Chapter 6

Transferring Advice into a Connectionist Reinforcement-Learning Agent

Once the teacher specifies advice for the reinforcement-learning agent, RATLE transfers the advice into the agent. To do this, RATLE must translate the teacher's instructions into a form that the agent can use. In standard reinforcement learning (RL), the agent senses the current world state, chooses an action to execute, and occasionally receives rewards and punishments. Based on these reinforcements from the environment, the agent's task is to improve its action-choosing module so that the total amount of reinforcement it receives is increased. Since my RL agent uses connectionist Q-learning (Lin, 1992; Sutton, 1988; Watkins, 1989), the agent's action-choosing module is a neural network. Therefore, to translate the advice of the teacher, RATLE must make changes to the agent's neural network that capture the meaning of the advice. This is done by translating statements in the RATLE language into new network units and links that correspond to the concepts represented by the statements.

Table 23 shows the main loop of an agent employing connectionist Q-learning, augmented (in italics) by my process for incorporating advice. The main difference between my approach and standard connectionist Q-learning is that the agent continually checks if the teacher has any recommendations to offer, and if so, incorporates those recommendations into its utility function. In order to understand how statements from the RATLE language are transformed, this chapter defines the steps of the subroutine *IncorporateAdvice* from Table 23. But first, I will discuss the steps that must be taken before the agent can begin the process shown in Table 23.

Recall that a connectionist Q-learner uses a neural network, where the input to the network is a vector of features describing the current state, and the outputs of the network are the predicted utilities for each of the agent's possible actions. In RATLE, I assume that the initial neural network is set up by a user (who need not be the teacher), referred to as

Table 23: The cycle of a RATLE agent. My additions to the standard connectionist Q-learning loop are Step 6 and the subroutine *IncorporateAdvice* (all shown in italics).

Agent's Main Loop	*IncorporateAdvice*
1. Read sensors. 2. Stochastically choose an action, where the probability of selecting an action is proportional to the log of its predicted utility (i.e., its current Q value). Retain the predicted utility of the action selected. 3. Perform selected action. 4. Measure reinforcement, if any. 5. Update utility function – use the current state, the current Q-function, and the actual reinforcement to obtain a new estimate of the expected utility; use the difference between the new estimate of utility and the previous estimate as the error signal to propagate through the neural network. 6. *Advice pending? If so, call IncorporateAdvice.* 7. Go to 1.	6a. *Parse advice.* 6b. *Operationalize any fuzzy conditions.* 6c. *Translate advice into neural-network components.* 6d. *Insert translated advice directly into RL agent's neural-network based utility function.* 6e. *Return.*

the *initializer*. As part of defining the agent's inputs and actions, the initializer must also provide names for each input feature using the language described in the previous chapter. The initializer also defines the set of fuzzy terms the teacher may employ in giving advice about the task the agent will be addressing. After the initializer finishes, he or she gives the teacher a description of the input features, actions, and fuzzy terms. The agent then starts the process shown in Table 23.

The routine *IncorporateAdvice* from Table 23 follows Hayes-Roth, Klahr, and Mostow's framework for advice-taking outlined in Chapter 4. Once the teacher gives a set of instructions, RATLE parses them using the standard Unix tools *lex* and *yacc* and the grammars shown in Chapter 5. RATLE then operationalizes any fuzzy conditions in the advice. Fuzzy conditions map to new units in the agent's neural network that capture the appropriate fuzzy membership function. After mapping fuzzy conditions, RATLE translates all remaining parts of the teacher's advice into new units and links. The resulting new units and links are added to the agent's neural network by RATLE and the agent then returns to connectionist Q-learning. The process of adding units and links to a neural network is accomplished by simple extensions to a neural-network simulator. Thus, the key parts of *IncorporateAdvice* are: (6b) how fuzzy conditions are mapped to network units, and (6c) how RATLE translates the other constructs of the RATLE language into neural-network components.

In the following sections, I define how each of the constructs in the RATLE language is translated into corresponding neural-network components. My methods are based on the KBANN algorithm (Towell et al., 1990) which is extended in the following ways: (1) advice can contain multi-step plans, (2) it can contain loops, (3) it can suggest actions to avoid, (4) it can contain fuzzy conditions, and (5) it can be given more than once. Details of these extensions are shown as I discuss how RATLE translates each construct.

The process of translating language constructs into neural-network parts can be divided into separate processes for translating the three basic language components of RATLE: statements, conditions, and actions. A condition represents the teacher's description of some state of the world that must be met in order for an action or actions to be taken. To translate a condition, RATLE creates a hidden unit that is highly active when the condition is true and inactive when the condition is false. RATLE may also create new hidden units combining conditions in order to translate certain types of statements. RATLE connects the resulting units to the action or actions specified by the teacher. In the following sections I show how RATLE combines conditions and actions to map statements, how RATLE turns conditions into corresponding units, and how the units from statements are connected to actions.

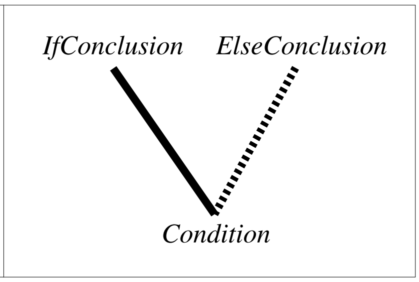
6.1 Translating Statements

RATLE maps statements into neural-network components[1] by connecting the conditions and actions of the statements in appropriate ways. Remember that RATLE maps a condition to a unit that is highly active when the condition is true. To map statements, RATLE adds links from the units representing conditions to the units representing actions as indicated in the statements. As part of representing more complex statements, RATLE may introduce new units that combine conditions.

[1]In demonstrating how RATLE represents the teacher's instructions I often show diagrams containing the construct being translated (on the left) and a corresponding neural-network diagram (on the right). In the neural network part of each of these diagrams, all of the links and units shown are new links and units introduced as part of mapping the construct, unless I specifically state otherwise. Links shown as solid lines represent links with large, positive weights (i.e., as in KBANN's highly weighted links). Straight, dashed lines represent links with large, negative weights, and curved, dashed lines represent recurrent links RATLE uses to connect state units. Arcs connecting links in the diagram indicate units that represent conjuncts, units without arcs show disjuncts. Recall also that the flow of activation in my neural-network diagrams is upwards, from the units at the bottom of the diagram to the units at the top.

6.1.1 Translating the IF Statement

IF *Condition* THEN
  [INFER | REMEMBER] *IfConclusion*
[ELSE
  [INFER | REMEMBER] *ElseConclusion*]
END



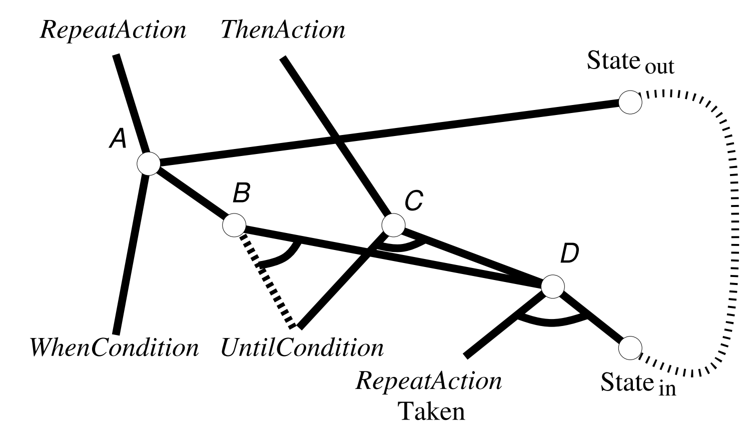
The teacher uses an IF statement to indicate some conclusion that the agent should reach when the unit for *Condition* is true. An ELSE condition can be used by the teacher to recommend a conclusion the agent should reach when *Condition* is false. Recall the three types of conclusions the teacher can suggest: an action the agent should take under that condition, an intermediate term that is true under that condition, and a term to remember. I will discuss the specifics of how these three conclusions are implemented in Section 6.3.1.

To map an IF statement, RATLE first creates a unit representing the condition of the IF (as described in Section 6.2), and then connects this condition to *IfConclusion*. When there is an ELSE part of the statement, RATLE creates a link from the unit representing *Condition* with a negative weight and connects the link to *ElseConclusion*.

Recall that conditions in RATLE can be negated. When this is the case, RATLE negates the weights of the links it constructs (i.e., the link to *IfAction* would have a negative weight and the link to *ElseAction* would have a positive weight). This process is essentially the mapping algorithm of KBANN, which was explicitly designed for IF-THEN rules.

6.1.2 Translating the REPEAT Statement

[WHEN *WhenCondition*]
REPEAT
  RepeatAction
UNTIL *UntilCondition*
[THEN *ThenAction*]
END



The teacher uses a REPEAT statement to indicate that the agent should execute *RepeatAction* as long as *UntilCondition* is false. The statement may have an optional *WhenCondition* that must be true for the agent to start the loop (i.e., to start executing *RepeatAction*). The teacher may also include a *ThenAction* to be taken after the agent finishes the loop, should *UntilCondition* be true.

In order to show how to map a REPEAT statement, I will assume that both a WHEN and THEN structure are used in the REPEAT (the diagram shown above is for this case). RATLE starts by constructing a unit (A) that is true if the loop should execute. It connects this unit to *RepeatAction*.

Unit A is a disjunction of the two possible ways the loop may start: (1) if *WhenCondition* is true, or (2) if the agent has just finished executing the loop and *UntilCondition* is false. RATLE determines that condition (1) holds by constructing a unit for *WhenCondition* and then creating a link from this condition to unit A. The second case requires a conjunction (unit B). Unit B is active when the agent has just finished executing the loop (unit D) and the *UntilCondition* is false.

In order to determine that the agent has just finished executing the loop (unit D), RATLE must check: (1) that the REPEAT statement indicated the loop should have started in the last step, and (2) that the agent actually took *RepeatAction* in the last step. RATLE knows the statement indicated the loop should start in the last step if unit A was true in the last step. Therefore, to know that the loop should have started in the last step, RATLE creates a state unit that remembers the value of unit A from the previous step. To know that the agent actually took *RepeatAction* in the last step, RATLE looks at the input features indicating the agent's last action. Unit D is the conjunction of the input feature indicating that action *RepeatAction* was taken in the previous step and the state unit that remembers the previous value of unit A.

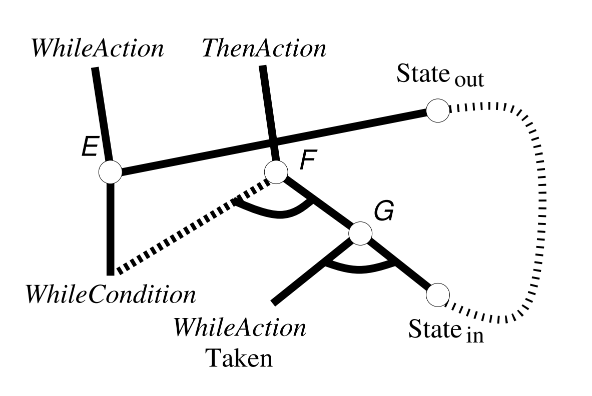
Unit C is constructed for the THEN part of the REPEAT. Unit C is a conjunction that checks that the loop has just finished executing (by making a link to unit D) and that *UntilCondition* is true. RATLE connects unit C to *ThenAction*. This unit is omitted when there is no THEN action.

If there is no WHEN part of the REPEAT statement, RATLE sets up unit A so that it is always true. This is done by setting the bias of the unit to a large positive number. The rest of the REPEAT statement is mapped as described above.

The action *RepeatAction* may be a multi-step plan in a REPEAT statement. If so, RATLE determines that the agent has just finished executing the loop (unit D) by remembering the value of the unit that predicts the *last* step of the multi-step plan, rather than remembering the value of unit A, which would predict the *first* step of the multi-step plan. For more details and an example of how this works, see Section 6.3.4.

6.1.3 Translating the WHILE Statement

WHILE *WhileCondition* DO
  WhileAction
[THEN *ThenAction*]
END



The teacher uses a WHILE statement to recommend that the agent take *WhileAction* as long as *WhileCondition* holds. A WHILE statement may have an optional THEN part, which specifies an action to take when the loop has been executing and *WhileCondition* is false.

A WHILE loop starts for two reasons: (1) *WhileCondition* is true, or (2) the agent has just finished executing the action of the loop and *WhileCondition* is still true. This second case is redundant, since the agent must start the loop when *WhileCondition* is true, regardless of whether the agent has just finished executing the loop. Therefore, RATLE does not check the second case. So, RATLE starts by creating a unit for the *WhileCondition*. A link is then made by RATLE to unit E, which is true when the loop should start. RATLE connects unit E to *WhileAction*.

If there is a THEN part of the WHILE loop (the diagram shows this case), RATLE creates unit F, which recommends *ThenAction*. Unit F is a conjunction that checks that *WhileCondition* is false, and the agent just finished executing the loop (unit G). Unit G is similar to unit D for the REPEAT statement. It is a conjunction that determines that the *WhileAction* was the agent's last action and whether the loop should have started at the last step. The agent determines that the loop should have started in the last step by remembering the value of unit E with a state unit.

6.2 Translating Conditions

RATLE represents a condition in a statement with a unit that is true (i.e., highly active) when the condition is true, and false (i.e., inactive) when the condition is false. RATLE's process is the same as KBANN's for mapping conditions, except that RATLE also supports fuzzy conditions. Since fuzzy conditions involve fuzzy membership functions, RATLE creates a unit that maps each fuzzy function – the more active the unit the better the fuzzy membership criterion is met.

6.2.1 Translating Conditions: Term Names

When the condition of a statement in RATLE is the name of a Boolean input unit or an intermediate term, RATLE represents this condition by simply using the unit that corresponds to the term.

6.2.2 Translating Conditions: Logical Combinations

"(" *Condition* ")"
NOT *Condition*
Condition1 AND *Condition2*
Condition1 OR *Condition2*

RATLE uses the standard KBANN method to map conditions that use the logical operators NOT, AND, and OR. Parentheses in the RATLE language are used by the teacher to indicate the precedence of operators. The negation operator NOT is implemented in RATLE by recursively calling the process for mapping conditions on the condition being negated. RATLE then remembers that links from the negated condition should have negative weights. To translate the AND and OR functions, RATLE maps conditions *Condition1* and *Condition2* to units and then creates a new unit that is a conjunction (AND) or a disjunction (OR) of these units. Note that RATLE also does some network compression when mapping logical conditions; if several units are combined with ORs or ANDs, RATLE will construct a single large disjunctive or conjunctive unit, rather than a series of binary units, up to the limits specified by Towell (1991).

6.2.3 Translating Fuzzy Conditions

To map fuzzy conditions, RATLE creates a hidden unit that captures the fuzzy membership function specified. The set of possible fuzzy membership functions available to the teacher is defined by the initializer before the instruction process begins. Recall the two types of fuzzy conditions: the first type makes use of fuzzy terms that I call descriptors, and the second type makes use of fuzzy terms that I call quantifiers and properties.

Translating Fuzzy Conditions: Descriptors

Object { IS | ARE } *SimpleDescriptor*

Recall that to translate a fuzzy condition of this form, RATLE matches the string *SimpleDescriptor* (and *Object* as well) to the set of predefined descriptors. An example of such a fuzzy condition from the previous chapter is:

`Tree1` IS `CloserThanTree2`

where `CloserThanTree2` was defined by the descriptor:

DESCRIPTOR `CloserThan?(Other)` ::=
(`DistTo?(Object)` - `DistTo?(Other)`) < 0

Instantiating this descriptor (with ?(Object) =`Tree1`) we get:

`CloserThanTree2` ::= (`DistToTree1` - `DistToTree2`) < 0

RATLE translates this term into a unit that calculates the specified sum and uses the threshold (0 in this case) as its bias.

To do this, RATLE uses an approach similar to Berenji and Khedkhar's (1992). RATLE uses the DESCRIPTOR function's possible range of values, the descriptor's operator, and the descriptor's threshold to determine the membership function. For example, assume that `DistToTree1` and `DistToTree2` are input variables that range from 0 to 100. The sum of the calculation above (`DistToTree1` - `DistToTree2`) ranges from -100 to 100. Since the threshold is 0, and the operator is <, I want a unit that is true when the sum is from -100 up to (but not including) 0. But, since the agent uses sigmoidal activation functions for its units, I cannot get a unit with such a discontinuity in the activation value. Instead RATLE selects two threshold values. One value represents the point at which the function will produce high activation and the other the point at which the function produces low activation.

For the above example, RATLE selects the value -0.5, below which the activation will be high, and the value 0.5, above which the activation will be low. For the region between -0.5 and 0.5 the function is "fuzzy" – values in this range partially meet the fuzzy condition. RATLE achieves this effect by weighting the input values being summed so as to calculate the appropriate function. If we define high activation as 0.9 and low activation as 0.1, RATLE would achieve the desired function by adding a link to `DistToTree1` with a weight of 4.39 and a link to `DistToTree2` with a weight of -4.39. RATLE determines these weights using the threshold values -0.5 and 0.5 and the expected activation values of 0.9 and 0.1 in Equation 3 in Chapter 2. Figure 26 shows a diagram of the resulting hidden unit.

The remaining question for RATLE is how to choose the values around which to base the membership function (the values -0.5 and 0.5 in the above discussion). To do this I again follow Berenji and Khedkhar's (1992) approach. I first calculate the range of possible values. Then I select a value up to which the membership function has low activity and a second value where the membership function starts being highly active. I chose to set these values one unit apart (unless the maximum range of the sum being calculated is less than 10, in which case I use two values that are 10% of the maximum range apart). For the "$>$ *Number*" test, RATLE sets the low value to ($Number\ -\ 0.5$) and the high value to ($Number\ +\ 0.5$). For the "$>=$ *Number*" test, RATLE sets the low value to ($Number\ -\ 1.0$) and the high value to $Number$. The range for $>=$ insures that the value $Number$ has a high activation, while for $>$ the value $Number$ is only somewhat active. The $<$ and $<=$ operations work similarly (with the obvious swaps of signs and high and low values). RATLE builds the $=$ operation

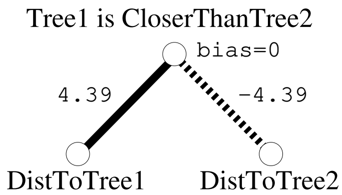


Figure 26: Hidden unit that represents the fuzzy condition `Tree1` IS `CloserThanTree2`. The resulting unit is a sum of the distance to `Tree1` minus the distance to `Tree2`. This unit is neither a conjunction nor a disjunction, since the link weights and the bias are determined by the fuzzy function being mapped. Note that RATLE only creates the unit for the fuzzy condition and the weights leading into that unit; the other units are assumed to be input units.

as a conjunction of two operations: a $>=$ and a $<=$ operation. Similarly, RATLE builds a $!=$ as a disjunction of a $<$ and a $>$ operation.

If more than one fuzzy term matches the fuzzy condition, RATLE builds a unit for each term that matches and constructs a disjunction of the resulting units.

Translating Fuzzy Conditions: Quantifiers

Quantifier Object { IS | ARE } *PropertyDescriptors*

RATLE translates the other type of fuzzy condition similarly. There are two main differences with this type of fuzzy condition: properties are used to select inputs that match those properties; and the resulting function may be applied to each input that matches separately or may be summed across all the inputs that match. When the teacher specifies a fuzzy condition of this form, RATLE starts by looking for inputs of the appropriate type, and then determines which inputs match the specified properties.

For example, when the fuzzy condition is:

`Many Trees` ARE (`Tall` AND `Leafy`)

RATLE would look for input features of type `Trees` and then determine how well each of these inputs match the properties `Tall` and `Leafy`. For instance, if an input represents `Trees` of heights 40 to 60 and `Tall` is defined by the initializer as being of height 50 to 80, this input unit matches with a value 0.5 (since half of the range 40 to 60 is considered `Tall`). When there is more than one property, RATLE multiplies the match values (e.g., in the above example, if the input described above matched the property `Leafy` with a value 0.6, the total match value of that unit is $0.3 = 0.5 * 0.6$). Once RATLE determines how well each input matches the properties in the teacher's fuzzy condition, it calculates the appropriate weight as described in the previous section, but it then multiplies the weight for each link by that input's match value.

