Experimental methods for robot learning

Jeremy Wyatt,
School of Computer Science,
University of Birmingham, U.K.
J.L.Wyatt@cs.bham.ac.uk

John Hoar
Department of Artificial Intelligence,
University of Edinburgh, U.K.
johnho@dai.ed.ac.uk

Gillian Hayes
Department of Artificial Intelligence,
University of Edinburgh, U.K.
gmh@dai.ed.ac.uk

5/9/97

Abstract

In this paper we outline some ideas as to how robot learning experiments might best be designed, based on practical experience. The principal finding arises from the fact that an observed robot behaviour will typically have several possible causes. In order to isolate the source of an error or result it is therefore necessary to employ several evaluation methods together. In particular we argue for a division of the evaluation of robot learning along several lines. First, both external and internal measures of a robot's performance are required. Second, both quantitative and qualitative methods of description should ideally be employed. Four methods for agent analysis are then presented. These are drawn from a variety of fields: psychology, ethology, machine learning, statistics and engineering. We describe the application of these techniques to the construction and analysis of a robot that learns to push boxes from reinforcement. The paper concludes with general remarks about the efficacy of each of these techniques. We also emphasize the importance of careful experimental design in order to integrate the various forms of evaluation.

1 Introduction

It is difficult to define a coherent experimental method for robot learning. This is partly because an observed phenomenon may be caused in one of a number of ways. The robot's behaviour may be the product of the learning algorithm; the robot's initial knowledge; some property of the robot's sensors; the environment; or of an interaction between some subset of these. How do we design robot learning experiments so as to generate meaningful results in the face of such complexity?

The answer arises out of the simple fact that the cornerstone of any experimental method is the form of evaluation used. Suprisingly, until recently little work has been done on the problems of evaluation in intelligent robotics[2, 1]. Indeed some authors in the learning literature have eschewed the use of real robots because of the difficulties in quantifying robot performance [4]. In this paper we argue that the correct approach for experiments in robot learning is to employ multiple forms of evaluation. This is the only way to disambiguate the source of an error or behaviour when there are many separate causes of any given behaviour. Fortunately, because robot learning is closely related to a number of fields, it is possible to draw on their experience and evaluation methods. In this work we adapt methods for evaluation and experimental design from the disciplines of psychology, ethology, statistics and engineering. Four separate methods for analysis are presented here. It is not yet clear which particular subset of techniques will prove most appropriate for the evaluation of robot learning. The main finding of this paper is that they are, however, complementary: the weakness of one is the strength of another.

Proc. Conference "Towards Intelligent Mobile Robots", TIMR 97, Manchester 1997.
Technical Report Series, Department of Computer Science, Manchester University, ISSN 1361 - 6161.
Report number UMCS-97-9-1. http://www.cs.man.ac.uk/csonly/cstechrep/titles97.html

2 A typology of evaluation

There are several ways in which forms of evaluation can be categorized. The distinctions we draw are between *internal* and *external* evaluation methods; between *quantitative* and *qualitative* measures of performance; and also between the *exploratory* and *hypothesis testing* stages of experimentation. These distinctions are now discussed in turn.

Internal analysis concerns the robot's internal state. This can include the perceived state of the world; the action the robot is trying to perform; the robot's estimate of the utility of particular course of action; and the robot's estimate of how well it is currently doing with respect to its task. Each of these has an external equivalent: the actual state of the world; the action the robot is really performing; the utility of a given course of action apparent to a human observer; and that observer's estimate of how well the robot is actually doing. Comparing the two allows to us to identify discrepancies between the robot's view and the view of the human observer. The external observer's view is typically assumed to be correct and thus the discrepancy is described in terms of some deficiency in the robot's perception, actuators, or internal performance measure.

The second division is between quantitative and qualitative forms of evaluation. In order to generate an ordering on different robot controllers it is necessary to have some quantitative measure of performance. Such measures are task specific. In the case of internal measures they will also depend upon the sensory modalities used. Equivalent external measures, however, should be independent of the individual robot or sensory modalities employed. Only in such a manner can we can compare the performance of different algorithms on different robots. These functions are commonly referred to as reinforcement, error or evaluation functions. It is important to note that the design of such functions (which map from the domain of robot behaviours into the set of real numbers) is typically not a trivial task. A badly designed function may mislead us concerning the robot's true performance level.

