimal controller for this task, and data collected from experiments with human subjects, have taught us that the NRL Navigation task challenges human learners because: (1) the states are only partially observable, (2) the time-critical nature of the task requires the determination of what is relevant to focus on when, and (3) predictions of the reward and/or sensor values are required for effective, time-constrained learning.

In this paper, we designed a cognitive model of skill acquisition on the NRL Navigation task that captures core similarities in task decomposition in our human subjects. We demonstrated the use of action models in human subjects and constructed variations in the types of predictions supported by these action models. The results from a systematic study of the task decomposition confirm the goodness of fit of our core model to human performance data. The results from our study of magnitude versus derivative predictions by the action models accounts for substantial differences in learning rates.

In the future, we plan to explore other design decisions in our model. We also plan to gather more detailed data about predictions made by subjects, as well as focus of attention (using an eyetracker) to sharpen our understanding of these issues. Related work along these lines evaluates the scanning behavior and mental workload of aircraft pilots, who also make decisions based on gauges (e.g., see Itoh et al., 1990). With more detailed human data, we plan to model individual subjects at a level that will enable us to predict the forms of their trajectories.

Acknowledgements

This research was sponsored by the Office of Naval Research N00014-95-WX30360, N00014-95-1-0846 and N00014-96-1-0538. Special thanks to Bill Spears and Sandy Marshall for their comments and suggestions, and to Helen Gigley and Susan Chipman for their encouragement and support.

References

Arbib, M. A., & Liaw, J.-S. (1995). Sensorimotor transformations in the worlds of frogs and robots. *Artificial Intelligence, 72*, 53-79.

Cassandra, A. R., Littman, M. L., & Kaelbling, L. (1994). Acting optimally in partially observable stochastic domains. *Proceedings of the 12th National Conference on Artificial Intelligence* (pp. 1023-1028). Seattle, WA.

Gallistel, C. R. (1990). *The organization of learning.* Cambridge, MA: MIT Press.

Gordon, D., & Subramanian, D. (1996a). Comparison of action selection learning methods. *Proceedings of the Third International Workshop on Multistrategy Learning, Harpers Ferry* (pp. 95-102). Fairfax, VA: George Mason University.

Gordon, D., & Subramanian, D. (1996b). Cognitive modeling of action selection learning. *Proceedings of the Eighteenth Annual Conference of the Cognitive Science Society* (pp. 546-551). Mahwah, NJ: Lawrence Elrbaum Associates.

Gordon, D., Schultz, A., Grefenstette, J., Ballas, J., & Perez, M. (1994). *User's guide to the navigation and collision avoidance task* (Tech. Rep. AIC-94-013). Washington, D.C.: Naval Research Laboratory.

Itoh, Y., Hayashi, Y., Tsukui, I., & Saito, S. (1990). The ergonomic evaluation of eye movement and mental workload in aircraft pilots, *Ergonomics, 33(6)*, 719-733.

Jaakkola, T., Singh, S. P., & Jordan, M. I. (1995). Reinforcement learning with soft state aggregation. *Advances in Neural Information Processing Systems 7* (pp. 361-368), MIT Press.

Jordan, M., & Rumelhart, D. (1992). Forward models: Supervised learning with a distal teacher, *Cognitive Science, 16*, 307-354.

Kent, E. (1981). *The brains of men and machines*. Peterborough, N.H.: Byte/McGraw Hill.

Moore, A. W. & Atkeson, C. G. (1995). The Parti-game algorithm for variable resolution reinforcement learning in multidimensional state-spaces, *Machine Learning, 21*.

Quinlan, J. R. (1986). Induction of decision trees. *Machine Learning, 1*, 81-107.

Subramanian, D., & Gordon, D. (1997). *Experiments with an optimal controller for the NRL Navigation Task* (Tech. Rep.) Houston, TX: Computer Science Department, Rice University.

Watkins, C.J.C.H. (1989). *Learning from delayed rewards.* Doctoral Thesis, Cambridge University, England.