The other difference for this type of fuzzy condition is that two different types of functions can be calculated. In one function (when SUM is specified in the quantifier) RATLE calculates the appropriate function by determining all of the input units that match and then calculating the total input across all of the matching inputs. If SUM is not specified in the quantifier, RATLE calculates the appropriate function with respect to each of the matching inputs, and then creates a disjunction of each of these units. Figure 27 shows a sample interpretation of

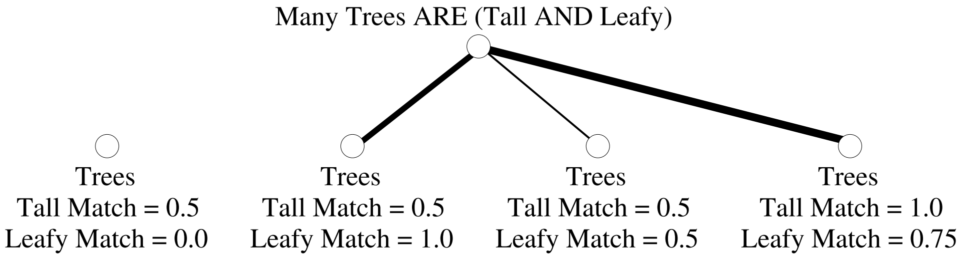


Figure 27: Unit that represents the fuzzy condition `Many Trees` ARE (`Tall` AND `Leafy`), assuming that `Many` is defined as a SUM quantifier. Note that the weights on the links from the inputs are affected by how the the input matches the properties `Tall` and `Leafy`. This unit totals the number of trees across the matching input units and then calculates the fuzzy operator associated with `Many`.

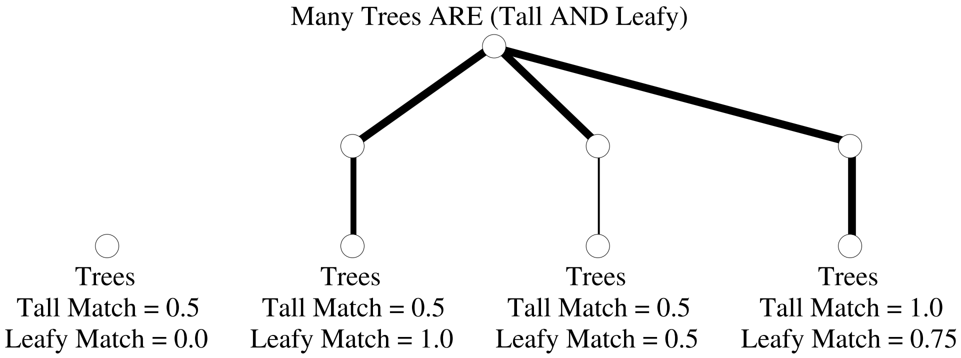


Figure 28: Unit that represents the fuzzy condition `Many Trees` ARE (`Tall` AND `Leafy`), assuming that `Many` is not defined as a SUM quantifier. RATLE creates a unit for each matching input unit to determine when individual input unit meets the condition, and then creates a disjunction to determine if there at least one matching input unit.

the condition above assuming that `Many` is defined as a SUM across all the matching input units. In this case, the resulting hidden unit totals how many `Tall` and `Leafy Trees` there are across all the input features to determine whether there are `Many Trees`. Figure 28 shows the same fuzzy condition should SUM not be specified in defining `Many`. Here, RATLE constructs hidden units that determine whether any of the input features that match the properties `Tall` and `Leafy` meet the criterion *Many*. The unit representing the fuzzy condition is the disjunction of all these tests.

Quantifiers may also use the operation IN. RATLE calculates an IN function as a conjunction of a >= and a <= operation.

6.3 Translating Actions

The teacher uses statements to indicate actions that the agent should take under certain conditions. As I have shown above, RATLE builds units to represent conditions that indicate when the agent should take an action. In this section, I give details on how RATLE connects conditions to the different types of actions: single actions, alternative actions, action prohibitions, and multi-step actions.

6.3.1 Translating Actions: Single Actions

Single actions occur in statements where the teacher gives a single name as an action in that statement. Recall that the teacher can also use an IF statement in two other ways, to create or add to the definitions of intermediate terms (using the INFER keyword), and to create and add to memory terms (using the REMEMBER keyword). So I will break down the set of single actions to five cases: (1) creating a new intermediate term, (2) creating a new memory term, (3) adding to the existing definition of an action, (4) adding to the definition of an intermediate term, and (5) adding to the definition of a memory term. There is no case for creating a new action since the set of possible actions is determined by the initializer before the agent starts learning.

The case where the single action is the name of a new intermediate term is straightforward. RATLE first creates a unit corresponding to the condition leading to the action. RATLE then attaches the name of the new intermediate term to this new unit. Figure 29 depicts a sample of this case. Note that should a condition correspond to a unit that already has a name (e.g., a condition that is just the name of an input), RATLE creates a new unit with a link to the condition. RATLE assigns the name of the intermediate term to this new unit.

```
IF (Small OR Medium) THEN
  INFER NotLarge
END
```

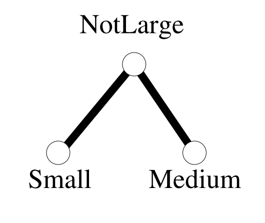


Figure 29: Translating a new intermediate term creates a new hidden unit that RATLE labels with the new intermediate term name (in this case, `NotLarge`).

When creating a new memory term, RATLE again constructs a unit representing the condition, but RATLE then adds a state unit with a recurrent link from the unit representing the condition. RATLE then gives the memory term name to this state unit. The unit labelled `NotLarge` in the advice in Figure 30 will be active when the condition `Small` OR `Medium` was active at the last step.

In the three remaining cases (adding a definition for an existing intermediate term, a memory term, or an action), RATLE has to add to the definition of a unit that already exists. For the first two cases (adding to an intermediate term or memory term) RATLE uses a similar method, while it uses a much different approach for adding to the definition of an action (see Figure 31).

For additional definitions of existing intermediate terms and memory terms, RATLE creates a new hidden unit. This new hidden unit is a disjunction of the old definition of the intermediate or memory term and the new definition. For actions, RATLE uses a different approach: it connects the condition associated with an action directly to the unit representing the action. I use this approach because the output units in a RL agent are not like units in a traditional KBANN network – an output unit represents the utility that the agent expects to receive by following the action associated with the output unit. Thus, creating a disjunction of the old action unit with the new condition is not appropriate, since the old action unit does not have a Boolean value.

For the units representing actions, RATLE must try to achieve the goal of the teacher. Unfortunately, there are many possible interpretations for a teacher's instruction. For example, when a teacher suggests an action, she could be indicating that the action will have a high utility. On the other hand, the teacher could be saying that the action is the best thing to do under those conditions, but that the action is merely the best of a set of bad choices – it may have a low expected utility even though it is the "best" action. A third

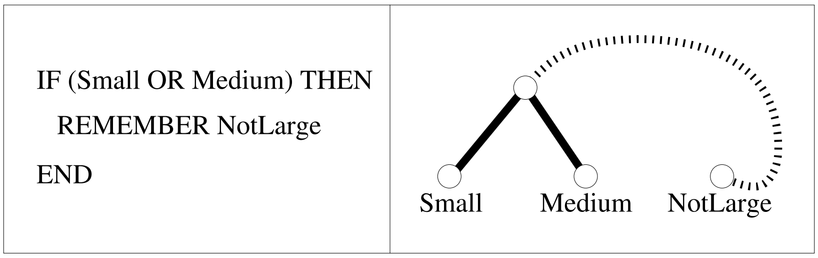


Figure 30: Translating a memory term creates a new state unit that RATLE labels with the new memory term name (e.g., `NotLarge`).

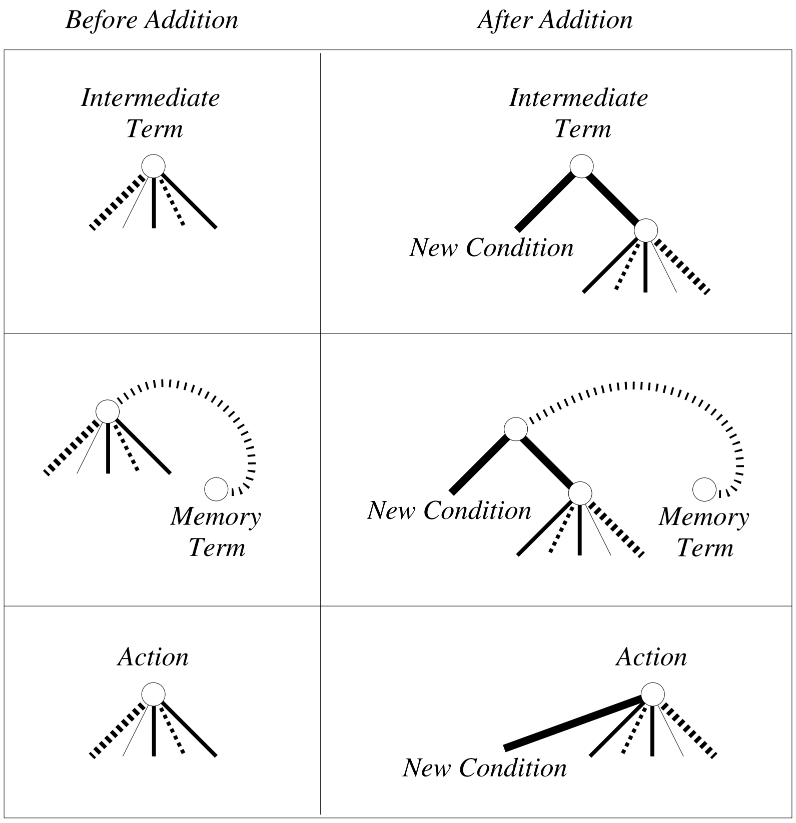


Figure 31: Adding a new definition of an existing intermediate term, memory term, or action changes the definition of an existing hidden unit in the agent's network. RATLE adds a definition to an existing intermediate term or memory unit by creating a unit that is a disjunction of the old term and the condition specifying the new definition. For an action, only a highly-weighted link from the condition to the action is added by RATLE.

interpretation is that RATLE should examine the network and reset the weights so that the suggested action is always the best choice. This approach can be faulty if the action is already considered the best of several bad choices. In this case, RATLE may conclude that no further work needs to be done since the action is already preferred. But the teacher's advice may be indicating that the action is not only the best choice, but that the utility of the action is high (i.e., the action is "good" in some sense). Therefore, simply determining that the action is the "best" choice may not always capture the teacher's intentions. This approach also requires that RATLE find an appropriate set of weights to ensure that the suggested action be selected first. Determining these weights appears to be equivalent to solving a non-linear programming problem.

I chose the simplest interpretation for the teacher's recommendation of an action – RATLE

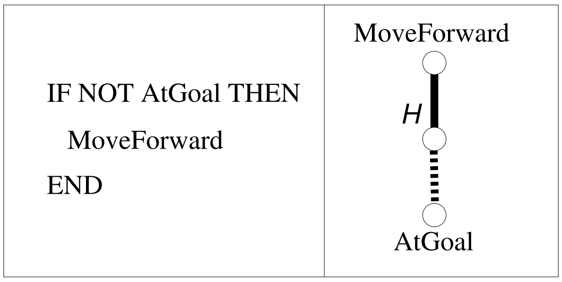


Figure 32: A negated condition leading to an action causes RATLE to introduce a new unit (H) that is true when the condition is false. RATLE then adds a positively weighted link from unit H to the recommended action.

simply connects the new condition to the action with a highly weighted link. This has the effect of increasing the predicted utility of the action under the conditions suggested by the teacher (but does not guarantee that the resulting utility will be the highest). Through empirical testing I found that weights of magnitude 2.0 are effective for these links.

Because of this approach for adding links to action units, RATLE may also have to introduce an extra hidden unit when the link being connected to the action is negated. Consider the IF statement shown in Figure 32. RATLE first maps the condition of the IF statement, representing `AtGoal` with a unit. It also remembers that the condition is negated, and therefore the weight from `AtGoal` should be negative. If RATLE then simply connected this negative weight to `MoveForward` this would cause the advice to have the effect "when `AtGoal` is true, do not move forward," which obviously does not match the instruction's intent. To avoid this problem, RATLE introduces a new hidden unit (H) that checks the condition (NOT `AtGoal`) – this unit is true when `AtGoal` is false. RATLE then creates a link from unit H to the unit for `MoveForward` with a positive weight. This link captures the idea that when unit H is true (which means that `AtGoal` is false) the utility of `MoveForward` should be increased.

6.3.2 Translating Actions: Alternative Actions

The teacher uses an alternative-action structure to indicate that more than one action could be taken under a particular condition. In the IF statement in Figure 33, the teacher indicates that both `MoveNorth` and `MoveEast` are appropriate when the condition `GoalIsNorthAndEast` is true. RATLE creates a link from the unit representing `GoalIsNorthAndEast` to the unit representing the action `MoveNorth` and another link from `GoalIsNorthAndEast` to `MoveEast`. One link is created by RATLE for each alternative action.

IF GoalIsNorthAndEast THEN
(MoveNorth OR MoveEast)
END

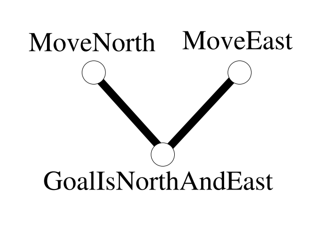


Figure 33: When the teacher indicates more than one action is appropriate given a condition, RATLE creates a link to each of the appropriate actions.

6.3.3 Translating Actions: Action Prohibitions

The teacher can indicate that the agent should not take an action. RATLE implements an action prohibition by connecting a negatively-weighted link to the action being prohibited. Note that, as with the case of single actions described above, RATLE may need to introduce a new unit if the link it is trying to attach is already negative. Figure 34 shows an example of such a piece of advice. The link from the condition `AtGoal` must be negated since the condition is (NOT `AtGoal`). Should RATLE negate the weight of this link again to represent the idea that the action is prohibited, connecting this link to `ApplyBrake` produces the effect that `AtGoal` implies `ApplyBrake`, which may be true, but does not capture the advice. RATLE solves this problem by creating a unit (J) to which it connects the negated link from `AtGoal` – unit J is true in Figure 34 if `AtGoal` is false. RATLE then adds a negative link from the unit J to `ApplyBrake` (this link reduces the utility of `ApplyBrake` when `AtGoal` is false).

IF NOT AtGoal THEN
DO_NOT ApplyBrake
END

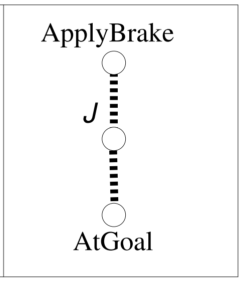
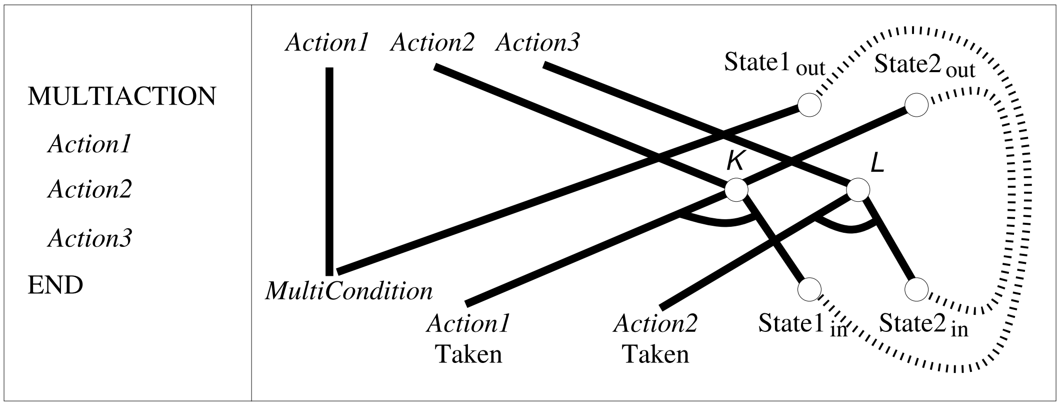


Figure 34: When the condition of an action prohibition is negated, RATLE creates a unit (J) that is true when the condition is false, and then connects unit J to the action with a negatively weighted link.

6.3.4 Translating Actions: Multi-Step Plans



A MULTIACTION is used by the teacher to indicate a sequence of steps that the agent should take. The teacher may use a MULTIACTION in any statement where an action is allowed. In the diagram above, I show just the MULTIACTION part of the network (the condition *MultiCondition* would be determined as part of mapping the rest of the statement).

RATLE translates a MULTIACTION using a set of state units to track the step of the plan. It connects the *MultiCondition* that indicates the start of the plan to the first step of the plan (i.e., *Action1*), and creates a copy of this link that it connects to a new state unit ($State1_{out}$). This state unit remembers that the condition *MultiCondition* was true at the previous time step. RATLE then creates a unit (K) that is the conjunction of the input value of *State1* plus a link to the input unit that indicates whether *Action1* was taken in the previous state. A link is created by RATLE from unit K to *Action2*, since K indicates that the second step of the plan should be taken. A copy of this link is connected to *State2* – state unit *State2* remembers the value of unit K from the previous step. RATLE then repeats the process, creating unit L which predicts *Action3*. Note that RATLE does not have to create a unit that remembers whether *Action3* was indicated, since for the MULTIACTION shown there is no fourth action.

MULTIACTIONs require RATLE to generalize the translation of REPEAT and WHILE statements presented earlier. This is needed because a loop does not end until the last action of a MULTIACTION has been taken. When a REPEAT loop predicts a single action, RATLE construct a state unit to remember the condition predicting that the first (and only) action of the loop should be taken. But when a loop contains a MULTIACTION, RATLE uses a state unit to remember the condition that predicts the *last* action of the multi-step plan. RATLE would use this state unit to determine if the loop has just finished executing.

As an example, consider the second piece of advice from Table 19. Figure 35 shows

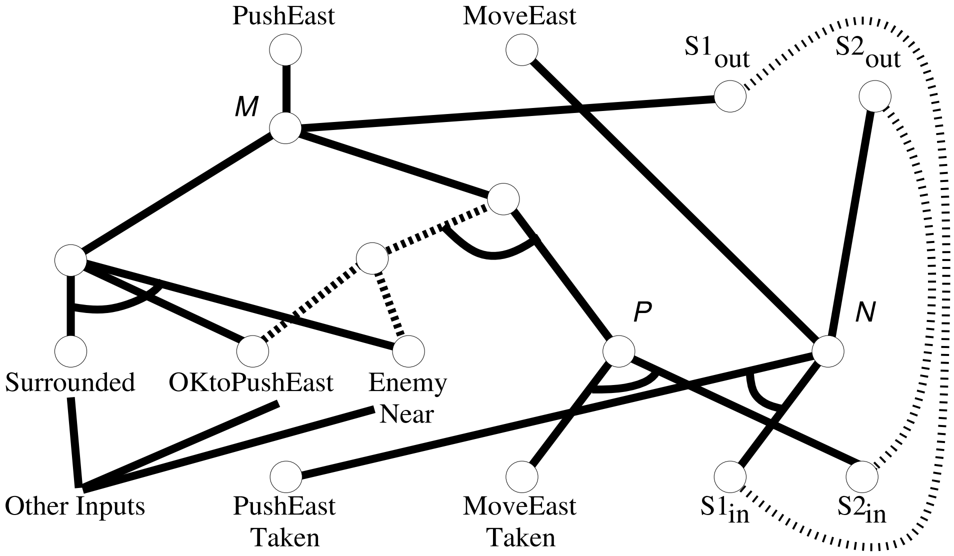


Figure 35: RATLE's translation of Table 19's second piece of advice.

RATLE's mapping of this piece of advice. This REPEAT statement contains a MULTIACTION within the loop. If the REPEAT statement had specified a single action, RATLE would have constructed a state unit to remember the value of unit M, the unit indicating that the first (and only) action of the loop should be taken. But this loop contains a MULTIACTION, so RATLE instead remembers the value of unit N with a state unit S2. Unit N recommends the last action of the MULTIACTION. RATLE creates a unit P to determine if the last action of the plan was recommended at the last step and that the agent actually took that action. Unit P is a conjunction of the input value of S2 and the unit that indicates whether the action `MoveEast` (the last action of the MULTIACTION) was taken in the last step. RATLE uses unit P to indicate that the loop has just ended.

6.4 Limitations of the RATLE Implementation

In implementing the RATLE system, I made a number of design decisions that greatly influence the performance of the system. In this section I outline the reasons for some of these decisions and discuss possible alternatives.

6.4.1 Mapping Action Recommendations

One basic issue, which I have already discussed in detail (see Section 6.3.1), was my decision to translate advice suggesting an action by simply increasing the utility of that action. There are many possible interpretations for a piece of advice, and therefore there are many ways to translate a piece of advice. My choice of increasing the utility by a fixed amount has the advantage of being simple (as opposed to some of the more complex interpretations I discussed previously). Besides using more complex mapping methods, I could also augment the RATLE language to require the teacher to indicate exactly what is meant by specifying an action (e.g., provide a keyword "BEST_ACTION" to indicate that the teacher thinks a particular action is the best). The problem with this approach is that it makes the process of building advice more difficult for the teacher.

6.4.2 Making Network Additions to Represent Advice

One possible problem with the current RATLE implementation is that each time the teacher provides advice, RATLE adds new hidden units and links to the agent's neural network. As the agent's neural network grows, the learning process becomes slower, since a larger network requires more computation. Also, an overly large neural network can often produce overfitting (Holder, 1991). One solution to this problem (see Chapter 2) introduces a penalty term to prevent overfitting, such as weight decay (Hinton, 1986), which can cause the learner to gradually remove unused links. A related approach periodically prunes the network (Le Cun et al., 1990) to find and remove hidden units that are no longer being used. Utilizing these approaches is a subject of future work. Other approaches to advice-taking that do not suffer from this problem will be discussed in Chapter 8.