Quantitative measures of performance may tell us when one method is better than another, or how far short of the optimum a behaviour falls, but they tell us very little about why a robot succeeded or failed. This shortcoming is the motivation behind explicit use of qualitative analysis. During the course of developing a robot learner, we typically make many informal observations and draw conclusions based on them. It is sensible to establish these, as far as possible, as a separate form of analysis, by formalising the process of recording qualitative observations. Qualitative analysis exhibits two important components lacking in the quantitative methods presented so far. First it requires a rich description of the robot's behaviour, incorporating knowledge not captured in numeric performance measures. Second it leads to the formation of hypotheses as to why the robot behaved as it did. These hypotheses guide the next round of experimentation.

The third distinction is drawn between exploratory experiments — typically used to form hypotheses to test — and experiments in which we test hypotheses. Normally only the latter type are reported. In fact it is just as useful to report the experiments which led to the formation of a hypothesis. In Section 4 experiments of both kinds are described. The exploratory experiments reported here were used to formulate a number of hypotheses to test. It should be clear that there are difficulties not only in designing appropriate evaluation methods, but also in the integration of different evaluation methods into one experimental framework. Sections 4 – 7 present four forms of evaluation. For three of the methods we describe their application to a specific robot learner. The robot, the learning task, and the learning algorithms employed are described in Section 3. The specific methods employed or suggested are: evaluation of sensors and sensory processing; analysis of reinforcement generated internally; quantitative external analysis of performance; and qualitative analysis of the robot's behaviour. In Section 8 additional remarks are made concerning the integration of these different evaluation techniques into a coherent experimental framework.

3 The agent, task and environment

This section outlines the learning task and the general architecture of the solutions evaluated. The robot used for these experiments, Asterix, was constructed from a Lego Technic kit (Figure 3 (left)). Asterix perceives the world by means of three kinds of sensors. Infra-red emitters/receivers are used to crudely assess the distance between it and nearby objects. A sprung bumper triggers microswitches when Asterix is in forwards contact with an object. A Hall effect sensor is used to encode wheel revolutions, and thus

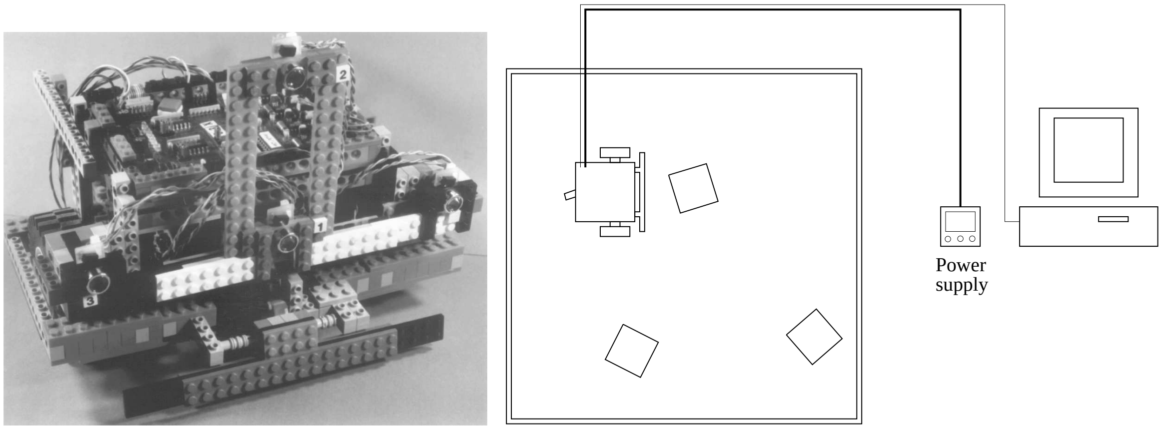

Figure 1: Asterix the robot (left) and the experimental setup (right).

to indicate when the robot is in motion. Each of these sensors possesses some subtle properties. These are formally analysed in Section 4.

The task was box pushing. The agent is placed within an arena ($2m^2$) with white walls (height $20cm$). The box pushing task is a reimplementation of an experimental setup used by Mahadevan and Connell [6]. It is worth noting that reimplementation in robotics has a particular purpose. Rather than than reproducing the same results using the same hardware, roboticists try to show how a technique generalises across slightly different environments and sensory modalities. By such means we can demonstrate the generality and robustness of the methods employed.

The box pushing task is divided into three subtasks: box finding, box pushing, and becoming unstuck. Mahadevan and Connell's hypothesis was that use of a behavioural decomposition simplified the learning task. The agent learned from reinforcement functions specified for each sub-task. They used Q-learning, in combination with two generalisation techniques. We reimplemented their work [3], using the statistical clustering method described in [6]; as well as the parameter values suggested. The only differences were that the learning rate for the Q-learner was initially set to be .95, declining according to an exponential law, until it reached a minimum value of .2; and that one-step rather than five-step Q-learning was employed.

The applicability functions and the arbitration network were retained as reported in [6]. The reinforcement functions initially employed were identical to those used in [6]. One of these, the reinforcement function for the box finding behaviour was altered, however, because the existing reinforcement function was found to be inadequate. The new reinforcement function is defined by:

$$r_{Finder_t} = \begin{cases} 3 & \text{if } a_t = \text{FW and BOX}(s_{t+1}) \\ 1 & \text{if } \neg\text{BUMP}(s_t) \text{ and BUMP}(s_{t+1}) \\ -1 & \text{if } \neg\text{BOX}(s_{t+1}) \\ 0 & \text{otherwise} \end{cases} \tag{1}$$

where FW specifies that the agent is travelling forward; s_t is the state at time t; and BOX(s) and BUMP(s) indicate that the robto is bumped against something and that it perceives a box.

The experimental setup used was slightly different from that reported in [6]. They used a robot with sonar as its primary sense; an infra-red sensor tuned to 4" as a bump sensor; and a motor current sensor to detect if the robot was stuck. In all, these sensors generated 2^{18} distinguishable states. The sensing on Asterix was similar, but rather more crude. In particular it is much harder to detect objects with the infra-red sensors, because their field of view is quite wide. Secondly, the Hall effect sensor provides slightly unreliable and delayed information about whether the robot is in motion. These changes in sensor modality can be shown to have a significant effect on the ability of the robot to learn, and on the reliability of its reward function.

As with [6] we carried out the learning computations using an off-board computer. We employed a physical tether, incorporating both a power feed and a data-line (see Figure 3 (right)). Although this

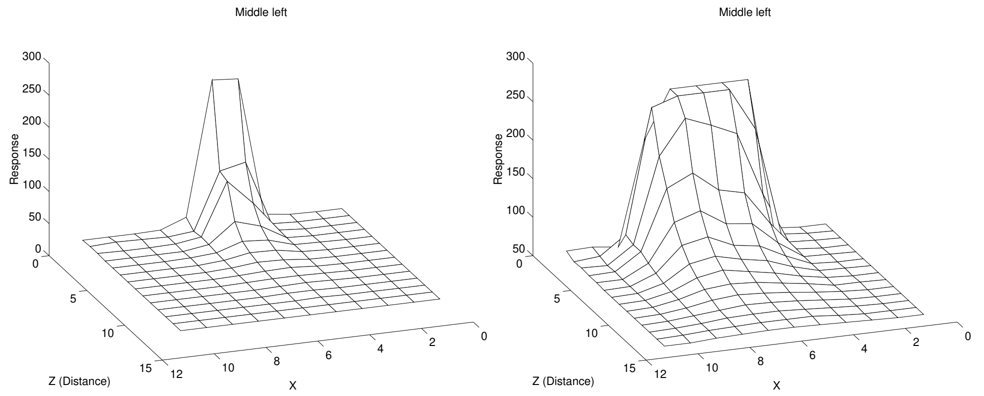

Figure 2: Response map of middle left IR in a passive array (left), and in an active array (right).

limits the autonomy of the robot, it extends the lifetime of the robot and allows the collection of large amounts of data. Finally it allows us to compare the perceived state of the world, with the actual state of the world; which in turn encourages explanations of why the robot behaved in a particular manner.

Finally in our setup we randomly placed both boxes and Asterix in the environment. Each 100 steps in a learning run we redistributed the boxes randomly. This contrasts with the approach taken in [6] where boxes and robot were given a fixed position, and the only random element was the orientation of the robot. We now turn to an account of the sensor evaluation we conducted.

4 Sensor evaluation

Whatever method is used for deriving a robot controller it is essential that the sensory features input to that controller contain sufficient information to distinguish all states relevant to the task in hand. If it can be guaranteed that all such states have been distinguished then an important condition for the convergence of a broad class of self-optimising methods has been satisfied. This class includes algorithms which learn from reinforcement. One way to try to provide such a guarantee is to carry out a detailed analysis of the sensors and sensory processing on the robot. This section presents just such an analysis.