Another method to deal with the problem of an ever-growing network would be to introduce a mechanism for the teacher to remove advice that is no longer in use. The teacher could give "FORGET" commands that would indicate antecedents or intermediate terms that RATLE should remove from the agent's neural network. One reason this approach is unsatisfactory is that it expects the teacher to better understand the internal workings of the agent so that the teacher can monitor when the agent's neural network has become unwieldy. I would prefer to reserve this method as a supplementary approach, allowing the teacher to suggest terms to remove, but not requiring that the teacher provide this type of advice.

6.4.3 Decaying State-Unit Activations

The activation of the state units in RATLE disappears after one time step. For example, when an agent is executing a multi-step plan, the agent must execute each step of the plan consecutively or the plan will no longer be active. This approach limits the ability of the agent to pause while executing a multi-step plan to perform opportunistic actions. In order to produce such a pause, the agent would have to use its Q-learning and experiences to learn that it is advantageous to pause, and refine the advice to reflect this understanding.

An alternative method would be to have the activation value of state units decay over a number of steps, so that the agent could resume a plan even after several intervening actions. While this approach is appealing for certain situations, in other situations the teacher may be indicating that the agent does indeed need to perform all of the steps of the plan in order, in which case having state units that decay would be undesirable. Another implementation problem of this approach is that we really want only the state unit corresponding to the current action in a multi-step plan to be active. If state units for previous actions are still partially active, the agent might take the action again, which would not capture the sequential aspect of the plan.

6.4.4 Adding "Extra" Hidden Units when Representing States

The reader may have noted that when I create state units I always add a hidden unit that records the value of the unit to be remembered, and then connect that hidden unit to an input with a recurrent link. I could simply connect the unit whose value is being remembered directly to the new input unit with a recurrent link. For example, in the diagram for translating a MULTIACTION in Section 6.3.4, I could leave out the unit $State1_{out}$ and simply connect *MultiCondition* to $State1_{in}$ directly. I chose not to implement states this way because in certain cases a state input could not be changed. If *MultiCondition* in the diagram in Section 6.3.4 was the name of input unit, and I made a direct connection from this input unit to $State1_{in}$, the activation value of this state would always be the same as the previous input unit value, since no learning occurs along the recurrent links. By introducing $State1_{out}$, I allow the agent to alter this state value, since the weight on the link between the unit representing *MultiCondition* and $State1_{out}$ can change.

6.4.5 Mapping Fuzzy Membership Functions with Sigmoidal Activation Functions

Another choice I made was to use sigmoidal activation functions to implement RATLE's fuzzy membership functions. I chose sigmoidal activation functions mostly for simplicity, since I was already using sigmoidal units for mapping other constructs. Other activation functions might make more sense for capturing fuzzy functions, especially radial basis functions (Moody & Darken, 1988, 1989). A unit using a radial basis function is defined by a prototype point in input space. The unit is more active the closer the current input vector is to the prototype (a second parameter determines how dispersed the activation function is). For a fuzzy function this would make sense, since the fuzzy membership function could be captured by selecting a point that is the "center" of the membership function and the dispersion parameter would depend on how much of the range of possible input values are to be considered part of the membership function. Choosing this type of activation function might simplify representing other types of fuzzy functions.

6.5 Summary

RATLE translates the teacher's statements in the RATLE language into additions to the agent's current neural network. It is important to remember that the resulting parts of the neural network do not have to be correct – the agent can adjust the advice, including fuzzy conditions, given its subsequent experiences.

To translate advice, RATLE works like the KBANN algorithm. RATLE converts the condition of a statement to a corresponding unit that is true when the condition is true (i.e., highly active) and false (i.e., inactive) otherwise. The resulting unit, which may be combined with other units as part of mapping a complex statement, leads to an increase in the the expected utility of the action suggested by the teacher.

RATLE extends the KBANN algorithm in a number of ways. One, advice in the RATLE language can contain multi-step plans that not only indicate the current action to take, but also subsequent actions to take. Two, RATLE uses state units to implement these plans in a way that is transparent to the teacher. Three, loops in the form of REPEAT and WHILE statements may also appear in the teacher's instructions – RATLE maps these statements using state units that remember the state of the loop from the previous step.

A teacher may also give advice about actions to avoid. RATLE uses weighted links to lower the utility of the prohibited action under the appropriate circumstances. The teacher can also make use of fuzzy terms in her statements. These fuzzy terms (which are generally problem-specific) are defined prior to the start of the instruction process. RATLE uses these fuzzy terms to build fuzzy membership functions that can then be adjusted during subsequent learning.

Once RATLE completes the process of translating the teacher's instructions, it performs KBANN's final process: RATLE adds connections between units at adjacent levels of the network that are not already connected (see Section 2.2 for more details). These connections allow RATLE to refine the teacher's recommendations by adding new antecedents to the conditions the teacher has specified.

Since advice is represented with differentiable sigmoidal units, the agent can refine the connections mapping the advice, including the connections defining fuzzy terms, using back-propagation learning. Thus the agent is able to evaluate and alter the teacher's advice through its experiences performing the task.

Finally, RATLE implements the advice translation process as a series of *additions* to the agent's network. This means that the teacher may instruct the agent any number of times, with each new set of advice introducing new units and links to the agent's network.

Chapter 7

Experimental Study of RATLE

In this chapter, I present experiments to evaluate the RATLE system described in Chapters 4, 5, and 6. The primary goal of these experiments is to explore the basic question of this thesis: does a technique for refining *procedural* domain theories show the same benefits (reviewed below) that techniques for refining non-procedural domain theories show? A secondary focus of my experiments is to exercise the features of RATLE, both to demonstrate their use, and to evaluate their effectiveness. I performed my experiments on two simulated environments: (1) an environment similar to the Pengo video game, and (2) a Soccer environment. Below I outline my hypotheses for the experiments and then discuss the important characteristics of these testbeds.

We generally expect to see two benefits from a theory-refinement technique: (1) instruction should provide a "head-start" for the inductive learner, and (2) the refined domain theory should be more accurate than the original theory. By "head-start" I mean that a student employing advice should see a gain in performance (i.e., cumulative reward) that a student without instruction would only achieve after seeing more task examples (e.g., a student who has driving explained to him should learn faster than a student who has to learn to drive without any instruction). Note that we cannot expect that the student with instruction will forever outperform the student without instruction. Since both students are able to learn, and there is generally a limit to how well one can perform a task, it is possible that both students will eventually achieve the same performance.

I use my testbeds to demonstrate that RATLE produces the two effects discussed above: (1) the RATLE agent shows a gain in performance after receiving advice that a standard RL agent would only achieve through lengthy exploration, and (2) the RATLE agent outperforms an agent that is unable to refine the advice (i.e. an agent that assumes that the advice is always correct in any situation where it applies). I will also analyze the results to show that RATLE agents produce effects that are consistent with the goals of the advice. Other experiments demonstrate that the effects of advice are independent of when the teacher presents the advice, that the agent can overcome the effects of "bad" advice, and that the agent is able to make use of a subsequent piece of advice (i.e., a piece of advice presented after the agent has received an initial piece of advice and returned to reinforcement learning).

In the next section of this chapter I present the extensive experiments I performed on the Pengo testbed. I will outline the important aspects of the task environment, the methodology I use in my experiments, and the results, then present some discussion of those results. Following that, I present supporting results on the Soccer testbed.

7.1 Experiments on the Pengo Testbed

I used the Pengo testbed to evaluate the hypotheses I present above. This testbed is similar to those explored by Agre and Chapman (1987) and Lin (1992). Using a similar testbed allows me to build on this previous work. In particular, I can make use of parameter settings empirically determined by Lin.

Like Agre and Chapman's work, my Pengo testbed is based on the Pengo video game. One major difference between my task and theirs is that my agent has only a limited sensor array – the agent is unable to see through objects in its maze-like world, it only receives agent-centered information about objects in direct line-of-sight. In Agre and Chapman's task the agent has an "aerial view" of the entire environment (i.e., the view of a player of this video game). I chose to provide limited sensor information to my agent so that the results would be applicable to real-world robot agents that have similarly limited sensors.

My Pengo testbed is also similar to Lin's in that he also uses a limited set of sensors, though his sensors are not constrained by line-of-sight limitations. My task differs from Lin's in that I have a more complex set of reinforcement signals. In Lin's task the agent focuses only on evading capture by enemies. In my task the agent receives reinforcements in three situations: when it is captured by an enemy, when it picks up food, and when it eliminates an enemy. This more complex set of reinforcement signals makes the Pengo task more difficult to learn, since the agent must learn to balance the different reinforcement signals.

7.1.1 The Pengo Environment

Figure 36a illustrates the Pengo environment. The agent in Pengo can perform nine actions: *moving* and *pushing* in each of the directions East, North, West and South; and *doing nothing*. Pushing moves the obstacles in the environment. A moving obstacle will destroy the food and enemies it hits, and will continue to slide until it encounters another obstacle or the

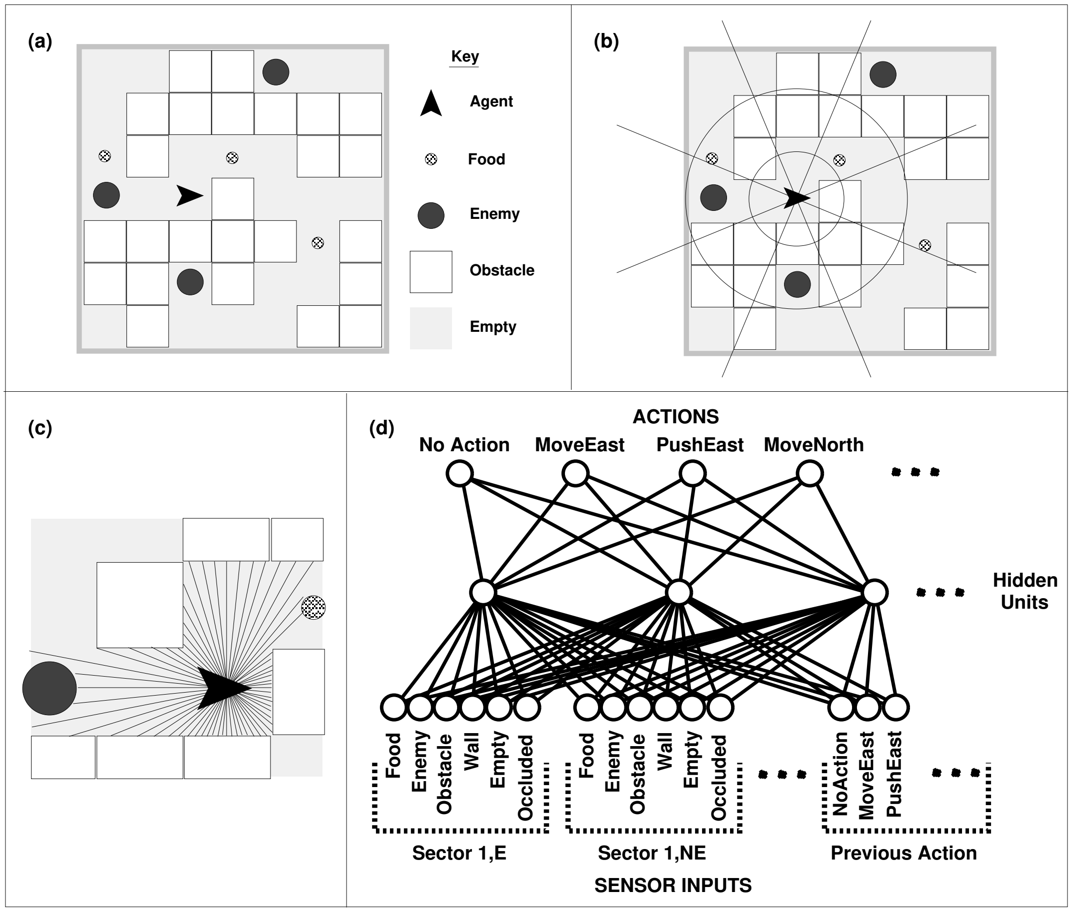


Figure 36: The Pengo environment: (a) sample configuration, (b) sample division of the environment into sectors, (c) distances to the nearest occluding object along a fixed set of arcs (measured from the agent), (d) a neural network that computes the utility of actions. (See text for further explanation.)

edge of the board. If the obstacle is unable to move (there is an obstacle or wall behind it), the obstacle disintegrates when pushed. Food is eaten when the agent or an enemy touches it.

Each enemy follows a fixed policy. It moves randomly unless the agent is in sight, in which case it moves toward the agent. Enemies may move off the board (they appear again after a random interval), but the agent is constrained to remain on the board. Enemies do not push obstacles.

The initial mazes are generated randomly using a maze-creation program that randomly lays out lines of obstacles and then creates connections between "rooms." The percentage of the total board covered by obstacles is controlled by a parameter, as are the number of enemies and food. The agent, enemies, and food are randomly deposited on the initial board, with the caveat that the enemies are required to be at least a fixed distance (another parameter) away from the agent at the start. More details on these processes appear in Appendix A.

The agent receives reinforcement signals when: (1) an enemy eliminates the agent by touching the agent (–1.0), (2) the agent collects one of the food objects (+0.7), and (3) the agent destroys an enemy by pushing an obstacle into it (+0.9). I chose these values empirically during my initial experimentation with the baseline learning agents (i.e., agents without advice).

As mentioned above, I do not assume a global view of the environment, but instead use an agent-centered sensor model. It is based on partitioning the world into a set of sectors around the agent (see Figure 36b). I define each sector by a minimum and maximum distance from the agent, and a minimum and maximum angle with respect to the direction the agent is facing. The agent calculates the percentage of each sector that is occupied by each type of object – food, enemy, obstacle, or wall. To calculate the sector occupancy, I assume the agent is able to measure the distance to the nearest occluding object along a fixed set of angles around the agent (see Figure 36c). This means that the agent is only able to represent the objects in direct line-of-sight from the agent (for example, the enemy to the south of the agent is out of sight). The percentage of each object type in a sector is just the number of sensing arcs that end in that sector by reflecting off an object of the given type, divided by the maximum number of arcs that could end in the sector. So for example, given Figure 36b, the agent's percentage for "obstacle" would be high for the sector to its right. The agent also calculates how much of each sector is empty and how much is occluded. The agent also records as input which action the agent took in the previous state. Appendix A presents the definitions of the input and and fuzzy terms used in these experiments.

The sector percentages and previous action features discussed above constitute the input to the agent's neural network (see Figure 36d). The agent's sensors for these experiments measure objects in 32 sectors (four "rings" divided into eight equal "wedges", where a ring is defined by a minimum and maximum distance and a wedge is defined by a a minimum and maximum angle). In each sector I count the percentage taken up by each of the six objects. There are nine output units for the network representing the nine possible actions.

7.1.2 Methodology

I train the RL agents for a fixed number of *episodes* for each experiment. An episode consists of placing the agent into a randomly generated, initial Pengo environment, and then allowing it to explore until it is captured or a threshold of 500 steps is reached. I report my results by training episodes rather than number of training actions because I believe episodes are a more useful measure of "meaningful" training done – an agent having collected all the food and eliminated all the enemies could spend a large amount of time in useless wandering (while receiving no reinforcements). Thus, counting actions might penalize such an agent since it gets to experience fewer reinforcement signals. In any case, for all of my results the results appear qualitatively similar when graphed by the number of training actions (i.e., the agents take a similar number of actions per episode during training).

Each Pengo environment contains a 7x7 grid with approximately 15 obstacles, 3 enemy agents, and 10 food items. I use three randomly generated sequences of initial environments as a basis for the training episodes. I train 10 randomly initialized networks on each of the three sequences of environments; hence, I report the averaged results of 30 neural networks. I estimate the average total reinforcement (the average sum of the reinforcements received by the agent)[1] by "freezing"[2] the network and measuring the average reinforcement on a testset of 100 randomly-generated environments. This same set of 100 environments is used for all of the test results reported in the next section.

I chose parameters for the Q-learning algorithm that are similar to those investigated by Lin (1992). The learning rate for the network is 0.15, with a discount factor of 0.9. To establish a baseline system, I experimented with various numbers of hidden units, settling on 15 since that number resulted in the best average reinforcement for the baseline system. I also experimented with giving the system recurrent units (recall that RATLE uses recurrent units to represent multi-step plan and loop statements), but these units did not lead to improved performance for the baseline system and, hence, my baseline results are for a system without recurrent links. I further experimented by giving the baseline network extra hidden units after initial training (to simulate the effect of adding units to the network when advice is given), but these extra hidden units did not produce gains in performance, so my

---

[1] I report the average total reinforcement rather than the average discounted reinforcement because this is the standard for the RL community. Graphs of the average *discounted* reward are qualitatively similar to those shown in this chapter.

[2] I freeze a network during testing by turning off learning, which means that the weights and biases in the network remain unchanged.

baseline system does not receive extra hidden units during training.

After choosing an initial network topology, I then spent time acting as a teacher in RATLE, observing the behavior of the agent at various times. Based on these observations, I wrote several collections of advice. For use in my experiments, I chose four sets of advice, two that use multi-step plan and loop plans (referred to as *ElimEnemies* and *Surrounded*), and two that do not (*SimpleMoves* and *NonLocalMoves*). Figures 37, 38, 39 and 40 pictorially depict these four sets of instructions. Appendix A shows the advice as RATLE statements, as well

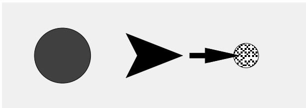


Figure 37: *SimpleMoves* advice for the Pengo agent. This advice suggests the agent move towards any `Food` near to the agent. The advice also suggests the agent move anyway from any `Enemy` near to the agent.

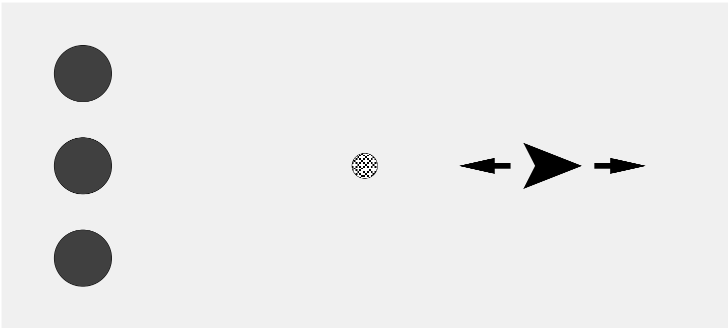


Figure 38: *NonLocalMoves* advice for the Pengo agent. This advice suggests running away if there are many `Enemies` in one direction, but it also suggests moving towards `Food` even if there is an `Enemy` in that direction (as long as the `Enemies` are distant).

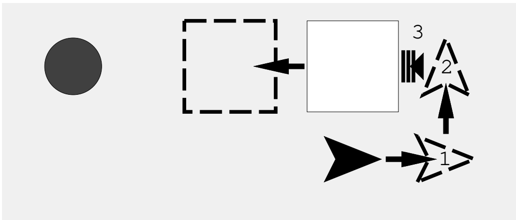


Figure 39: *ElimEnemies* advice for the Pengo agent. This advice gives a three-step plan for eliminating `Enemies` involving moving behind convenient `Obstacles` and pushing them towards the chosen `Enemy`.

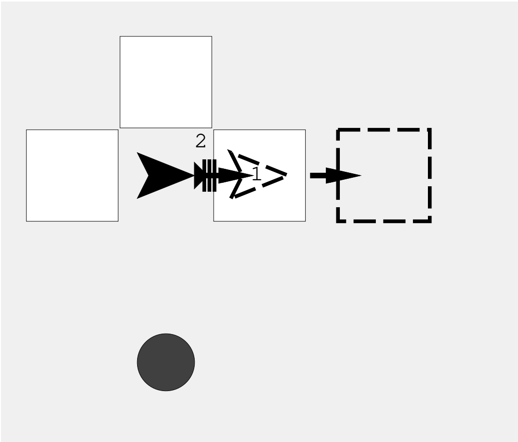


Figure 40: *Surrounded* advice for the Pengo agent. This advice uses a loop and a two-step plan for escaping when the agent is surrounded by `Obstacles` and `Enemies`. The plan involves pushing the nearest `Obstacle` out of the way and moving into the vacated spot until the agent is no longer surrounded.

as the input and fuzzy terms used in these statements.

The advice *SimpleMoves* provides advice about short-term reinforcement situations: when the agent is near to a food item or an enemy. *NonLocalMoves* gives advice about more strategic situations: moving away from groups of enemies and collecting food as long as the food is not too closely guarded. Eliminating enemies is extremely difficult since the agent is unable to see through obstacles to determine when there is an enemy behind the obstacle; the advice *ElimEnemies* provides a multi-step plan to overcome this difficulty. *Surrounded* provides advice in the form of a multi-step plan within a loop that applies to environments where the agent is trapped; the advice indicates how to get out of this trap.