In designing sensory processing functions we must have a clear understanding of what the sensory signals mean within the environment in which the agent is situated. Four experiments were carried out on Asterix to understand some of the properties of its infra-red sensors. Three of these experiments were *exploratory* in nature. On the basis of these experiments, possible sensory functions were designed. The final experiment evaluated their performance within the task environment.

The first experiment investigated the sensor response to boxes. Asterix has four infra-red sensors, which were initially arranged in a row. One of the boxes used for the box-pushing task was moved across a 10 by 12 grid of 5cm squares — in front of the robot — readings being taken from each IR at each position. In this experiment the face of the box was normal to the array of sensors. Each sensor's response was recorded under two circumstances: with only that sensor emitting (passive array — Figure 4 (left)); and with all sensors emitting (active array — Figure 4 (right)).

We can see two things from this. First, such sensors are not homogeneous; a reading may decay at different rates with respect to distance, and have an asymmetric response with respect to lateral translation. Furthermore it can be seen (Figure 4 (right)) that when all the IRs are emitting the sensors have a higher repsonse. This is because the field of view for each sensor is quite broad. Thus the effective field of view for each sensor is expanded by the reflected emissions from neighbouring sensors.

In the second experiment we investigated how the sensor response changed with respect to changes in orientation of a box. A box was placed at a distance of 15cm in front of the robot, with a face parallel to the front of the robot. It was rotated anti-clockwise through 90° in 5° increments. All sensors were emitting. Figure 3 shows that the response as the box rotates is roughly sinusoidal in shape, the minima roughly corresponding to an orientation of 45°, and the maxima to when the box is facing the sensor.

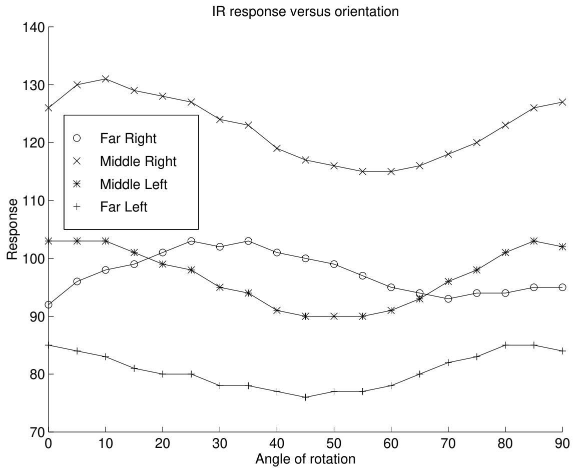

Figure 3: Response from IRs as orientation changes.

Since the sensors are in an array, these orientations are different for each sensor, and thus the responses have different phase. From this we can conclude that the orientation of the box affects the sensor response, and that the lateral position of the box with respect to the sensor modifies this behaviour.

In the third experiment we examined the sensor response to walls. The robot was placed at varying distances and orientations from the arena wall. The distance started at 10cm and finished at 130cm, measurements being taken at 10cm increments. The angle between the robot's face and the box was varied in 15° increments from 0° (parallel) through 75°. The sensor response was measured at each point. It transpires that the response saturates the sensors when the robot is less than 40cm from the wall, the orientation being insignificant. The consequence is that the robot is blind when close to the wall. From our previous experiments we concluded that the robot would not be able to see a box any further than 40cm away. It should now be clear that the sensory capabilities of the robot are rather limited.

Using the insights gained from these exploratory experiments we designed three possible filters to detect a box more or less in front of the robot. Two of these filters employ the original arrangement of four IRs in a row; the third filter uses an upside down T-shape. The first filter is the conjunctive filter. If the inner two IRs detect something but the outer two do not, then the object is narrow and likely to be a box. The difference of averages filter takes the difference between the averages of the two inner and two outer IRs. If this is greater than some threshold then a box is indicated. The T-shape filter thresholds the infra-red sensors individually, as for the conjunctive filter. If any of the lower infra-red sensors detect something, but the top one does not, then a box is signalled.

To compare these filters we conducted a statistical analysis based upon their observed identification rates. This allowed us to determine the absolute utility of each filter, and thereby identify the most successful filter. Thus this fourth experiment constitutes a testof any hypothesis that a particular filter is best[1]. Such an analysis is central to our method, because it allows us to state the extent to which relevant states can be distinguished and thus that to which the learning algorithm's convergence criteria hold. To test each filter we placed the robot in the arena. On thirty occasions it was placed far from the wall, and on thirty occasions close to the wall. On half of each of the observations in each position a box was placed in a random orientation and position, in front of the robot. When the robot was close to a wall, the robot's position and orientation were also slightly randomised. Each time we noted whether the filter's output indicated the presence of a box or not.

Table 1 gives the Pearson product-moment correlation coefficient for each filter. This gives a measure of the rate of a filter's rate of success. A value of -1 indicates that the filter is likely to be active when a box is absent, $+1$ indicates that the filter is likely to be active in the presence of a box. A value of 0

[1] Although the tests are not hypothesis tests in the sense of providing significant differences they do provide an ordering on the filters based on a reasonable sample size.

indicates no correlation. As can be seen, each of the filters has a moderate rate of success when the robot is far from the wall. Performance near the wall is much poorer. On the basis of these results we chose the filter that functioned best away from the wall: the T-shaped filter.

Correlation coefficient		
	Open Space	Near Wall
Conjunctive	0.6030	0.0673
Difference of Averages	0.6030	0.2673
T-shape	0.6804	0.1857

Table 1: Correlation between positive filter response and the presence of a box, measured using the Pearson product-moment correlation coefficient

5 Evaluation by internal reward

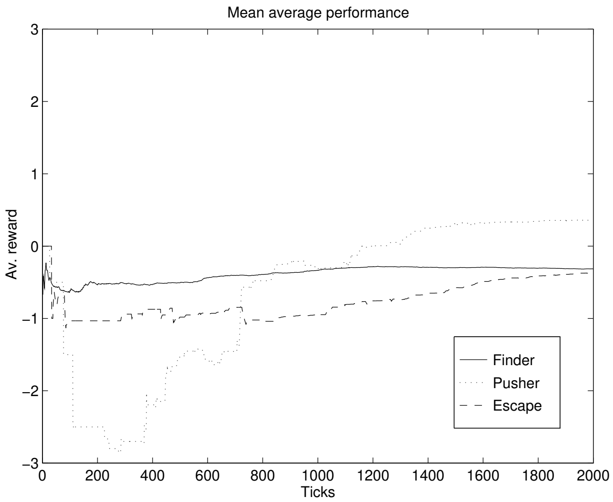

Figure 4: The mean average rewards for behaviours over 3 runs. The extents of the y-axis are set to the minimum and maximum reward possible

Having investigated the sensors, and developed appropriate filters based on our investigations we constructed the learning controller and ran learning experiments as described in Section 3. At each time step we recorded the average performance of the robot in the trial so far according to the internal reward functions. Figure 4 shows the average of this over three runs for the Q-learning agent described previously. It can be clearly seen that each behaviour improves during the learning period. The box pushing and escape behaviours improve considerably with respect to the reward function defined, and the box finding behaviour shows a small improvement. These results should not be surprising given the unreliability of the box filters described in Section 4.

These are the standard performance measures employed in the Machine Learning community for this class of problem. As previously mentioned, however, the accuracy of an internal measure depends upon reliable perception of the environment. If the agent's perception is unreliable this performance measure can be misleading. From our observations of the robot's actual behaviour it became clear that there was in fact a discrepancy between the number of boxes the agent believed it had pushed and the number of boxes it had actually pushed. We now turn to the problem of identifying the source of such an error.