7.1.3 Results

Hypothesis: An agent can make use of advice

In my first experiment, I evaluate the hypothesis that a RATLE agent can in fact make use of advice. After 1000 episodes of initial learning, I judge the value of (independently) providing each of the four sets of advice to my agent via RATLE. I train the system for 2000 more episodes after adding the advice, then measure the average cumulative reinforcement on the testset. (The baseline is also trained for 3000 episodes.) Table 24 reports the averaged testset

Table 24: Testset results for the baseline and the four different types of advice. Each of the four gains over the baseline is statistically significant.

Advice Added	Average Total Reinforcement
None (baseline)	1.32
SimpleMoves	1.91
NonLocalMoves	2.01
ElimEnemies	1.87
Surrounded	1.72

Table 25: Mean number of enemies captured, food collected, and number of actions taken for the experiments summarized in Table 24.

Advice Added	Enemies	Food	Survival Time
None (baseline)	0.15	3.09	32.7
SimpleMoves	0.31	3.74	40.8
NonLocalMoves	0.26	3.95	39.1
ElimEnemies	0.44	3.50	38.3
Surrounded	0.30	3.48	46.2

reinforcement; all gains over the baseline system are statistically significant[3]. Note that the gain is higher for the simpler pieces of advice, *SimpleMoves* and *NonLocalMoves*, which do not incorporate MULTIACTION or loop constructs. This suggests the need for further work on taking complex advice; however the multi-step advice may simply be less useful.

Each of my pieces of advice to the agent addresses specific subtasks: collecting food (*SimpleMoves* and *NonLocalMoves*); eliminating enemies (*ElimEnemies*); and avoiding enemies, thus surviving longer (*SimpleMoves*, *NonLocalMoves*, and *Surrounded*). Hence, it is natural to ask how well each piece of advice meets its intent. Table 25 reports statistics on the components of the reward. These statistics show that the pieces of advice do indeed lead to the expected improvements. For example, the *ElimEnemies* advice leads to a much larger number of enemies eliminated than the baseline or any of the other pieces of advice.

Hypothesis: The effects of advice are independent of when the advice is given

In my second experiment I investigate the hypothesis that the observer can beneficially provide advice at any time during training. To test this, I insert the four sets of advice at

[3] All results reported in this chapter as statistically significant are significant at the $p < 0.05$ level (i.e., with 95% confidence).

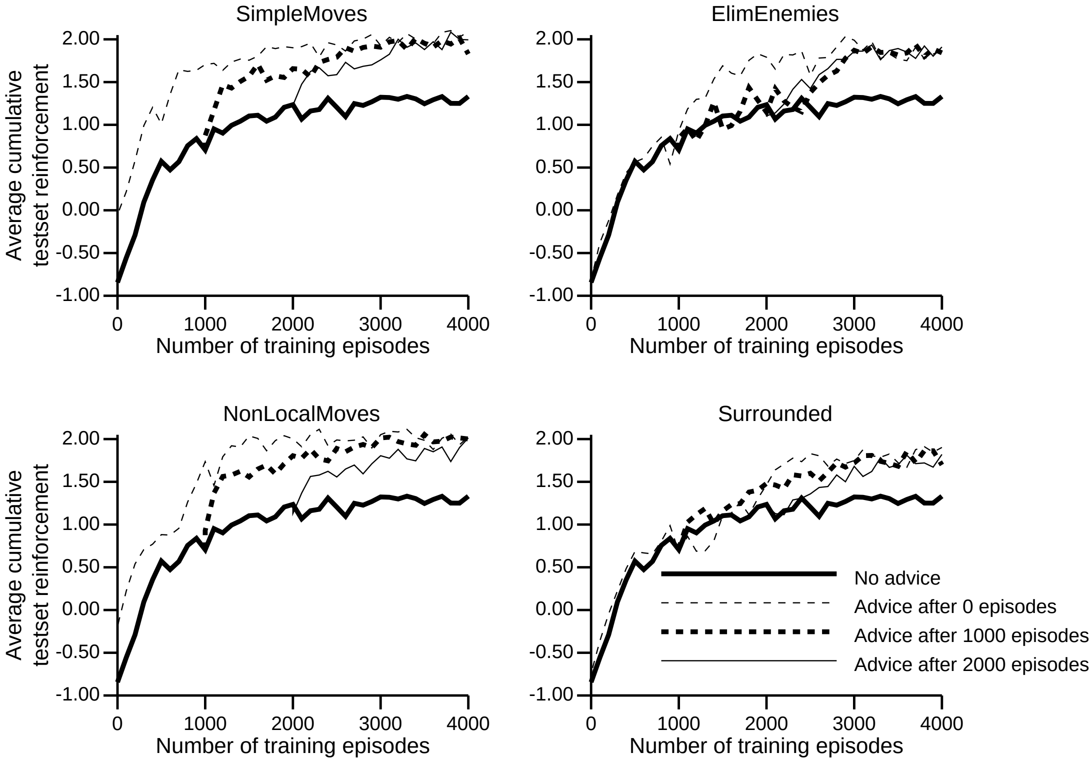


Figure 41: Average total reinforcement for my four sample pieces of advice as a function of amount of training and point of insertion of the advice.

different points in training (after 0, 1000, and 2000 episodes). Figure 41 contains the results for the four pieces of advice. They indicate the learner does indeed converge to approximately the same expected reward no matter when the advice is presented.

Hypothesis: The ability to refine advice is important

In my third experiment, I investigate an approach for using advice where the advice is not refined. This simple strawman algorithm exactly follows the observer's advice when it applies; otherwise it uses a "traditional" connectionist Q-learner to choose its actions. When evaluated on the test set, the strawman employs a loop similar to that shown in Table 23, with the following differences: Step 6 (incorporating advice) is left out; and Step 2 (selecting an action) is replaced by the following:

```
Evaluate the advice to see if it suggests any actions:
   If any actions are suggested, choose one of these randomly,
   Else choose the maximum utility action from the connectionist Q-network.
```

Table 26: Average testset reinforcement using the strawman approach of literally following advice compared to the RATLE method of refining advice based on subsequent experience.

Advice	STRAWMAN	RATLE
SimpleMoves	1.63	1.91
NonLocalMoves	1.46	2.01
ElimEnemies	1.28	1.87
Surrounded	1.21	1.72

The performance of this strawman is reported in Table 26. In all cases, RATLE performs statistically significantly better than the strawman on cumulative reinforcement received. In fact, in two of the cases, *ElimEnemies* and *Surrounded*, the resulting method for selecting actions is actually worse than simply using the baseline network (whose average performance is 1.32).

Hypothesis: An agent can profitably make use of subsequent advice

In my fourth experiment, I investigate the hypothesis that subsequent advice (i.e., advice provided after the agent has already incorporated and refined an initial piece of advice) will lead to further gains in performance. To test this hypothesis, I supplied each of my four pieces of advice to an agent after 1000 episodes (as in my first experiment), supplied one of the remaining three pieces of advice after another 1000 episodes, and then trained the resulting agent for 2000 more episodes. These results are averaged over 60 neural networks instead of the 30 networks used in the other experiments in order to get statistically significant results. Table 27 shows the results of this experiment.

In all cases, adding a second piece of advice leads to improved performance. However, the resulting gains when adding the second piece of advice are not as large as the original gains over the baseline system. I suspect this occurs due to a combination of factors: (1) there is an upper limit to how well the agents can do – though it is difficult to estimate this upper bound; (2) the pieces of advice interact – they may suggest different actions in different situations, and in the process of resolving these conflicts, the agent may use one piece of advice less often than it would if there was only one piece of advice; and (3) the advice pieces are related, so that one piece may cover situations that the other already covers. Also interesting to note is that the order of presentation affects the level of performance achieved in some cases (e.g., presenting *NonLocalMoves* followed by *SimpleMoves* achieves higher performance than *SimpleMoves* followed by *NonLocalMoves*).

Table 27: Average testset reinforcement for each of the possible pairs of my four sets of advice. The first piece of advice is added after 1000 episodes, the second piece of advice after an additional 1000 episodes, and then trained for 2000 more episodes (for total of 4000 training episodes). Shown in parentheses next to the first pieces of advice are the performance results from my first experiment where only a single piece of advice was added. All of the resulting agents show statistically significant gains in performance over the agent receives just the first piece of advice.

	Second Piece of Advice			
First Piece of Advice	*SimpleMoves*	*NonLocalMoves*	*ElimEnemies*	*Surrounded*
SimpleMoves (1.91)	-	2.17	2.10	2.05
NonLocalMoves (2.01)	2.27	-	2.18	2.13
ElimEnemies (1.87)	2.01	2.26	-	2.06
Surrounded (1.72)	2.04	2.11	1.95	-

Hypothesis: RATLE agents can overcome the effects of "bad" advice

In my fifth experiment, I investigate the effect that "bad" advice has on a RATLE agent. To form the advice used in these experiments I altered the advice *SimpleMoves*. I produced one set of bad advice, *NotSimpleMoves*, that is similar to *SimpleMoves*, except that whenever *SimpleMoves* suggests an action, *NotSimpleMoves* suggests that the agent should *not* take that action (using the DO_NOT construct). My second set of bad advice, *OppositeSimpleMoves*, is also similar to *SimpleMoves*, except that whenever *SimpleMoves* suggests an action, *OppositeSimpleMoves* suggests that the action in the opposite compass direction (e.g., when *SimpleMoves* suggests the agent should `MoveEast`, *OppositeSimpleMoves* suggest the agent should `MoveWest`). The actual constructs used to define all three pieces of advice appear in Appendix A.

I tested these two pieces of "bad" advice by separately adding each after 1000 training episodes and then training the agent for 3000 more training episodes. Figure 42 presents the results of these experiments. These results demonstrate that while bad advice does in fact cause an initial immediate drop in performance, the agent is quickly able to refine the advice to return to reasonable behavior. It is also interesting to note that, over time, the agent may actually produce gains in performance from the "bad" advice (compared to learning without advice). After 4000 training episodes, the advice *OppositeSimpleMoves* produces a statistically significant gain in performance over the agents with no advice, but produces significantly worse performance than the original *SimpleMoves* advice. The advice *NotSimpleMoves* does not produce a significant gain over not receiving advice after 4000 training

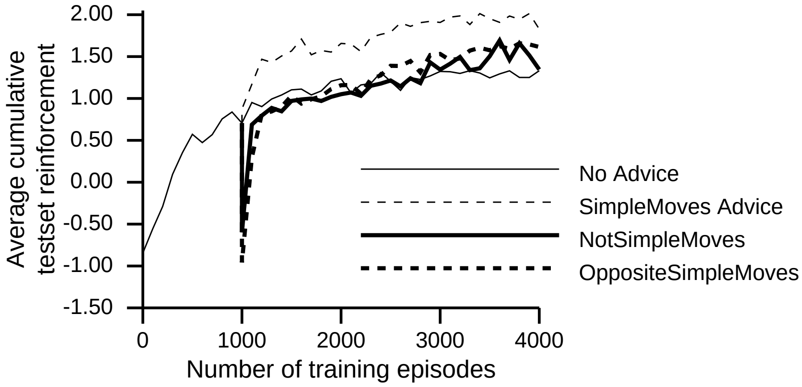


Figure 42: Average total reinforcement for the baseline as well as agents with the advice *SimpleMoves*, *NotSimpleMoves*, and *OppositeSimpleMoves*.

episodes. The improvements over the no advice case are likely due to the fact that, while I changed the predicted actions of *SimpleMoves* to produce the bad advice, the intermediate terms constructed in forming the bad advice (i.e., the hidden units that trigger the actions) are the same in all three sets of advice; the agent can take advantage of these intermediate inferences to learn the correct actions to take.

Hypothesis: Replay will improve the performance of RATLE's agents

In my final experiment, I evaluate the usefulness of combining my advice-giving approach with Lin's "replay" technique (1992). Lin introduced the replay method as a means to make use of "good" sequences of actions provided by a teacher. In replay, the agent trains on the teacher-provided sequences frequently to bias its utility function towards these good actions. Thrun (1994) reports that replay can be used even when the remembered sequences are not teacher-provided sequences – in effect, by training multiple times on each state-action pair the agent is "leveraging" more value out of each example. Hence, my experiment addressed two related questions: (1) does the advice provide any benefit over simply reusing the agent's experiences multiple times?; and (2) could my approach benefit from replay, for example, by needing fewer training episodes to achieve a given level of performance? My hypothesis was that the answer to both questions is "yes."

To test my hypothesis, I implemented two approaches to replay in RATLE and evaluated them using the *NonLocalMoves* advice. In one approach, which I will call *Action Replay*, I simply keep the last N state-action pairs that the agent encountered, and on each step train with all of the saved state-action pairs in a random order. A second approach (similar

Table 28: Average total reinforcement results for the advice *NonLocalMoves* using two forms of replay. The advice is inserted and then the network is trained for 1000 episodes. Replay results are the *best* results achieved on 1000 episodes of training (at 600 episodes for Action Replay and 500 episodes for Sequence Replay). Results for the RATLE approach without replay are also shown; these results are for 1000 training episodes.

Training Method	Average Total Reinforcement
Standard RATLE (no replay)	1.74
Action Replay Method	1.48
Sequence Replay Method	1.45

to Lin's), which I will call *Sequence Replay*, is more complicated. Here, I keep the last N *sequences* of actions that ended when the agent received a non-zero reinforcement. Once a sequence completes, I train on all of the saved sequences, again in a random order. To train the network with a sequence, I first train the network on the state where the reinforcement was received, then the state one step before that state, then two steps before that state, etc., on the theory that the utility of states nearest reinforcements are best estimated by the network. Results for keeping 250 state-action pairs and 250 sequences[4] appear in Table 28. Due to time constraints, I trained these agents for only 1000 episodes. Note that for replay this represents much more training than the standard system receives, since after each actions or episodes the agent is trained with multiple actions or multiple episodes. For example, with the Action Replay approach, each state-action pair is used in training 250 times, so there is 250 times as much network training for this approach.

Surprisingly, replay did not help in this testbed. After examining networks during replay training, I hypothesize this occurred because I am using a single network to predict all of the Q values for a state. During training, to determine a target vector for the network, I first calculate the new Q value for the action the agent actually performed. I then activate the network, and set the target output vector for the network to be equal to the actual output vector, except that I use the new prediction for the action taken. For example, assume the agent takes action two (of three actions) and calculates that the Q value for action two should be 0.7. To create a target vector, the agent activates the network on the current state (assume that the resulting output vector is [0.4,0.5,0.3]), and then creates a target vector that is the same as the output vector except for the new Q value for the action taken

[4] I also experimented with keeping only 100 pairs or sequences; the results using 250 pairs and sequences were superior.

(i.e., [0.4,**0.7**,0.3]). This causes the network to have error at only one output unit (the one associated with the action taken). For replay this is a problem because I will be activating the network for a state a number of times, but only trying to correctly predict one output unit (the other outputs are essentially allowed to take on any value), and since the output units share hidden units, changes made to predict one output unit may affect others. If I repeat this training a number of times, the Q values for other actions in a state may become greatly distorted. One possible solution to this problem is to use separate networks for each action, but this means the actions will not be able to share concepts learned at the hidden units. I plan to further investigate this topic, since replay intuitively seems to be a valuable technique for reducing the amount of experimentation on the external world that an RL agent has to perform.

7.1.4 Discussion of the Pengo Experiments

My experiments demonstrate that: (1) advice can in fact improve the performance of an agent, (2) advice produces approximately the same resulting performance no matter when it is added, (3) it is important for the agent to be able to refine advice based on subsequent experience, (4) a second piece of advice can produce further gains, (5) "bad" advice can produce an initial drop in performance that the agent can overcome with learning, and (6) a straightforward approach to replay does not produce significant benefits. One key question that might arise from the results presented is why the baseline agent does not eventually achieve the same results that the system with advice achieves? To answer this question, I will address a more general question – what effect do I expect advice to have on the agent?

A connectionist Q-learner uses gradient-based learning to select a function that minimizes the error between the predicted and estimated Q values:

$$Error \;=\; \sum_{e \in Examples} \left[\tilde{Q}(s_e, a_e) \;-\; \hat{Q}(s_e, a_e)\right]^2 \qquad (13)$$

where $\hat{Q}$ is the estimate obtained by actually taking an action and looking at the predicted utility for all of the actions in the resulting state, and $\tilde{Q}$ is the current Q value predicted by the learner's Q function, and s_e and a_e are the state and action chosen in example e. One problem with this approach is that the learner may select a minimum through gradient-based learning that is "good," but sub-optimal with respect to the optimal Q function (represented by a Q table). For example, assume that for the vast majority of the states and actions that the reward is zero, and that, as a consequence, most of the Q values are zero. In this case it may be advantageous to the network to always predict zero, since the total error for predicting zero all the time would be relatively small. One hopes that the network will be able to find features of those states that have non-zero Q values that distinguish those states, but if these states are very similar to other states with zero Q values this may not occur. Clearly, it is possible for a connectionist Q learner to settle on a Q function that is less than optimal.

When I introduce "good" advice into an agent, I expect it to have one or more of several possible effects. One possible effect is that the advice will change the network's predictions of some of the Q values to ones that are closer to the desired "optimal" values. By reducing the overall error the agent may be able to converge more quickly towards the optimal Q function. A second related effect is that by increasing (or decreasing) certain Q values the advice changes which states are explored by the agent. In this case, good advice will cause the agent to explore states that are useful in finding the optimal plan (or ignoring states that are detrimental). Focusing on the states that are important to the optimal solution may lead to the agent converging more quickly. A third possible effect is that the addition of advice alters the weight space of possible solutions that the learner is exploring. This occurs because the new weights and hidden units locally change the space that the learner is in. For example, the advice may construct an intermediate term (represented by a hidden unit) with very large weights that could not have been found by gradient-based search. In the resulting altered weight space, the learner may be able to explore functions that were unreachable before the advice is added (and these functions may be closer to the optimal).

Given these possible effects of good advice, I conclude that advice can both cause the agent to converge more quickly to a solution, and that advice may cause the agent to find a better solution than it may have otherwise found. For my experiments, I see the effect of speeded convergence in the graphs of Figure 41, where the advice, generally after a small amount of training, leads the agent to achieve a high level of performance quickly. These graphs also demonstrate the effect of convergence to a better solution. Note that these effects are a result of what I would call "good" advice. It is possible that "bad" advice could have equally deleterious effects. So, a related question is how do I determine whether advice is "good" or not?

Unfortunately, determining the "goodness" of advice appears to be a fairly tricky problem, since even apparently useful advice can lead to poor performance in certain cases. Consider, for example, the simple problem shown in Figure 43. This example demonstrates

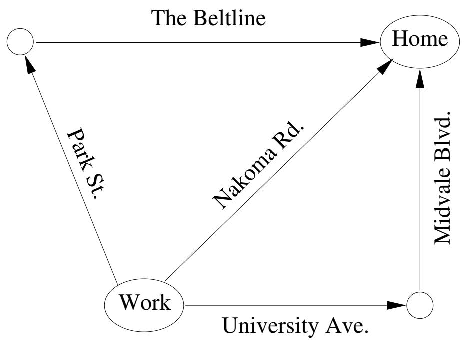


Figure 43: A sample problem where advice can fail to enhance performance. Assume the goal of the agent is to go from Work to Home, and that the agent receives a large reward for taking Nakoma Road, and a medium-sized reward for taking University Avenue followed by Midvale Boulevard. If the agent receives advice that University followed by Midvale is a good plan, the agent, when confronted with the problem of going from Work to Home will likely follow this plan (even during training, since actions are selected proportional to their likelihood of achieving reward). Thus it may take a long time for the learner to try Nakoma Road often enough to learn it is a better route, since every time the learner follows the advice it will not learn much. A learner without advice might try both routes equally often and quickly learn the best route.

a case where advice, though providing useful information, could in fact cause the agent to take longer to converge to an optimal policy. This example suggests that I may want to re-think the stochastic mechanism I use to select actions, but other mechanisms seem to suffer from similar problems. In any case, it thus appears that defining the properties of "good" advice is a topic for future work. As a first (and admittedly vague) approximation, I would expect advice to be "good" when it causes one of the effects mentioned above: (1) it reduces the overall error in the agent's predicted Q values, (2) it causes the agent to pick actions that lead to states that are important in finding a solution, or (3) it transforms the network's weight space so that the agent is able to perform gradient-based learning to find a better solution.

7.2 Experiments on the Soccer Testbed

I chose to experiment on a Soccer testbed to confirm the results I obtained in the Pengo testbed. In designing the Soccer testbed, I deliberately set out to build a task that was significantly different from the Pengo environment. Below I discuss the differences between these tasks. Using this testbed, I will show that an agent receiving advice outperforms an agent that does not receive advice. I also discuss why the agent refining the advice outperforms an agent that does not refine the advice.

One major difference between the Pengo and Soccer testbeds is that the agent in the Soccer task receives complete information; no objects are occluded in the Soccer environment. The agent's sensors are also different. The agent in Pengo received information about the occupancy of various sectors around the agent; the agent in Soccer receives a set of measures indicating the distance and direction to all of the other objects in the environment. The tasks also differ in that Soccer has cooperating agents in addition to competing agents. Finally, the reinforcement signals are more sparse in the Soccer environment; the agent receives a positive signal when it scores and a negative signal when the other agent's score. These signals are made more difficult to interpret by the agent because there is more than one agent per team and the positive reinforcement for a goal is received by all of the agents on the scoring team (similarly, the negative reinforcement for allowing a goal is received by all the players on a team).

7.2.1 The Soccer Environment

Figure 44 shows the basic Soccer environment. Each player can perform eleven actions: the agent may *move* in each of the directions Forward, Back, Left, and Right; the agent may

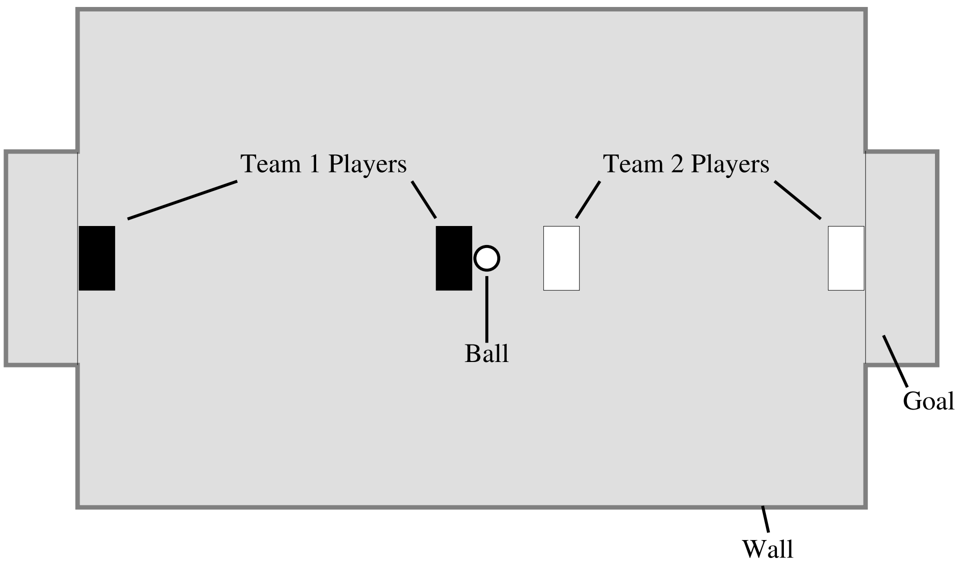


Figure 44: The Soccer test environment. Each player wants to kick the ball into the opponent's goal while preventing the ball from going into their own goal.

kick the ball in each of the directions Forward, Back, Left, Right, ForwardLeft (diagonally) and ForwardRight; and the agent may *do nothing*. Moves cause the agent to change position on the field, while kicking actions cause the ball to leave the agent (who stays still) and roll on the field. All of the moves (both by players and by the ball) take place simultaneously. Moves that would cause two players to collide are resolved randomly (only one player may occupy a location on the field at the time). I assume that the Soccer game is played in an indoor (walled) arena, so that the agent is not able to move off the field of play, and that a ball kicked against the wall ricochets off. A kicking action affects the ball only when the agent actually controls the ball. An agent controls the ball when the agent moves in the path of the moving ball or when the agent moves into the location where the ball currently rests. Should two players collide while one has the ball, the other player gains control of the ball half of the time. A score occurs when the ball rolls into one of the goals.

The players receive a positive reinforcement (1.0) when a goal for their team is scored and a negative reinforcement (-1.0) when a goal is scored against them. I do not differentiate goal-scoring situations; a player receives a positive reinforcement when a goal is scored for its team even if the player had no part in scoring the goal (either because some other player on its team scored, or because some player on the other team mistakenly kicked the ball in their own goal).

In the experiments I performed, there are two agents on each team. In order to reduce the complexity of the task, I focused on learning in situations which I call *breakaways* (see Figure 45). A breakaway occurs when one team has control of the ball and only one of the players on the opposing team is in position to defend their goal. I will refer to the team that initially has control of the ball as the *attacking team* and the other team as the *defending team*. The breakaway is defined to last for 20 actions or until either team scores. I further limited the complexity of the task by only allowing the attacking team to learn. The attacking team uses reinforcement learning, while the defending team follows a hand-crafted policy I developed. In my hand-crafted policy, the player that is initially in position to defend acts as a goalie, keeping itself between the ball and the goal. The other player rushes back until it too can act like a goalie.

Recall that in the Soccer testbed, unlike the Pengo testbed, each agent has complete information about the environment. Each agent receives information describing the distance and direction to the following objects in the environment: the ball; the players on the opposing team; the other player on the same team; and the left post, right post, and center of each goal. Each agent also has input features representing the distance to each wall,

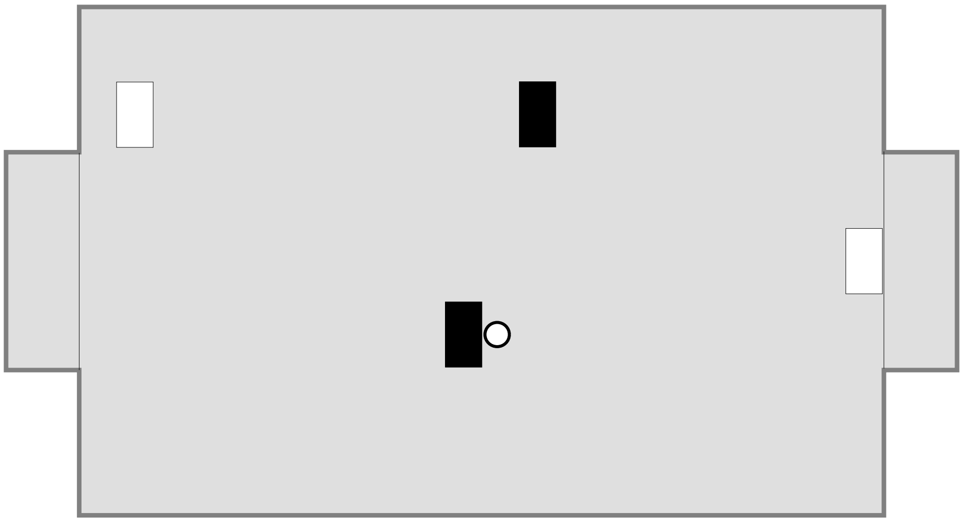


Figure 45: An example of a *breakaway* Soccer environment. The attacking team (black) has control of the ball, and only one of the defenders is between the attackers and the defenders goal.

whether or not that player has the ball and a counter indicating how much time is left for the team to score. I simultaneously represent each distance in two ways: as the x and y differences between the agent's current position and the object observed; and as the distance, plus the *sine* and *cosine* of the angle to the object. I include both representations because different values are useful in different situations, and the agents do no currently have the ability to calculate one representation from the other. Appendix B describes the input and fuzzy terms used in these experiments.

7.2.2 Methodology

I trained the agents for 400,000 games. Each game starts with an initial randomly generated breakaway situation, and continues until either team scores or until 20 actions have been taken. Similar to the Pengo experiments, I report results by the number of breakaway situations rather than the number of training actions taken; the results are qualitatively similar when graphed by number of actions.

Each breakaway environment is an 11x7 grid with two players on each team. Initially, one of the attacking team players has the ball. In a breakaway environment, I distribute the attacking team players randomly in the center five columns of the field, and the defending team players are split, one randomly in the three far-left columns of the field, and one randomly in the three far-right columns.

I generated 40 sequences of random breakaway situations and trained one attacking team for each sequence; my results are averaged over 40 attacking teams. To test the teams, I freeze the attacking team's neural networks and count the number of times the different teams win for a sequence of 100 randomly generated breakaway environments. The identical environments are used for all tests. I report the net number of goals for the attacking team (i.e. the number of games the attacking team scored minus the number of games where the defending team scored).

I use the same parameters for connectionist Q-learning that I used in the Pengo experiments. The learning rate for the network is 0.15, with a discount factor of 0.9. My baseline network (chosen empirically) has five hidden units for each action (for a total of 55 hidden units), and I connect each hidden unit to all of the inputs. I chose to use a policy network that has a separate sub-network for each action because tests with a completely connected

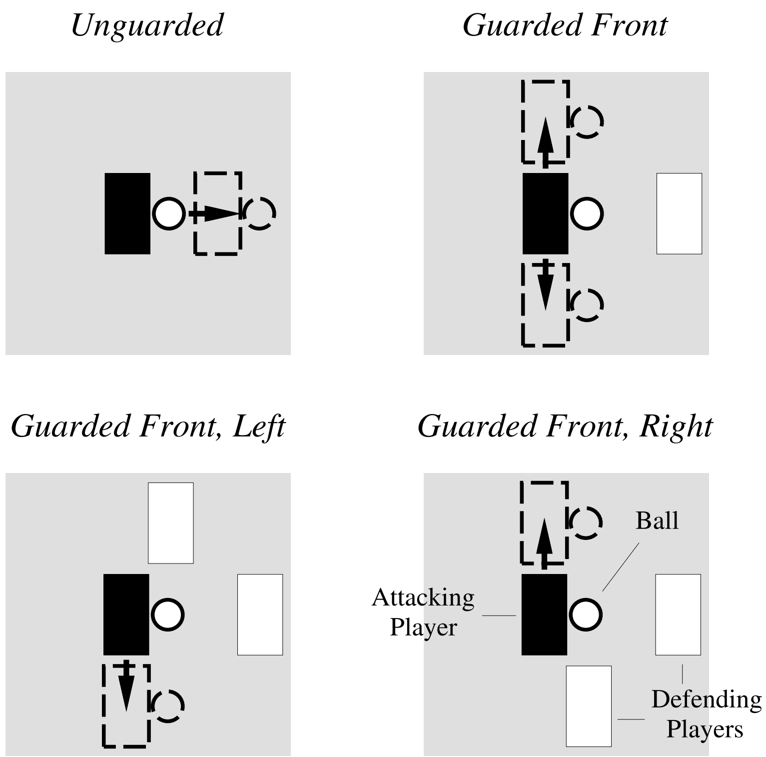


Figure 46: Advice for soccer players. The advice characterizes situations where the player has the ball according to how well guarded the player is; it suggests a move for the various situations.

network produced very poor performance, apparently for the reasons I mention in the discussion in Section 7.1.4 about the value of advice – the reinforcements signals are so sparse in this environment that the network often settles on a solution that simply predicts zero future reward. Separating the actions ameliorates this effect.

The advice I gave to the attacking team is shown pictorially in Figure 46. I give the same advice to both players on the attacking team. The advice suggests moves that a player should make when they have the ball and depending on how many opposing players are guarding the player. Appendix B includes the actual statements used to give the advice.

7.2.3 Results and Discussion

Figure 47 shows the average net testset games won by the team using the advice pictured in Figure 46 and the net test games won by a baseline team (without advice). Note that both teams play against my hand-crafted defending team (described above). The resulting gains in performance for the team using advice over the team without advice are significant for all of the results after 90,000 training games. These results show that agents can profitably

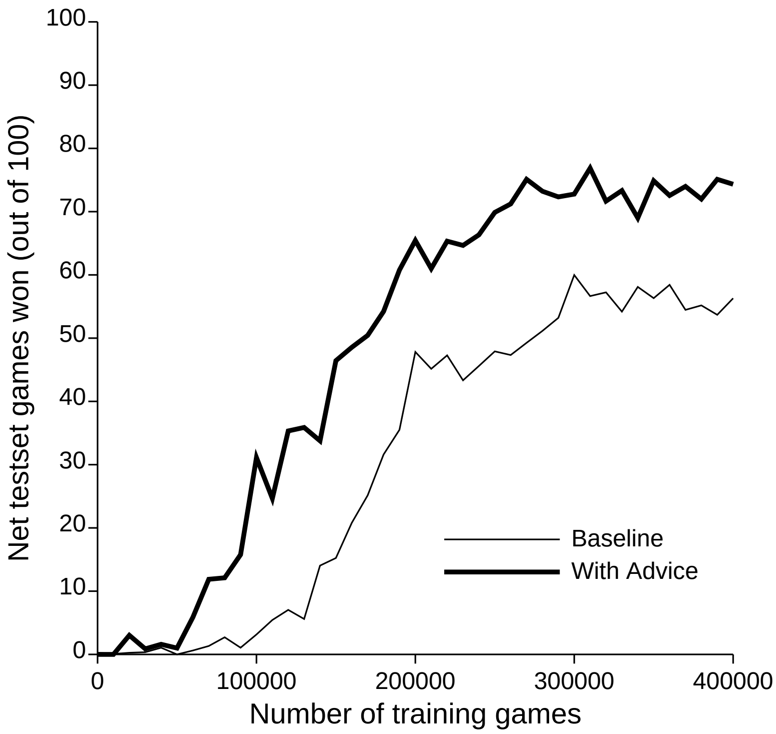


Figure 47: Net testset games won, as a function of training games, by an RL team with advice and a baseline RL team without advice.

make use of advice in an environment where multiple agents learn.

I also performed tests to determine how a team of agents that are not able to refine the advice performs. As with the strawman I constructed in the Pengo task, I constructed agents that follows the advice when any is applicable, and otherwise choose the highest utility action according to the networks trained for the baseline above. In experiments with this strawman, a team of agents following this approach *never* scored. In my advice I suggested that a player move forward whenever it has the ball and it is unguarded. But this prevents the player from shooting when they are near the goal (and the player cannot carry the ball into the goal). Thus, this strawman, when applying the advice, will never shoot in situations where it is appropriate to do so. This result demonstrates the need to refine a teacher's advice when that advice is not precisely correct.

7.3 Summary

My experiments with the Pengo and Soccer testbeds show that my system, RATLE, which allows a reinforcement-learning agent to refine procedural domain theories. An agent in RATLE, when receiving good advice, achieves performance that is superior to the performance that an agent achieves without the advice. My experiments also demonstrate that an agent that is able to refine the advice it receives is able to outperform an agent that receives the same advice but does not refine it. Thus, my experiments support the central hypothesis of this thesis.

I also investigated other aspects of the RATLE system. Analyses of the effects of advice show that advice seems to produce effects that are consistent with the focus of the advice (e.g., advice designed to eliminate enemies produces agents that eliminate more enemies). In the Pengo testbed, I performed experiments that indicate that the gains in performance due to advice occur no matter when the agent receives advice. I also show that an agent is able to profitably make use of a second piece of advice, which demonstrates that the interaction of the teacher and agent in RATLE need not stop after a single piece of advice. I also demonstrated that an agent supplied with "bad" advice may see an initial drop in performance, but that with learning the agent can overcome the effects of "bad" advice. Experiments with the technique *replay*, a method for reusing the experience that the agent obtains during exploration, indicate that a straightforward approach to this technique does not produce the gains in performance reported by others in their testbeds (Lin, 1992; Thrun, 1994).

My tests with the Soccer task support my results from the Pengo task. The Soccer task is interesting because it is different from the Pengo task in a number of ways: the agents in Soccer have complete world information, where the Pengo agent has only limited line-of-sight information; the agents in Soccer work cooperatively against a set of competitive agents, while the Pengo agent works alone against a set of competitive agents; and the set of reinforcements in the Soccer environment (goals scored) are much more sparse than those in the Pengo environment. Despite these differences, the results for the Soccer task are similar to those from the Pengo task; the agents receiving advice in the Soccer task also show gains in performance over agents that do not receive instruction. These results suggest that the RATLE system is applicable to a wide range of reinforcement-learning tasks.

Chapter 8

Additional Related Work

My work relates to research in a number of areas, some of which I presented in Chapter 2. In this chapter, I describe other research that is closely related to my own. As I present other research, I will discuss their differences with respect to RATLE, since it encompasses much of FSKBANN (related work for the protein-folding problem was discussed in Chapter 3).

One related research area involves methods for giving advice to a problem solver – techniques that use knowledge provided by a teacher. I first discuss these systems, and then discuss ones specifically designed to give advice to a reinforcement learner.

Research on refining domain theories relates both to advice-giving research in general and to my work in particular. These techniques take advice in the form of an initial theory about a task and refine it using samples of the task. An especially relevant theory-refinement approach is to create knowledge-based neural networks, since my work augments methods from this field. I first present a number of alternatives for incorporating advice into a neural network and discuss how these approaches relate to mine. I then give an overview of some work using inductive-logic programming (Muggleton, 1992) to refine a domain theory.

Next, I present research on inducing finite-state automata. Early work in this field focused on theoretical limits for learning algorithms and methods using constructs such as oracles (i.e., a device that can always produce an answer to a query if an answer exists). More recently, techniques such as recurrent neural networks and knowledge-based neural networks have been applied to this problem, and I describe the relation of these approaches to mine.

Finally, I discuss some work on developing programming languages for interacting with agents, and discuss how RATLE differs from these efforts.

8.1 Providing Advice to a Problem Solver

As I previously noted, the concept of giving advice has a long history in AI. Many advice-taking systems focus on the problems of understanding teacher instructions, including how to operationalize them into a form that the computer learner can use. Generally these approaches focus on understanding rather than refining the advice provided by the teacher.

One early example of an algorithm that makes use of advice for planning is ABSTRIPS (Sacerdoti, 1974). In ABSTRIPS, a human assigns initial "criticalities" to preconditions to cause the planner to focus on making key planning steps. In both ABSTRIPS and RATLE a human is in some way directing the search; however, while the ABSTRIPS mechanism requires that the teacher have some understanding of the search process, RATLE does not need the teacher to understand how the agent works internally. Also, RATLE agents are able to learn.

In FOO (Mostow, 1982), general advice is *operationalized* by reformulating the advice into search heuristics. FOO applies these search heuristics during problem solving. In FOO, Mostow assumes that the advice is correct, and the learner's task is to convert the general advice into an executable plan based on its knowledge about the domain. RATLE is different from FOO in that RATLE tries to directly incorporate general advice, but does not provide a sophisticated means of operationalizing advice – RATLE only operationalizes fuzzy terms like "big" and "near." Also, RATLE does not assume that the advice is correct; instead, it uses reinforcement learning to refine and evaluate the advice.

More recently, Laird et al. (1990) created an advice-taking technique called ROBO-SOAR that is based on the SOAR architecture (Laird et al., 1987). SOAR uses its knowledge to select operators to achieve goals. When it is unable to select from a set of operators, or no operator is applicable, an *impasse* arises. When this happens, SOAR creates a subgoal and applies itself to that subgoal. In ROBO-SOAR, an observer can provide advice to the system during an impasse by suggesting which operators to explore in an attempt to resolve the impasse. As with FOO and ABSTRIPS, ROBO-SOAR uses advice to guide the learner's reasoning process, while RATLE directly incorporates the advice into the learner's knowledge and then refines that knowledge with subsequent experience.

Huffman and Laird (1993) developed another SOAR-based method called INSTRUCTO-SOAR that allows an agent to interpret simple imperative statements such as "Pick up the red block." INSTRUCTO-SOAR examines these instructions in the context of its current task, and uses SOAR's form of explanation-based learning (Mitchell et al., 1986) to generalize the instruction into a rule that it uses in similar situations. RATLE differs from INSTRUCTO-SOAR in that it provides a language for providing general advice rather than attempting to generalize specific advice.

8.2 Providing Advice to a Problem Solver that Uses Reinforcement Learning

Recently, several researchers have defined methods for providing advice to computer learners and, in particular, reinforcement-learning agents. A number of these approaches are closely related to mine.

Lin

Lin (1991, 1992, 1993) uses RL on a simulation task that is similar to my Pengo task. Both of our agents employ connectionist Q-learning and explore a robot world using a set of sensors providing limited information. Lin (1992) designed the technique "replay" that uses advice expressed as sequences of teacher's actions. In his approach, the agent periodically trains on the sequence of teacher-provided actions. Thus, the resulting agent should be biased towards the actions chosen by the teacher. One difference between my approach and Lin's is that RATLE accepts advice in a general form. Also, RATLE directly installs advice into the RL agent; in Lin's system, the agent must spend time training its policy function on the teacher's actions to try to determine the conditions that lead to these actions.

Lin (1993) has also investigated the idea of having a learner use prior state knowledge. He uses an RL agent whose input contains not only the current input description, but also some number of the previous input descriptions. The difference between Lin's approach and mine is that his agents retain entire input descriptions, while in RATLE the teacher gives advice that suggests which information to retain. The advice thus limits the amount of information the agent has to process and greatly focuses the learning process.

Clouse and Utgoff

Utgoff and Clouse (1991) developed a learner that consults a set of teacher actions when the action the learner chose resulted in significant error. Their system has the advantage that the agent determines the situations in which it requires advice, but is limited in that it may require advice more often than the observer is willing to provide it. In RATLE the advisor provides advice whenever she feels she has something to say.