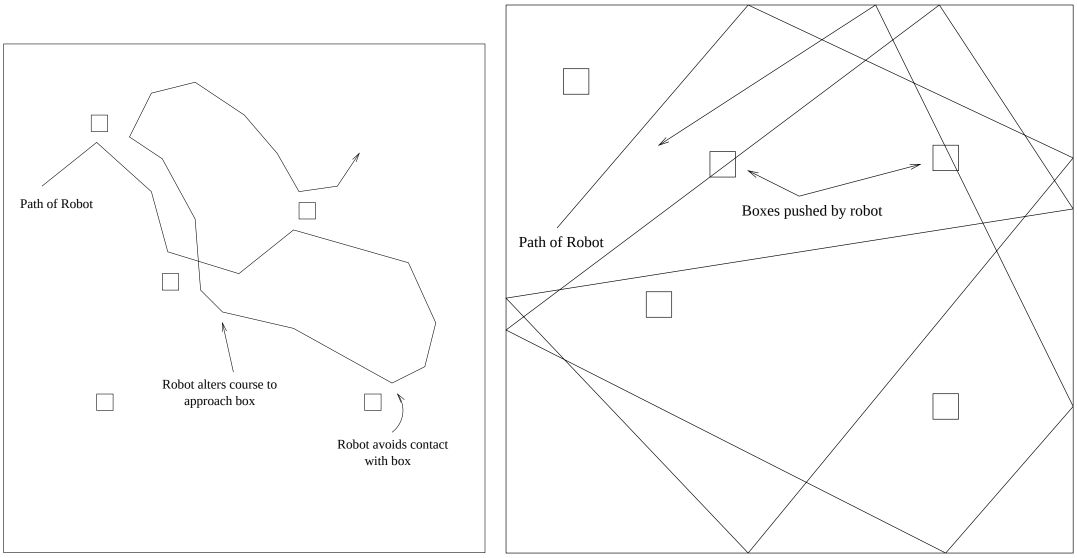

Figure 5: The box avoider (left) and the wall pusher (right)

6 Qualitative analysis

Having realised from our observations of the robot's actual behaviour that the robot was not in fact performing in the manner suggested by the internal reward function we were prompted to carry out a qualitative analysis of the form described in Section 2. It was discovered that the agent converges to one of two overall behaviours. We refer to these as the "box avoider" and the "wall pusher" respectively (see Figure 5).

In the box avoiding behaviour the agent does actually recognise boxes. It then generates reward by moving toward them. Rather than push the box the robot then turns away. We hypothesise that this is because the reward function for the finder (Equation 1) weights seeing boxes more heavily than bumping into them. Thus the agent generates more finder rewards by wandering through the environment and staring at boxes than by bumping into them; although when a box is bumped the box pusher module takes over and the box will be pushed. This is an example of an inappropriate reinforcement function. Designing appropriate reward functions is notoriously hard. Hence this is a phenomenon that we should expect to see frequently during the development of a robot learner.

In the wall pushing behaviour we found that the agent traversed the environment in straight lines, bumping into walls, pushing them, and then turning away. By comparing the robot's internal state to this behaviour, we found that the box pushing module became active and was rewarded when the robot bumped against a wall. This is because the Hall effect sensor — unlike the motor current sensor used in [6] — has a delay in reporting the cessation of motion. In addition it is unreliable. Because of this the agent is rewarded for pushing walls for a short period prior to the unwedging behaviour being activated. This is an example of subtle changes in the sensors used making perception unreliable, and a reasonable reward function misleading.

7 Quantitative analysis

Although we reported that the improvement in the robot's behaviour as indicated by the internal reward function was different from the actual behaviour that we observed we did not discuss in detail methods for detecting such discrepancies. In the case of our experiments this discrepancy became apparent by informal observation. In general a more rigorous method is required to detect such discrepancies. In this section we discuss the requirements for such a method and possible techniques for satisfying them.

When we design a reinforcement function from which a robot should learn we carry out two transformations on the minimal definition of optimal behaviour. First we need to correlate the conditions that define success with the robot's sensory states. As we have seen, this can lead to problems if the perception of the robot is inadequate. Second, we typically provide a more detailed performance measure than is strictly necessary in the hope that this will speed convergence. When carrying out external analysis, we can use a reward function devoid of these additions. In the case of the box pushing task we can therefore score the robot on its overall behaviour rather than for each sub-task — as is done by the decomposed learner. Equally the human observer is able to tell when the robot is actually pushing a box. A simple performance measure for the box-pushing task is to time the total period of the run during which the robot was pushing a box[2].

But the impartial evaluation of performance, especially by a experimenter with an a priori expectation of outcome, is difficult to achieve. It is thus necessary to introduce blind evaluation to robot learning. How can we separate the operational and observational roles in our experimental work? The route we propose taking is to separate the observer and operator. The observer is ideally someone with minimal knowledge of the aims of the project, i.e. the observer is blind to the hypothesised outcome, and to the different treatments employed. To physically separate the observer and the operator we suggest employing a video based evaluation method. At the end of each run the agent's policy is fixed, and the agent is run for a period of time with that fixed policy. This is carried out for each agent, learning or non-learning. Each run is recorded and then viewed 'blind' by the observer, who marks the robot's performance on the whole task according to a marking scheme that corresponds as closely as possible to the objective behaviour desired.