Clouse and Utgoff (1992) created a second technique that takes advice in the form of actions suggested by the teacher. Whenever the teacher feels it is appropriate, she can suggest an action that the agent should take. Their approach is therefore similar to RATLE in that the teacher determines when to provide advice. The main difference is that in Clouse and Utgoff's approach the advice is specific to a single situation (i.e., the action the teacher suggests for a state), whereas the advice in RATLE is broadly applicable. In Clouse and Utgoff's approach, the agent must employ empirical learning to generalize the advice given.

Whitehead

Whitehead (1991) examined an approach similar to both Lin's and Utgoff & Clouse's. In his approach the agent can learn by receiving advice in the form of *critiques*, which are rewards indicating whether the agent's current action is optimal or not. The agent can also learn by observing the actions of the teacher. As with the above approaches, RATLE differs from Whitehead's approach in that RATLE is able to accept general advice about actions to take in multiple states.

Gordon and Subramanian

Gordon and Subramanian (1994) developed a system that is closely related to mine. Their system employs genetic algorithms (Holland, 1975), an alternate approach to learning from reinforcements. Their agent accepts high-level advice of the form IF *conditions* THEN ACHIEVE *goal*. It operationalizes these rules using its background knowledge about goal achievement. Their agent then uses genetic search to explore its environment and alter its initial background knowledge.

My work primarily differs from Gordon and Subramanian's in that RATLE uses connectionist Q-learning instead of genetic algorithms. Also, my advice language focuses on actions to take rather than goals to achieve. Finally, RATLE allows advice to be given at any time during the training process. One drawback, however, is that RATLE does not have the operationalization capability of Gordon and Subramanian's approach.

Thrun and Mitchell

Thrun and Mitchell (1993) investigated a method that allows RL agents to make use of prior knowledge in the form of neural networks. In their approach they assume the existence of neural networks that have been previously trained on similar tasks. They focus on using the knowledge already embedded in these networks to give their learner an initial theory for the new task. This proves to be effective, but requires previously-trained neural networks that are related to the task being addressed. RATLE takes advice in a more general form and does not require previous training.

8.3 Refining Prior Domain Theories

There has been a growing literature on automated theory refinement (Cohen, 1994; Fu, 1989; Ginsberg, 1988; Ourston & Mooney, 1994; Pazzani & Kibler, 1992; Towell & Shavlik, 1994). Theory refinement is closely related to the idea of advice taking, since the domain theory provided to the learner can be thought of as advice to the learner. Theory refinement differs from most advice-taking approaches in that theory-refinement systems generally focus on the process of refining the advice, rather than on understanding advice.

My work differs from standard theory-refinement approaches by its novel emphasis on refinement of procedural domain theories in multi-actor worlds, as opposed to refinement of theories for categorization and diagnosis. Finally, unlike previous approaches, I allow domain theories to be provided at any time during the training process, as the need becomes apparent to the teacher. In complex tasks where learning is slow, it is not desirable to simply restart learning from the beginning whenever one wants to add something to the domain theory.

8.3.1 Incorporating Advice into Neural Networks

One popular method for refining domain theories is to refine them with a neural network. KBANN, the system RATLE and FSKBANN are built on, is an example of a system for refining domain theories with neural networks. I presented details on KBANN in Chapter 2. Below I discuss a number of other techniques for using advice in neural networks and how these techniques relate to my work.

One approach closely related to my own is Siegelman's (1994) technique for converting programs expressed in a general-purpose, high-level language into a type of recurrent neural network. Her system is especially interesting in that it provides a mechanism for performing arithmetic calculations. She also provides mechanisms for looping constructs and other statements that are very similar to the constructs of RATLE. Siegelman's approach differs from mine in that she uses network units with piecewise linear activation functions. Also, her work has not been empirically demonstrated, due to her reliance on these non-standard network units.

Gruau (1994) developed a compiler that translates Pascal programs into neural networks. While his approach has so far only been tested on simple programs, it may prove applicable to the task of programming agents. Gruau's approach includes two methods for refining the networks he produces: (1) a genetic algorithm, and (2) a hill-climber. The main difference between Gruau's technique and RATLE is that the networks RATLE produces can be refined using standard connectionist techniques such as backpropagation. Also, Gruau's networks require the development of a specific learning algorithm, since they require integer weights (-1,0,1) and incorporate functions that do not have derivatives.

Noelle and Cottrell (1994) suggest a novel approach for making use of advice in neural networks. In their approach, the connectionist model itself performs the process of incorporating advice, where advice is expressed as inputs to the network, and the "knowledge" of the network is maintained in a set of memory units employing recurrent links. This contrasts with RATLE's approach of directly adding new "knowledge-based" units to the neural network. RATLE leads to faster assimilation of advice, though Noelle and Cottrell's approach is arguably a better psychological model.

Diederich (1989) devised a method that accepts instructions in a symbolic form. He uses instructions to create examples and then incorporates the instructions by training a neural network with these examples. RATLE differs from Diederich's approach in that it directly installs instructions into the neural network.

Abu-Mostafa (1995) uses an approach similar to Diederich's to encode "hints" in a neural network. A hint is a piece of knowledge provided to the network that indicates some important general aspect for the network to have. For example, a hint might indicate to a network trying to assess people as credit risks that a "monotonicity" principle should hold (e.g., when one person is a good credit risk, then an identical person with a higher salary should also be a good risk). Abu-Mostafa uses these hints to generate examples that will cause the network to have this property, then mixes these examples in with the original training examples. As with Diederich's work, my work differs from Abu-Mostafa's in that I directly install the advice into the network.

Suddarth and Holden (1991) investigated another form of "hint" for a neural network. In their approach, a hint is an extra output value for the neural network. For example, a neural network using sigmoidal activation units to try to learn the difficult XOR function might receive a hint in the form of the output value for the OR function. The OR function is useful as a hint because it is simple to learn. The network can use the hidden units it constructs to predict the OR value to construct a simple function for the XOR (i.e., the hint serves to decompose the problem for the network). Suddarth and Holden's work however only deals with hints in the form of useful output signals, and still requires network learning, while my approach incorporates advice immediately.

8.3.2 Refining Domain Theories via Inductive Logic Programming

In inductive-logic programming (ILP) (Muggleton & Feng, 1990; Quinlan, 1990), the learner induces a set of Horn clauses to represent a desired relation. It receives as input a set of facts, in the form of predicates (e.g., `sibling(tom,jane)` and `male(tom)`), defining a set of samples of the relation to be induced. For example, an ILP system could induce a set of clauses to define the relationship "uncle" from background facts extensionally defining useful predicates for this relationship (e.g., parent, sibling, male, etc.). ILP is related to RATLE in that ILP addresses problems that may require procedural solutions, such as ones that require recursive solutions like the "ancestor" relationship. My work mostly differs from ILP systems in that I focus on reinforcement learning rather than learning logical relationships, and in that I use connectionism as my induction method.

Recent work in ILP (Cohen, 1994; Pazzani & Kibler, 1992; Richards & Mooney, 1995) has produced methods that can refine an initial set of clauses with respect to a set of examples defined intensionally by background predicates. These systems introduce operators that can make changes to the current theory. They work by selecting operators that refine the current domain theory in order to better fit the background predicates. They are also able to make use of other advice such as indications of the types of the arguments for a clause. As with other ILP approaches, my work differs in that RATLE applies to RL tasks, rather than Horn-clause learning. Also, RATLE is able to assimilate advice provided at any time during the learning process.

8.4 Inducing Finite-State Information

My work with FSKBANN and RATLE relates to the problem of learning an finite-state automaton (FSA) from examples. Early theoretical work on this problem looked at the problem of learning an FSA from a set of example sequences. Gold (1972) designed an algorithm that makes a hypothesis about a set of states and then suggests a test string to further define the automaton. An oracle identifies this test string as an accepted or rejected string, and Gold's algorithm then adjusts its hypothesis. Gold (1978) later showed that the problem of inferring the *minimum* (in terms of number of states) automaton from a set of data is an *NP*-complete problem. Angluin (1987) designed an alternate learning method involving hypotheses and counterexamples. Her system makes a hypothesis about the correct automaton and then asks an oracle to suggest a counterexample. The system uses the counterexample to fix the original automaton.

Rivest and Schapire (1987) investigated an alternate method. Their technique is based on the idea that there are a limited number of sets of actions resulting in new states, where states are defined as combinations of environmental sensor values. The structure they produce, an *update graph*, specifies the compositionality of series of actions. Rivest and Schapire developed a method for learning an update graph by exploration. An FSA can be exhaustively built from the update graph, starting with the start state and applying each possible operator to determine the set of possible result states.

More recently, researchers have proposed neural-network architectures for incorporating information about state. Jordan (1989) and Elman (1990, 1991) introduced the Simple Recurrent neural Network (SRN) architecture that I presented in Chapter 2. The idea of retaining a state or context across training patterns occurs primarily in work addressing natural language problems (Cleeremans et al., 1989; Elman, 1990). The idea of using the type of network introduced by Jordan to represent a finite-state automaton was first discussed by Cleeremans et al. (1989). They show that this type of network can perfectly learn to recognize a grammar derived from an FSA. The major difference between my research and Cleeremans et al. is that we focus on using an initial domain theory expressed as a finite-state automaton, rather than attempting to learn it solely from training examples.

Omlin and Giles

Giles et al. (1992) demonstrated the idea of learning an FSA with a second-order recurrent neural network (see Figure 48). A second-order recurrent neural network's input vector consists of the current input character to the FSA and the current state (mapped with recurrent links). The next layer of a second-order recurrent neural network is a fixed set of hidden units (the "second-order" units) that calculate the product of each of the input character units with each of the state units (i.e., there are $\#inputs\ x\ \#states$ of these units). These units correspond to the set of all possible transitions in an FSA. Finally, there are a set of links from these second-order units to the set of output units representing states. The learning algorithm must learn the links from the second-order units to the states in order to

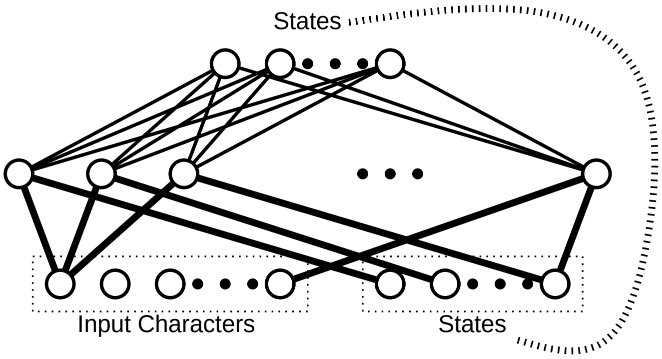


Figure 48: A second-order recurrent neural network. This type of network has a set of hidden units ("second-order" units) that calculate the multiplicative combination of each of the input values with each of the states. These second-order units are fixed; the only links that may be altered in a second-order network are the links from the second-order units to the outputs (states).

produce a network that predicts whether a string is accepted or rejected by the FSA. Giles et al. (1992) demonstrated that such an architecture can be used to both learn an FSA and to extract the learned FSA. As with the Cleeremans et al. work, RATLE differs from the Giles et al. work in that I focus on inserting prior knowledge about the finite-state task.

Omlin and Giles (1992) present an approach that closely relates to FSKBANN. In their work they employ a second-order recurrent neural network to learn FSAs. They also make use of "hints" in the form of instructions that indicate some of the transitions in the FSA. To map these hints, Omlin and Giles simply make the link corresponding to the hint a strong link. For example, a hint stating that there is a transition from state A to B on character x would cause a strong connection from the second-order unit representing conjunction of state A and input x to state B. My work differs from Omlin and Giles' work in several ways. In FSKBANN, I can translate complex transitions that are dependent on combinations of input features, while Omlin and Giles can only map conditions dependent on a single input feature. FSKBANN can map a domain theory containing more than one FSA. Finally, the general instructions of RATLE allow the teacher to provide information about multiple state transitions at once.

Frasconi, Gori, Maggini and Soda

Frasconi et al. (1995) have also examined an an approach for learning a finite-state automaton where some information about the FSA is known. In their approach, they use Local

Feedback Multi-Layered Networks (Frasconi et al., 1992). This type of network has recurrent connections within the hidden layer of the network. They show that this type of network will achieve a stable activation state for each input value in a finite number of steps. To encode the set of transitions in an initial FSA, they define a set of constraints on the weights and biases of the network based on the initial transitions and use linear programming to select a feasible point in weight space that meets those constraints. My work differs from theirs in that I insert the knowledge about the finite-state process into the network directly, without having to use linear programming to select the weights. Also, I allow the teacher to specify general information about multiple transitions.

8.5 Developing Robot-Programming Languages

A number of researchers have introduced languages for programming robot-like agents (e.g. Brooks, 1990; Chapman, 1991; Gat, 1991; Kaelbling, 1987; Nilsson, 1994). In general, these languages focus on turning a set of instructions, often in the form of a programming language, into a form that is easy for the agent to execute. My work differs from these approaches in that I focus on how to give instruction to an agent that is able to learn, thus the instructions need not be completely correct.

Kaelbling and Rosenschein (Kaelbling, 1987; Kaelbling & Rosenschein, 1990) built the language REX that produces a reactive plan for a robot agent. The REX language, like RATLE's, is based on a programming language. REX contains constructs to specify the effects of acts that the agent may take and goals the agent may want to achieve. Kaelbling and Rosenschein (1990) built a compiler that could turn a set of constructs in REX into a circuit that could efficiently determine the appropriate next action given the current world state. Thus, REX and RATLE are related in that they both produce reactive agents. The main difference between them is that in RATLE the agent can refine its instructions, while REX must receive correct instructions.

Crangle and Suppes (1994) investigated how a robot can understand a human's instructions that is expressed in ordinary English. Their work focuses on mapping certain types of English statements, such as simple imperative sentences, into corresponding robot commands. However, unlike RATLE they do not address corrections, by the learner, of approximately correct instruction.

8.6 Summary

The work in this thesis relates to research in a number of areas, which I presented in this chapter. In the first section, I described approaches to giving advice to a problem solver. My work primarily differs from these approaches in that they focus on understanding the teacher's instructions, while my techniques try to refine the instructions. Thus my work is able to deal with advice that is not completely correct. However, my work is less able to perform the elaborate operationalization that allows these approaches to understand vague advice.

More recently, a number of researchers have investigated methods for supplying hints to a reinforcement-learning agent, techniques which closely relate to my work. My work differs from most of these methods in that I allow the teacher to provide general advice, which may refer to an action or actions to take in many situations, while the teacher in other approaches provides a recommendation about a single action to take in a state. These techniques require the agent to use empirical learning in order to generalize the teacher's recommendation.

Theory refinement is a second major area of related research. My work differs from typical theory-refinement approaches in that I focus on procedural domain theories. Also, my work addresses reinforcement-learning tasks in multi-agent environments, while most theory-refinement techniques have been designed for classification tasks. A major difference with my work and other theory-refinement approaches is that my work incorporates advice provided at any time by the teacher, rather than only at the beginning of training.

One closely related area of theory refinement is knowledge-based neural networks. My work augments the knowledge-based neural network technique KBANN (Towell et al., 1990), to translate a much broader instruction language. My work differs from some techniques for producing knowledge-based networks in that I produce networks that can be refined with standard learning algorithms, rather than requiring new learning methods. My work differs from other techniques because I directly incorporate advice into the network, rather than using neural-network training to incorporate advice.

Inductive logic programming has been applied to theory refinement. Such approaches differ from mine in that they focus on learning logical relationships, rather than learning utility functions for RL agents.

A third major area of related research are techniques for learning finite-state automata. Several researchers have developed recurrent network mechanisms that can induce a finite-state automaton. My work differs from these mainly in that I am able to provide *general* advice to the network about the finite-state process.

Finally, my work differs from most languages for instructing robots in that I refine the instructions given to the agent (robot), where other techniques assume that the provided instructions are correct.

Chapter 9

Conclusions

In this thesis, I define and evaluate two systems, FSKBANN and RATLE, that expand the language a teacher can use to instruct a computer learner. Both FSKBANN and RATLE are techniques for refining domain theories with training examples. These systems are novel because they accept *procedural* domain theories from their teacher; RATLE also allows the teacher to continuously give advice to the learner. To evaluate FSKBANN and RATLE, I performed several experiments. The central question of these experiments is: does a technique for refining procedural domain theories demonstrate the same type of appealing benefits as do techniques for refining non-procedural domain theories? That is, does the learner with instruction learn faster than a learner without instruction, and is the learner able to produce a refined theory that is more accurate than the original domain theory? My experiments with my two systems indicate that the answer to these questions are yes.

In the next section, I outline the specific contributions made by this thesis. Following that I will discuss some limitations and future directions for my work.

9.1 Contributions of this Thesis

The contributions of this thesis are as follows:

- My FSKBANN system allows a teacher to provide domain theories in the form of propositional rules augmented with finite-state automata (FSAs). The teacher uses an FSA to capture the contextual aspects of a procedural task. For example, the teacher can specify a rule for predicting the current secondary-structure of an amino acid that is based on the prediction the learner made for the previous amino acid. The states of the FSA act as the learner's "memory" for information from previous steps of the task.

  FSKBANN inserts the knowledge from the propositional rules and FSAs into a knowledge-based neural network. My approach is based on the KBANN technique (Towell et al., 1990), which I extend to represent the states of the FSAs. The states of the teacher's

FSAs are represented in FSKBANN using a Simple Recurrent neural Network (Elman, 1990; Jordan, 1989). Each state corresponds to a unit with a recurrent link that remembers the activation of the state from the previous step of the network. FSKBANN translates the transitions from the FSAs to rules that determine whether the units representing states are active or inactive.

- Tests of FSKBANN on the Chou-Fasman (1978) algorithm, a method for predicting the secondary structure of globular proteins, show that FSKBANN is able to successfully refine a procedural domain theory. My tests also show that the refined theory is more accurate than FSKBANN's underlying inductive learning method, neural networks, would be without the domain theory.

  My in-depth analysis of these results indicates that the refined domain theory does a good job of predicting all three secondary structure classes, whereas the neural network approach achieves good performance only by focusing on predicting the largest secondary-structure class (coil, the default class). This is important, since biologists prefer an approach that focuses on predicting the other, more meaningful, classes of secondary structure (α-helices and β-sheets).

- I also define the RATLE advice language, which allows a teacher to communicate instructions to a reinforcement-learning agent. Instructions in this language take the form of simple programming-language constructs. These constructs indicate actions the agent should take under certain conditions in the world. The teacher defines these conditions using logical combinations of input features and teacher-defined intermediate terms. The resulting language provides a simple, yet powerful, mechanism for instruction. Novel features of the RATLE advice language include:

  - The advice language allows the teacher to describe general states of the world, thus the teacher is able to give instruction about the appropriate action to take in multiple situations. This is an improvement, since most other techniques for advising a reinforcement learner only accept recommendations about the action to take in a specific situation, and require the learner to generalize the recommendation.

  - RATLE has statements that the teacher can use to define loops (i.e., REPEAT-UNTIL) the agent should perform. RATLE represents these loops using state units in a way that is completely transparent to the teacher.

  - A teacher can give the agent instructions indicating multiple-step plans the agent should take. These multiple-step plans are composed of sequences of single actions, and, as with loops, the plans are implemented with state units.
  - The teacher can also indicate actions that the agent should NOT take under certain circumstances.
  - RATLE includes conditions that allow the teacher to use fuzzy terms like "small" and "light." The set of fuzzy terms available to the teacher is defined by a separate person that I refer to as the initializer. The initializer uses RATLE's fuzzy-term language to create fuzzy terms for a task. RATLE automatically maps any fuzzy terms the teacher specifies to the definition supplied by the initializer.
  - Advice in RATLE can be supplied continuously. RATLE translates language statements into *additions* to the agent's neural network. RATLE maps as many instructions as the teacher is willing to supply. The teacher can thus select advice that addresses current problems with the agent's behavior.

- I tested RATLE on two testbeds: Pengo and Soccer. In Pengo, an agent explores a maze-line environment, trying to eat food, while avoiding or eliminating enemies. The agents in Soccer are part of a team attempting to score a goal while preventing the other team from scoring. In the Soccer testbed the agent receives complete information about the environment, while in Pengo the agent only receives information about objects in line-of-sight around it. In both testbeds, the agents work against a group of competitive agents, but in the Soccer testbed there are also cooperative agents (i.e., players on the same team).

  Tests on these two testbeds indicate that a RATLE agent receiving advice outperforms an agent that does not. My experiments also show that it is important for RATLE's agent to be able to refine the instructions, since when the agent refines instructions, it outperforms a version that does not refine the teacher's statements. An analysis of the performance of the agents after refining advice shows that the agent produces results that are consistent with intent of the statement (e.g., a statement aimed at helping the agent collect food does indeed lead the agent to collect more food).

  Other experiments I performed indicate that the agent achieves similar results no matter when the teacher provides instruction. I also tested RATLE by providing an agent with a second piece of advice to show that agents are able to profitably make use of subsequent instruction. I further showed that a RATLE agent provided with "bad" advice may experience an initial loss of performance, but that with learning the agent can overcome the effects of the "bad" advice.

9.2 Limitations and Future Directions

Based on my experiences with RATLE and FSKBANN, I have identified a number of limitations and possible future directions for my research, some of which I presented previously in Sections 3.5, 5.5, and 6.4. Below I discuss other possible future research topics.

One important future topic is to evaluate my approach in other domains. In particular, I intend to explore giving advice to agents in situations where there are other competing agents that are able to learn. Preliminary tests I have performed in such situations show that agents receiving advice demonstrate initial gains in performance, but that the opposing agents adjust their performance to counter the teacher's instructions (Shavlik & Maclin, 1995). This suggests that the ability of the teacher to give recommendations continuously is crucial in this type of situation. Such a domain would also be interesting in that the teacher could provide instructions about how learners on the same team can cooperate.

Another domain of interest is software agents (Dent et al., 1992; Maes & Kozierok, 1993; Riecken, 1994). For example, a human could advise a software agent that looks for "interesting" papers on the World-Wide Web.

I also see a number of useful algorithmic extensions to RATLE. These extensions fit into three broad classes:

1. Broadening what I allow the teacher to "say" to the learner.

2. Improving the technical details involving the mapping of advice into a neural network and then refining it with connectionist Q-learning.

3. Designing algorithms that convert refined advice back into human-readable terms, thereby allowing human inspection of the refined advice and also allowing communication of the learned knowledge to other machine learners.

9.2.1 Broadening the Advice Language

My experience with RATLE has led me to consider a number of extensions to the current advice language, many of which I discussed in Section 5.5. Another form of information the teacher could supply are statements that indicate the relative value of two actions in a given situation (e.g., one action is *preferred* under some conditions). The teacher could also provide indications of the utility of actions, perhaps using fuzzy terms such as "good" or "poor." I also plan to explore mechanisms for specifying multi-user plans when I further explore domains with multiple agents.

RATLE could also be extended to work similarly to Sutton's (1991) DYNA system. The agent in RATLE could simultaneously learn both a utility function and a model of the world. The agent would then use the world model to perform simulated actions in order to obtain more training experiences, which can be important in domains where examples are very difficult or costly to obtain. The teacher could give advice about the effects of the agent's actions (i.e., the world model) in addition to advice about actions to take. This information would be used to improve the agent's world model.

Another method the teacher could use to help the learner would be to select configurations of the environment that would be particularly useful for the agent to experience, given the skill level of the learner. For example, a teacher selecting environments for a novice driver might choose to bring the driver to a large, open parking lot so that the novice could learn to handle the car without other considerations (other drivers, traffic signals, etc.). With a more advanced student such as an expert airline pilot, a teacher might select environments that the student might not ordinarily experience (e.g., taking the pilot to a flight simulator to experience emergency situations like the loss of power to an engine). In selecting the environments the agent experiences, the teacher is indirectly selecting features of the environment for the agent to focus on, and counts on the agent's inductive learning mechanism to determine those features. This type of approach would be complementary to my current work, and investigation of the potential synergy is a promising topic for future research.

9.2.2 Improving the Algorithmic Details

There are four issues involving the mapping or refining of advice that I have not previously discussed. Firstly, in my approach to connectionist Q-learning, I use a neural network as the agent's utility function, where the input vector to the network includes the state of the world, as well as one output for each of the agent's possible actions. Another possible approach would be to use a network where the input vector contains features describing a possible action in addition to the description of the current state of the world. The output of the network would then be a single unit indicating the utility of the represented input action. While this approach would make the learning task more difficult in some ways, it might be useful for tasks where actions can be parameterized. For example, a robot agent might be able to rotate itself by differing degrees. This type of action is difficult to represent as a set of discrete actions. With a representation where the action is an input to the network, the angle of rotation could be an input feature to the network. The agent would then select an action by sampling the space of possible rotations to find an appropriate rotation. The agent could also analyze the derivatives of the the weights to the features representing actions with respect to the predicted utility in order to find an appropriate action.

Mapping the RATLE language to this form of network would be straightforward. To translate a recommended action, RATLE would create a unit that is a conjunction of the unit representing the condition leading to the action and the input feature corresponding to the action, similar to Omlin and Giles' (1992) approach. It would then connect this unit to the output unit with a strong link.

Secondly, in another approach to reinforcement learning, the agent predicts the utility of a state rather than the utility of an action in a state (Sutton, 1988); in this approach the learner has a model of how its actions change the world. In this approach the agent determines the action to take by checking the utility of the states that are reachable from the current state. Applying RATLE to this type of reinforcement-learning system would be difficult, since RATLE statements suggest actions to take. In order to map a statement indicating an action, RATLE would first determine the set of states that meet the condition of the statement, then calculate the set of states that would result by following the suggested action. RATLE would then increase the utility of the states that follow from the suggested action. Other types of advice would be more straightforward under this approach. For example, if the teacher gave advice about a goal the agent should try to achieve (i.e., as in Gordon and Subramanian's (1994) approach), RATLE could determine the set of states corresponding to the goal and simply increase the utility of all of these states.

Thirdly, further tests with the replay technique (Lin, 1992) are important, since this technique allows the agent to re-use the experiences it accumulates. This ability is crucial for domains where the cost of exploration is prohibitive.

Finally, RATLE currently maintains no statistics that record how often a piece of advice was applicable and how often it was followed. I intend to add such statistics-gatherers and use them to inform the teacher that a given piece of advice was seldom applicable or followed. I also plan to keep a record of the *original* advice and compare its statistics to the *refined* version. Significant differences between the two should cause the learner to inform the teacher that some instruction has substantially changed (I plan to use the rule-extraction techniques described below when I present the refined statement to the teacher).

9.2.3 Converting Refined Advice into Human Comprehensible Terms

One interesting area of future research is the "extraction" (i.e., conversion to a easily comprehensible form) of learned knowledge from my connectionist utility function. In order to extend previous rule-extraction techniques (Craven & Shavlik, 1994; Towell & Shavlik, 1993) to advice-taking tasks, I plan to examine two tasks: extending Towell and Shavlik's (1993) language for extracted rules, and employing rule extraction to transfer learned knowledge between agents operating in the same or similar domains.

Previous rule-extraction work has been applied to domains with discrete-valued input features and has used a language of propositional inference rules to express extracted concept representations. I plan to extend these techniques so that they can handle real-valued input features, and so that they can express extracted concept representations using the same language that is used to give advice to the learners. The RATLE advice language incorporates aspects, such as multi-step operators and fuzzy conditions, for which current rule-extraction techniques provide no support. I plan to extend these rule-extraction techniques to extract concept descriptions expressed in this richer language.

I also plan to investigate the use of rule-extraction methods for knowledge transfer among machine learners. For example, one learner might ask another what it would do given a particular sensor reading. Transferring acquired knowledge from one learning system to another can speed up learning, which is particularly important in domains where training is slow or where training examples are sparse or expensive. My approach to knowledge transfer would use rule-extraction methods to elicit knowledge from trained networks, and my advice-giving method to insert transferred knowledge into other networks. That is, my advice-taking language would serve as the interlingua for communication between machine learners.

9.3 Final Summary

In this thesis I present the FSKBANN and RATLE systems. These techniques allow a teacher to give advice to a computer learner solving a procedural task. The learner can then refine the teacher's instructions. Experiments presented in Chapters 3 and 7 demonstrate that the refined domain theories produced by FSKBANN and RATLE outperform both the original domain theories and the underlying inductive learning mechanism, neural networks, when that underlying mechanism does not receive the teacher's instructions.

In Chapter 3 I outlined my initial approach, FSKBANN, which allows a teacher to give advice to a computer learner in the form of propositional rules and FSAs. FSKBANN maps the knowledge from the rules and FSAs into a neural network that represents the states from the FSAs with recurrent links. Tests with FSKBANN on the Chou-Fasman (1978) algorithm produce a refined theory that outperforms the non-learning Chou-Fasman algorithm as well as a standard neural-network approach.

I defined RATLE in Chapters 4, 5, and 6. RATLE allows a teacher to communicate advice to a connectionist, reinforcement learning agent in the form of statements in a simple programming language. The advice language allows the teacher to indicate conditions of the environment and actions the agent should take under those conditions. Novel features of the language include the following: (1) a statement may indicate *loops* the agent should execute, (2) it may suggest *multi-step* plans for the agent to follow, (3) it may recommend actions the agent should *not* take, (4) it can include references to *fuzzy* terms such as "near" and "short", and (5) advice may be given *continuously*. RATLE translates the teacher's statements into additions to the RL agents neural-network policy function, using recurrent units to represent loops and multi-step plans.

In Chapter 7, I presented experiments applying RATLE to two simulated testbeds involving multiple agents: Pengo and Soccer. These tests demonstrated that a RATLE agent receiving advice outperforms both an agent that does not receive advice and an agent that receives the instruction but is unable to refine it. Analysis of these results indicate that the effects of the instruction conform to my expectations (e.g., recommendations about avoiding enemies helps the agent survive longer). Other experiments show that the effect of instruction seems to be independent of when the teacher supplies the instruction to the agent, that the agents are able to profitably make use of a subsequent piece of advice, and that the agents can overcome the effects of "bad" advice with learning.

In conclusion, I have defined an appealing approach for learning from both instruction and experience in procedural tasks. This work widens the "information pipeline" between humans and machine learners, without requiring that the human provide absolutely correct information to the learner.

Appendix A

Details of the Pengo Testbed

In this appendix I present further details about the Pengo testbed investigated in Chapter 7. I start by describing the processes used to create Pengo environments. I then define input features, fuzzy terms, and advice used in Chapter 7's experiments.

A.1 Creating Initial Pengo Environments

In order to perform experiments in the Pengo environment, I randomly construct a series of initial Pengo environments. To construct an initial environment, I start with an empty 7x7 grid and then to deposit a series of randomly generated "rooms" on the grid. Each "room" is a hollow rectangle whose border is composed of obstacle objects, and where the length and width of the rectangle are random numbers between 1 and 7 (inclusive). I randomly select coordinates for one of the corner points and place the room on the grid (any portion of the room that is off the grid is then discarded). I continue this process until the grid is at least 35% full of obstacles.

I then use a standard algorithm for connected components (Sedgewick, 1983, pp. 384–385) to insure that all of the open spaces in the grid are reachable from any other. To do this, I construct the connected components of the grid, and then remove obstacles in a line between the two closest components until they are connected. I repeat this process until all of the components are connected. The resulting grid generally has approximately 15 obstacles.

After setting up an initial set of obstacles, I choose a location for the agent randomly among the remaining open spots of the grid. I then choose initial locations for the foods from among the remaining open spots on the grid. Finally, I choose initial locations for the enemies from the remaining open spots, with the caveat that these locations must be at least two grid units distant from the agent's initial location.

A.2 Input Features

As I discussed in Chapter 7, the sensors of the agent in the Pengo consist of a series of values indicating how many of each kind of object exist in a set of sectors around the agent. The agent measures six values for each sector: the number of enemies, foods, walls, and obstacles, as well as how much of each sector is empty and how much is occluded. I define each sector by four values – a minimum and maximum angle and a minimum and maximum distance from the agent. There are eight pairs of minimum and maximum angles around the agent, with the 360° range of values evenly divided. There are four sets of minimum and maximum distances: {0,1}, {1,3}, {3,6}, and {6,10}. Thus, there are 192 sensor input features. To this I add nine Boolean input features representing, in an 1-of-N encoding, the previous action taken by the agent. The final total is 201 input features.

Sensor inputs are REALs of the form:

REAL `Enemy` `Angle` IN [0.393 .. 1.178] `Distance` IN [3 .. 6] [0 .. 10.6 , 0 .. 1]

This example describes an input unit that represents up to 10.6 `Enemies` (parts of an object may be in multiple sectors), where the value is normalized to be between 0 and 1. These `Enemies` are from three to six units distant from the agent, and the angle to these `Enemies` is between 0.393 ($\frac{\pi}{8}$) to 1.178 ($\frac{3\pi}{8}$).

A.3 Fuzzy Terms

The fuzzy terms set up for the Pengo environment are as follows:

QUANTIFIER `No` ::= SUM < 1
QUANTIFIER `A` ::= SUM = 1
QUANTIFIER `An` ::= SUM = 1
QUANTIFIER `Many` ::= SUM > 2

PROPERTY `NextTo` ::= `Distance` IN [0 .. 1]
PROPERTY `Near` ::= `Distance` IN [0 .. 2.5]
PROPERTY `Medium` ::= `Distance` IN [2.5 .. 5]
PROPERTY `Far` ::= `Distance` IN [5 .. 10]

PROPERTY `East` ::= `Angle` IN [-0.785 .. 0.785]
PROPERTY `West` ::= `Angle` IN [0.785 .. 2.356]
PROPERTY `North` ::= `Angle` IN [2.356 .. 3.927]
PROPERTY `South` ::= `Angle` IN [3.927 .. 5.498]

Note that the thresholds for these functions can be changed by the agent in its network through training.

A.4 Advice Provided

The statements I use to create the six pieces of advice used in the Pengo experiments in Chapter 7 appear below. Figures 37, 38, 39, and 40 pictorially depict the first four pieces of advice, respectively.

Good Advice

SimpleMoves:

```
FOREACH dir IN { East, North, West, South }
  IF An Obstacle IS (NextTo AND $(dir)) THEN
    INFER OkPush$(dir)
  END;
  IF No Obstacle IS (NextTo AND $(dir)) AND
    No Wall IS (NextTo AND $(dir)) THEN
    INFER OkMove$(dir)
  END;
  IF An Enemy IS (Near AND $(dir)) THEN
    DO_NOT Move$(dir)
  END;
  IF OkMove$(dir) AND
    A Food IS (Near AND $(dir)) AND
    No Enemy IS (Near AND $(dir)) THEN
    Move$(dir)
  END;
  IF OkPush$(dir) AND
    An Enemy IS (Near AND $(dir)) THEN
    Push$(dir)
  END
ENDFOREACH
```

NonLocalMoves:

```
FOREACH dir IN { East, North, West, South }
  IF No Obstacle IS (NextTo AND $(dir)) AND
    No Wall IS (NextTo AND $(dir)) THEN
    INFER OkMove$(dir)
```

```
    END;
    IF Many Enemy ARE (NOT $(dir)) AND
      No Enemy IS (Near AND $(dir)) AND
      OkMove$(dir) THEN
      Move$(dir)
    END;
    IF OkMove$(dir) AND
      An Enemy IS ($(dir) AND {Medium OR Far}) AND
      No Enemy IS ($(dir) AND Near) AND
      A Food IS ($(dir) AND Near) THEN
      Move$(dir)
    END
  ENDFOREACH
```

ElimEnemies:

```
  FOREACH dir IN { East, North, West, South }
    IF No Obstacle IS (NextTo AND $(dir)) AND
      No Wall IS (NextTo AND $(dir)) THEN
      INFER OkMove$(dir)
    END;
  ENDFOREACH;
  FOREACH ( ahead, back, side1, side2 ) IN
          {( East, West, North, South ), ( East, West, South, North ),
          ( North, South, East, West ), ( North, South, West, East ),
          ( West, East, North, South ), ( West, East, South, North ),
          ( South, North, East, West ), ( South, North, West, East )}
    IF OkMove$(ahead) AND
      An Enemy IS (Near AND $(back)) AND
      An Obstacle IS (NextTo AND $(side1)) THEN
      MULTIACTION
          Move$(ahead)
          Move$(side1)
          Move$(side1)
          Move$(back)
          Push$(side2)
      END
    END
  ENDFOREACH
```

Surrounded:

```
  FOREACH dir IN { East, North, West, South }
```

```
    IF An Obstacle IS (NextTo AND $(dir)) THEN
      INFER OkPush$(dir)
    END;
    IF An Enemy IS (Near AND $(dir)) OR
      A Wall IS (NextTo AND $(dir)) OR
      An Obstacle IS (NextTo AND $(dir)) THEN
      INFER Blocked$(dir)
    END;
    WHEN Surrounded AND OkPush$(dir) AND An Enemy IS Near
      REPEAT
          MULTIACTION
            Push$(dir)
            Move$(dir)
          END
      UNTIL NOT OkPush$(dir)
    END
  ENDFOREACH;
  IF BlockedEast AND BlockedNorth AND
    BlockedSouth AND BlockedWest THEN
    INFER Surrounded
  END
```

Bad Advice

NotSimpleMoves:

```
  FOREACH dir IN { East, North, West, South }
    IF An Obstacle IS (NextTo AND $(dir)) THEN
      INFER OkPush$(dir)
    END;
    IF No Obstacle IS (NextTo AND $(dir)) AND
      No Wall IS (NextTo AND $(dir)) THEN
      INFER OkMove$(dir)
    END;
    IF An Enemy IS (Near AND (NOT $(dir))) THEN
      DO_NOT Move$(dir)
    END;
    IF OkMove$(dir) AND
      A Food IS (Near AND $(dir)) AND
      No Enemy IS (Near AND $(dir)) THEN
      DO_NOT Move$(dir)
    END;
    IF OkPush$(dir) AND
```

```
    An Enemy IS (Near AND $(dir)) THEN
    DO_NOT Push$(dir)
  END
ENDFOREACH
```

OppositeSimpleMoves:

```
FOREACH (dir, oppdir) IN { (East, West), (North, South),
                           (West, East), (South, North) }
  IF An Obstacle IS (NextTo AND $(dir)) THEN
    INFER OkPush$(dir)
  END;
  IF No Obstacle IS (NextTo AND $(dir)) AND
    No Wall IS (NextTo AND $(dir)) THEN
    INFER OkMove$(dir)
  END;
  IF An Enemy IS (Near AND $(dir)) THEN
    Move$(dir)
  END;
  IF OkMove$(dir) AND
    A Food IS (Near AND $(dir)) AND
    No Enemy IS (Near AND $(dir)) THEN
    Move$(oppdir)
  END;
  IF OkPush$(dir) AND
    An Enemy IS (Near AND $(dir)) THEN
    Push$(oppdir)
  END
ENDFOREACH
```

Appendix B

Details of the Soccer Testbed

In this appendix I present the input terms, fuzzy terms, and advice used in the Soccer testbed investigated in Chapter 7.

B.1 Input Features

The input features for each Soccer agent describe the current location for every object on the field. The location and direction of an object is calculated relative to the agent's current position. I describe each object with five values: the *distance* to the object, the x component of the distance to the object, the y component of the distance to the object, the *sine* of the angle to the object, and the *cosine* of the same angle. Some of these features are redundant, but they each provide information that can be useful. The name for an input feature is formed using one of the five prefixes (`DistTo`, `XDistTo`, `YDistTo`, `CosTo`, or `SinTo`), followed by the name of the object. The objects whose distances are measured are: the ball (`Ball`); the opposing players (`Opposition1` and `Opposition2`); the player on the same team (`TeamMate1`); and the left post, right post, and center of each goal (`MyGoalLeftPost`, `OppositionGoalLeftPost`, `MyGoalCenter`, `OppositionGoalCenter`, `MyGoalRightPost`, and `OppositionGoalRightPost`). A sample input feature is `YDistToMyGoalCenter`, which is the distance in the y plane to the player's own goal. The input vector also includes distances to each wall (`LeftEnd`, `RightEnd`, `FarEnd`, and `NearEnd`). All of these features are REAL input features. There is also a BOOLEAN input feature that is true when the player has the ball `HasBall` and a REAL value representing the clock `TimeLeft`. Finally, there are BOOLEAN input features that indicate the previous action taken (in the form *Action*`Taken`). There are 67 total input features.

B.2 Fuzzy Terms

The fuzzy terms for the Soccer environment are:

```
DESCRIPTOR Near  ::= DistTo?(Object) <  2.0
DESCRIPTOR East  ::= CosTo?(Object)  >  0.95
DESCRIPTOR North ::= SinTo?(Object)  >  0.95
DESCRIPTOR West  ::= CosTo?(Object)  < -0.95
DESCRIPTOR South ::= SinTo?(Object)  < -0.95
```

B.3 Advice Provided

The advice used in the Soccer experiments in Chapter 7 appears below. Figure 46 pictorially depicts this advice.

```
FOREACH opp IN { Opposition1, Opposition2 }
  FOREACH dir IN { East, North, West, South }
    IF $(opp) IS Near AND
      $(opp) IS $(dir)) THEN
      INFER Guarded$(dir)
    END
  ENDFOREACH
ENDFOREACH;

IF HaveBall AND NOT GuardedForward THEN
  MoveForward
END;

IF HaveBall AND GuardedForward THEN
  (MoveRight OR MoveRight)
END;

IF HaveBall AND GuardedForward AND GuardedLeft THEN
  MoveRight
END;

IF HaveBall AND GuardedForward AND GuardedRight THEN
  MoveLeft
END;
```

Bibliography

Abu-Mostafa, Y. (1995). Hints. *Neural Computation*, 7:639–671.

Agre, P. & Chapman, D. (1987). Pengi: An implementation of a theory of activity. In *Proceedings of the Sixth National Conference on Artificial Intelligence*, (pp. 268–272), Seattle, WA.

Anderson, C. (1987). Strategy learning with multilayer connectionist representations. In *Proceedings of the Fourth International Workshop on Machine Learning*, (pp. 103–114), Irvine, CA.

Angluin, D. (1987). Learning regular sets from queries and counterexamples. *Information and Computation*, 75:87–106.

Atlas, L., Cole, R., Muthusamy, Y., Lippman, A., Connor, J., Park, D., El-Sharkawi, M., & Marks, R. (1990). A performance comparison of trained multilayer perceptrons and trained classification trees. *Proceedings of the IEEE*, 78:1614–1619.

Barto, A., Bradtke, S., & Singh, S. (1995). Learning act using realtime dynamic programming. *Artificial Intelligence*, 72:81–138.

Barto, A., Sutton, R., & Anderson, C. (1983). Neuronlike adaptive elements that can solve difficult learning control problems. *IEEE Transactions on Systems, Man, and Cybernetics*, 13:834–846.

Barto, A., Sutton, R., & Watkins, C. (1990). Learning and sequential decision making. In Gabriel, M. & Moore, J., editors, *Learning and Computational Neuroscience*. MIT Press, Cambridge, MA.

Berenji, H. & Khedkar, P. (1992). Learning and tuning fuzzy logic controllers through reinforcements. *IEEE Transactions on Neural Networks*, 3:724–740.

Brooks, R. (1990). The behavior language; user's guide. AI Memo 1227, Artificial Intelligence Laboratory, Massachusetts Institute of Technology, Cambridge, MA.

Chapman, D. (1991). *Vision, Instruction, and Action*. MIT Press, Cambridge, MA.

Chou, P. & Fasman, G. (1978). Prediction of the secondary structure of proteins from their amino acid sequence. *Advances in Enzymology*, 47:45–148.

Cleeremans, A., Servan-Schreiber, D., & McClelland, J. (1989). Finite state automata and simple recurrent networks. *Neural Computation*, 1:372–381.

Clouse, J. & Utgoff, P. (1992). A teaching method for reinforcement learning. In *Proceedings of the Ninth International Conference on Machine Learning*, (pp. 92–101), Aberdeen, Scotland.

Cohen, B., Presnell, S., Cohen, F., & Langridge, R. (1991). A proposal for feature-based scoring of protein secondary structure predictions. In *Proceedings of the AAAI91 Workshop on AI Approaches to Classification and Pattern Recognition in Molecular Biology*, (pp. 5–20), Anaheim, CA.

Cohen, P. & Feigenbaum, E. (1982). *The Handbook of Artificial Intelligence (volume 3)*. William Kaufmann, Los Altos, CA.

Cohen, W. (1994). Grammatically biased learning: Learning logic programs using an explicit antecedent description language. *Artificial Intelligence*, 68:303–366.

Cost, S. & Salzberg, S. (1993). A weighted nearest neighbor algorithm for learning with symbolic features. *Machine Learning*, 10:57–78.

Crangle, C. & Suppes, P. (1994). *Language and Learning for Robots*. CSLI Publications, Stanford, CA.

Craven, M. & Shavlik, J. (1994). Using sampling and queries to extract rules from trained neural networks. In *Proceedings of the Eleventh International Conference on Machine Learning*, (pp. 37–45), New Brunswick, NJ.

Dent, L., Boticario, J., McDermott, J., Mitchell, T., & Zabowski, D. (1992). A personal learning apprentice. In *Proceedings of the Tenth National Conference on Artificial Intelligence*, (pp. 96–103), San Jose, CA.

Diederich, J. (1989). "Learning by instruction" in connectionist systems. In *Proceedings of the Sixth International Workshop on Machine Learning*, (pp. 66–68), Ithaca, NY.

Dietterich, T. (1991). Knowledge compilation: Bridging the gap between specification and implementation. *IEEE Expert*, 6:80–82.

Elman, J. (1990). Finding structure in time. *Cognitive Science*, 14:179–211.

Elman, J. (1991). Distributed representations, simple recurrent networks, and grammatical structure. *Machine Learning*, 7:195–225.

Fahlman, S. & Lebiere, C. (1990). The cascade-correlation learning architecture. In Touretzky, D., editor, *Advances in Neural Information Processing Systems (volume 2)*. Morgan Kaufmann, Palo Alto, CA.

Fisher, D. & McKusick, K. (1989). An empirical comparison of ID3 and back-propagation. In *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence*, (pp. 788–793), Detroit, MI.

Frasconi, P., Gori, M., Maggini, M., & Soda, G. (1995). Unified integration of explicit knowledge and learning by example in recurrent networks. *IEEE Transactions on Knowledge and Data Engineering*, 7:340–346.

Frasconi, P., Gori, M., & Soda, G. (1992). Local feedback multi-layered networks. *Neural Computation*, 4:120–130.

Fu, L. M. (1989). Integration of neural heuristics into knowledge-based inference. *Connection Science*, 1:325–340.

Garnier, J. & Robson, B. (1989). The GOR method for predicting secondary structures in proteins. In Fasman, G., editor, *Prediction of Protein Structure and the Principles of Protein Conformation*. Plenum Press, New York.

Gat, E. (1991). ALFA: A language for programming reactive robotic control systems. In *IEEE International Conference on Robotics and Automation (volume 2)*, (pp. 1116–1121).

Giles, C., Miller, C., Chen, D., Chen, H., Sun, G., & Lee, Y. (1992). Learning and extracting finite state automata with second-order recurrent neural networks. *Neural Computation*, 4:393–405.

Ginsberg, A. (1988). *Automatic Refinement of Expert System Knowledge Bases*. Pitman, London.

Gold, E. M. (1972). System identification via state characterization. *Automatica*, 8:621–636.

Gold, E. M. (1978). Complexity of automaton identification from given data. *Information and Control*, 37:302–320.

Gordon, D. & Subramanian, D. (1994). A multistrategy learning scheme for agent knowledge acquisition. *Informatica*, 17:331–346.

Gruau, F. (1994). *Neural Network Synthesis using Cellular Encoding and the Genetic Algorithm*. PhD thesis, Ecole Normale Superieure de Lyon, France.

Hayes-Roth, F., Klahr, P., & Mostow, D. J. (1981). Advice-taking and knowledge refinement: An iterative view of skill acquisition. In Anderson, J., editor, *Cognitive Skills and their Acquisition*. Lawrence Erlbaum, Hillsdale, NJ.

Hinton, G. E. (1986). Learning distributed representations of concepts. In *Proceedings of the Eighth Annual Conference of the Cognitive Science Society*, (pp. 1–12), Amherst, MA.

Holder, L. B. (1991). *Maintaining the Utility of Learned Knowledge Using Model-Based Control*. PhD thesis, Computer Science Department, University of Illinois at Urbana-Champaign.

Holland, J. (1975). *Adaptation in Natural and Artificial Systems*. University of Michigan Press, Ann Arbor, MI.

Holland, J. (1986). Escaping brittleness: The possibilities of general-purpose learning algorithms applied to parallel rule-based systems. In Michalski, R., Carbonell, J., & Mitchell, T., editors, *Machine Learning: An AI Approach (volume 2)*. Morgan Kaufmann, San Mateo, CA.

Holley, L. & Karplus, M. (1989). Protein structure prediction with a neural network. *Proceedings of the National Academy of Sciences (USA)*, 86:152–156.

Hopcroft, J. & Ullman, J. (1979). *Introduction to Automata Theory, Languages, and Computation*. Addison Wesley, Reading, MA.

Huffman, S. & Laird, J. (1993). Learning procedures from interactive natural language instructions. In *Machine Learning: Proceedings on the Tenth International Conference*, (pp. 143–150), Amherst, MA.

Jensen, K. & Wirth, N. (1975). *PASCAL: User Manual and Report*. Springer-Verlag, New York.

Jordan, M. (1989). Serial order: A parallel, distributed processing approach. In Elman, J. & Rumelhart, D., editors, *Advances in Connectionist Theory: Speech*. Erlbaum, Hillsdale, NJ.

Kaelbling, L. (1987). REX: A symbolic language for the design and parallel implementation of embedded systems. In *Proceedings of the AIAA Conference on Computers in Aerospace*, Wakefield, MA.

Kaelbling, L. & Rosenschein, S. (1990). Action and planning in embedded agents. *Robotics and Autonomous Systems*, 6:35–48.

Laird, J., Hucka, M., Yager, E., & Tuck, C. (1990). Correcting and extending domain knowledge using outside guidance. In *Proceedings of the Seventh International Conference on Machine Learning*, (pp. 235–243), Austin, TX.

Laird, J., Newell, A., & Rosenbloom, P. (1987). SOAR: An architecture for general intelligence. *Artificial Intelligence*, 33:1–64.

Lang, K., Waibel, A., & Hinton, G. (1990). A time-delay neural network architecture for isolated word recognition. *Neural Networks*, 3:23–43.

Le Cun, Y., Denker, J., & Solla, S. (1990). Optimal brain damage. In Touretzky, D., editor, *Advances in Neural Information Processing Systems (volume 2)*. Morgan Kaufmann, Palo Alto, CA.

Leng, B., Buchanan, B., & Nicholas, H. (1993). Protein secondary structure prediction using two-level case-based reasoning. In *Proceedings of the First International Conference on Intelligent Systems for Molecular Biology*, (pp. 251–259), Washington, DC.

Levine, J., Mason, T., & Brown, D. (1992). *Lex & Yacc*. O'Reilly, Sebastopol, CA.

Lim, V. (1974). Algorithms for prediction of α-helical and β-structural regions in globular proteins. *Journal of Molecular Biology*, 88:873–894.

Lin, L. (1991). Programming robots using reinforcement learning and teaching. In *Proceedings of the Ninth National Conference on Artificial Intelligence*, (pp. 781–786), Anaheim, CA.

Lin, L. (1992). Self-improving reactive agents based on reinforcement learning, planning, and teaching. *Machine Learning*, 8:293–321.

Lin, L. (1993). Scaling up reinforcement learning for robot control. In *Proceedings of the Tenth International Conference on Machine Learning*, (pp. 182–189), Amherst, MA.

Maclin, R. & Shavlik, J. (1991). Refining domain theories expressed as finite-state automata. In *Proceedings of the Eighth International Machine Learning Workshop*, (pp. 524–528), Evanston, IL.

Maclin, R. & Shavlik, J. (1992). Using knowledge-based neural networks to improve algorithms: Refining the Chou-Fasman algorithm for protein folding. In *Proceedings of the Tenth National Conference on Artificial Intelligence*, (pp. 165–170), San Jose, CA.

Maclin, R. & Shavlik, J. (1993). Using knowledge-based neural networks to improve algorithms: Refining the Chou-Fasman algorithm for protein folding. *Machine Learning*, 11:195–215.

Maclin, R. & Shavlik, J. (1994a). Incorporating advice into agents that learn from reinforcements. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, (pp. 694–699), Seattle, WA.

Maclin, R. & Shavlik, J. (1994b). Refining algorithms with knowledge-based neural networks: Improving the Chou-Fasman algorithm for protein folding. In Hanson, S., Drastal, G., & Rivest, R., editors, *Computational Learning Theory and Natural Learning Systems (volume 1)*. MIT Press, Cambridge, MA.

Maclin, R. & Shavlik, J. (1996). Creating advice-taking reinforcement learners. *Machine Learning*.

Maes, P. & Kozierok, R. (1993). Learning interface agents. In *Proceedings of the Eleventh National Conference on Artificial Intelligence*, (pp. 459–465), Washington, DC.

Mahadevan, S. & Connell, J. (1992). Automatic programming of behavior-based robots using reinforcement learning. *Artificial Intelligence*, 55:311–365.

Mataric, M. (1994). Reward functions for accelerated learning. In *Proceedings of the Eleventh International Conference on Machine Learning*, (pp. 181–189), New Brunswick, NJ.

Mathews, B. (1975). Comparison of the predicted and observed secondary structure of T4 Phage Lysozyme. *Biochimica et Biophysica Acta*, 405:442–451.

McCarthy, J. (1958). Programs with common sense. In *Proceedings of the Symposium on the Mechanization of Thought Processes (volume I)*, (pp. 77–84). (Reprinted in M. Minsky, editor, 1968, *Semantic Information Processing.* Cambridge, MA: MIT Press, 403–409.).

Minsky, M. & Papert, S. (1969). *Perceptrons.* MIT Press, Cambridge.

Mitchell, T., Keller, R., & Kedar-Cabelli, S. (1986). Explanation-based generalization: A unifying view. *Machine Learning*, 1:47–80.

Moody, J. & Darken, C. (1988). Learning with localized receptive fields. In Hinton, G., Sejnowski, T., & Touretzky, D., editors, *Proceedings of the 1988 Connectionist Models Summer School*, (pp. 133–143), San Mateo, CA. Morgan Kaufmann.

Moody, J. & Darken, C. (1989). Fast learning in networks of locally-tuned processing units. *Neural Computation*, 1:281–294.

Mostow, D. J. (1982). Transforming declarative advice into effective procedures: A heuristic search example. In Michalski, R., Carbonell, J., & Mitchell, T., editors, *Machine Learning: An Artificial Intelligence Approach (volume 1).* Tioga Press, Palo Alto.

Muggleton, S. (1992). *Inductive logic programming.* Academic Press, London.

Muggleton, S. & Feng, C. (1990). Efficient induction of logic programs. In *Proceedings of the First Conference on Algorithmic Theory*, (pp. 1–14), Tokyo, Japan.

Nilsson, N. (1994). Teleo-reactive programs for agent control. *Journal of Artificial Intelligence Research*, 1:139–158.

Nishikawa, K. (1983). Assessment of secondary-structure prediction of proteins: Comparison of computerized Chou-Fasman method with others. *Biochimica et Biophysica Acta*, 748:285–299.

Noelle, D. & Cottrell, G. (1994). Towards instructable connectionist systems. In Sun, R. & Bookman, L., editors, *Computational Architectures Integrating Neural and Symbolic Processes.* Kluwer Academic, Boston.

Nowlan, S. & Hinton, G. (1992). Simplifying neural networks by soft weight-sharing. *Neural Computation*, 4:473–493.

Omlin, C. & Giles, C. (1992). Training second-order recurrent neural networks using hints. In *Proceedings of the Ninth International Conference on Machine Learning*, (pp. 361–366), Aberdeen, Scotland.

Opitz, D. & Shavlik, J. (1993). Heuristically expanding knowledge-based neural networks. In *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence*, (pp. 1360–1365), Chambery, France.

Ourston, D. & Mooney, R. (1990). Changing the rules: A comprehensive approach to theory refinement. In *Proceedings of the Eighth National Conference on Artificial Intelligence*, (pp. 815–820), Boston, MA.

Ourston, D. & Mooney, R. (1994). Theory refinement combining analytical and empirical methods. *Artificial Intelligence*, 66:273–309.

Pazzani, M. & Kibler, D. (1992). The utility of knowledge in inductive learning. *Machine Learning*, 9:57–94.

Pineda, F. (1987). Generalization of back-propagation to recurrent neural networks. *Physical Review Letters*, 59:2229–2232.

Qian, N. & Sejnowski, T. (1988). Predicting the secondary structure of globular proteins using neural network models. *Journal of Molecular Biology*, 202:865–884.

Quinlan, J. (1990). Learning logical definitions from relations. *Machine Learning*, 5:239–2666.

Richards, B. & Mooney, R. (1995). Automated refinement of first-order Horn-clause domain theories. *Machine Learning*, 19:95–131.

Richardson, J. & Richardson, D. (1989). Principles and patterns of protein conformation. In Fasman, G., editor, *Prediction of Protein Structure and the Principles of Protein Conformation*. Plenum Press, New York.

Riecken, D. (1994). Special issue on intelligent agents. *Communications of the ACM*, 37(7).

Rivest, R. & Schapire, R. (1987). A new approach to unsupervised learning in deterministic environments. In *Proceedings of the Fourth International Workshop on Machine Learning*, (pp. 364–375), Irvine, CA.

Rost, B. & Sander, C. (1993). Prediction of protein secondary structure at better than 70% accuracy. *Journal of Molecular Biology*, 232:584–599.

Rumelhart, D., Hinton, G., & Williams, R. (1986). Learning internal representations by error propagation. In Rumelhart, D. & McClelland, J., editors, *Parallel Distributed Processing: Explorations in the microstructure of cognition. (volume 1)*. MIT Press, Cambridge, MA.

Sacerdoti, E. (1974). Planning in a hierarchy of abstraction spaces. *Artificial Intelligence*, 5:115–135.

Salzberg, S. & Cost, S. (1992). Predicting protein secondary structure with a nearest-neighbor algorithm. *Journal of Molecular Biology*, 227:371–374.

Samuel, A. (1959). Some studies in machine learning using the game of checkers. *IBM Journal on Research and Development*, 3:210–229. (Reprinted in E. Feigenbaum and J. Feldman, eds., 1963, *Computers and Thought*. McGraw-Hill, New York).

Schoppers, M. (1994). Estimating reaction plan size. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, (pp. 1238–1244), Seattle, WA.

Sedgewick, R. (1983). *Algorithms*. Addison Wesley, Redding, MA.

Sejnowski, T. & Rosenberg, C. (1987). Parallel networks that learn to pronounce English text. *Complex Systems*, 1:145–168.

Shastri, L. (1988). A connectionist approach to knowledge representation and limited inference. *Cognitive Science*, 12:331–392.

Shavlik, J. & Maclin, R. (1995). Learning from instruction and experience in competitive situations. In *Proceedings of the ML95 Workshop on Agents that Learn from Other Agents*, Tahoe City, CA.

Shavlik, J., Mooney, R., & Towell, G. (1991). Symbolic and neural learning algorithms: An experimental comparison. *Machine Learning*, 6:111–143.

Shavlik, J. & Towell, G. (1989). An approach to combining explanation-based and neural learning algorithms. *Connection Science*, 1:233–255.

Siegelmann, H. (1994). Neural programming language. In *Proceedings of the Twelfth National Conference on Artificial Intelligence*, (pp. 877–882), Seattle, WA.

Suddarth, S. & Holden, A. (1991). Symbolic-neural systems and the use of hints for developing complex systems. *International Journal of Man-Machine Studies*, 35:291–311.

Sun, R. (1992). On variable binding in connectionist networks. *Connection Science*, 4:93–124.

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

Sutton, R. (1991). Reinforcement learning architectures for animats. In Meyer, J. & Wilson, S., editors, *From Animals to Animats: Proceedings of the First International Conference on Simulation of Adaptive Behavior*. MIT Press, Cambridge, MA.

Tesauro, G. (1992). Practical issues in temporal difference learning. *Machine Learning*, 8:257–277.

Thrun, S. (1994). Personal communication.

Thrun, S. & Mitchell, T. (1993). Integrating inductive neural network learning and explanation-based learning. In *Proceedings of the Thirteenth International Joint Conference on Artificial Intelligence*, (pp. 930–936), Chambery, France.

Towell, G. (1991). *Symbolic Knowledge and Neural Networks: Insertion, Refinement, and Extraction*. PhD thesis, Computer Sciences Department, University of Wisconsin, Madison, WI.

Towell, G. & Shavlik, J. (1993). Extracting refined rules from knowledge-based neural networks. *Machine Learning*, 13:71–101.

Towell, G. & Shavlik, J. (1994). Knowledge-based artificial neural networks. *Artificial Intelligence*, 70:119–165.

Towell, G., Shavlik, J., & Noordewier, M. (1990). Refinement of approximate domain theories by knowledge-based neural networks. In *Proceedings of the Eighth National Conference on Artificial Intelligence*, (pp. 861–866), Boston, MA.

Utgoff, P. & Clouse, J. (1991). Two kinds of training information for evaluation function learning. In *Proceedings of the Ninth National Conference on Artificial Intelligence*, (pp. 596–600), Anaheim, CA.

Watkins, C. (1989). *Learning from Delayed Rewards*. PhD thesis, King's College, Cambridge.

Watson, J. (1990). The Human Genome Project: Past, present, and future. *Science*, 248:44–48.

Whitehead, S. (1991). A complexity analysis of cooperative mechanisms in reinforcement learning. In *Proceedings of the Ninth National Conference on Artificial Intelligence*, (pp. 607–613), Anaheim, CA.

Wilson, I., Haft, D., Getzoff, E., Tainer, J., Lerner, R., & Brenner, S. (1985). Identical short peptide sequences in unrelated proteins can have different conformations: A testing ground for theories of immune recognition. *Proceedings of the National Academy of Sciences (USA)*, 82:5255–5259.

Zadeh, L. (1965). Fuzzy sets. *Information and Control*, 8:338–353.

Zhang, X., Mesirov, J., & Waltz, D. (1992). Hybrid system for protein secondary structure prediction. *Journal of Molecular Biology*, 225:1049–1063.