8 An experimental framework

This paper has now outlined four different forms of evaluation. It has, however, not yet been made explicit how these can be tied together into a coherent experimental framework. The general structure of our actual experiments is apparent from the sequence given in the paper. The experiments described in this paper are, however, incomplete. In this section some more detailed remarks are made about the manner in which the complete process of devising, designing and executing robot learning experiments might be conducted.

One possible sequence of events for robot experiments is as follows. First the general hypothesis to be tested should be outlined. In our case it was to compare the performance of $Q(\lambda)$ learning with simple one-step Q-learning[3]. The hypothesis was that Q-learning with eligibility traces would learn faster than one-step Q-learning on a real robot task. The next stage is the choice of task. This is inextricably linked to the robot type, and to the available sensors and actuators. Currently there is no canonical set of robot learning bench-mark tasks, although this may change as the sensors and capabilities of off the shelf robots become increasingly standard. Once the task has been selected then the internal and external performance measures can be more precisely defined. As has already been stated the internal performance measure may depend on particular sensors, and thus may influence the choice of robot and environment.

The next stage involves the construction of the robot if necessary, and of the environment. Following this the properties of the sensors must be analysed. There are two relationships that must be studied. First the properties of the sensors with respect to a static environment should be assessed. Second, the behaviour of the sensors when the robot interacts with the environment might also need to be studied[4].

Following this preparatory work the actual comparison of a set of learning techniques can be carried out. Typically, in order to understand the properties of the robot and the environment a human-designed non-adaptive controller is also tested in order to check that the task can be performed using the robot. We did not carry out such an experiment and suggest that although using hand-designed controllers as experimental controls is a laudable idea, they should not be developed prior to the conduct of the experiments with the learning controllers themselves. This is for the reason that in the process of implementing the hand-designed controller a great deal is typically learned about the task and this can

[2] To the best of our knowledge Mahadevan and Connell did not carry out this type of analysis, on the assumption that the sensors on Obelix were accurate.

[3] The results of the experiments with the $Q(\lambda)$ learner are not described here, as the object is to focus on techniques for evaluation. For the results obtained see [3].

[4] We carried out no such experiments here, though a good example is given in [5].

seep into the design of the robot learner in various ways, e.g. in the choice of sensor modalities; the design of the sensory pre-processing; any decisions about the amount of state that should be available to the learning controller; or the design of the internal performance measure. All these are informed by a detailed understanding of the interaction between the robot, the environment and the task. If we are to reject the hypothesis that an adaptive is no more powerful than a handcoded controller then we can only do so if no unseen advantages have been given to the learning method. This need to aviod an order effect arises from the fact that our alternative hypothesis will typically be that the learning controller is better. Additionally the process of understanding the robot's interaction with its environment needs to be formalised as far as possible, and the findings reported. The implementation of a non-learning controller can assist in this regard. To these ends we suggest that it is sensible to implement any non-adaptive controller after the experiments with the learning controllers have been conducted.

The order in which the various quantitative evaluation methods are conducted may not matter, given that the observer for external evaluation should be separated from the robot operator. Clearly at each stage sufficient data should be collected in order that statistically significant results can be obtained wherever possible[5]. If discrepancies are discovered between the internal and external performance measures then the source of the discrepancy must be identified. We suggest that the cause may be one of three. Either the internal performance measure is inappropriate; the sensory information available to the robot is inadequate; or additional knowledge used to constrain the learning task may be inappropriate. This paper has given examples of the first two. Examples of the third would be if the decomposition of the controller into a hierarchy of simpler controllers was inappropriate and prevented learning; or if the learning controller's experience was biased using a teaching method, and the bias thus induced prevented the robot subsequently converging to the optimal behaviour. This is akin to the problem of choosing an appropriate training set in supervised learning.

Finally following the detection of any such discrepancy the experiments must be reconducted once the source of the error has been identified and eliminated. In order to eliminate the discrepancy between the internal and external performance measures hypotheses must be formed as to the source, and these hypotheses tested in turn. The method we used to form hypotheses concerning the source of the error was based on a qualitative analysis of the robot's behaviour. Thus the experimental process may have to undergo several iterations of adjusting the sensory processing and the internal performance measure until the learning techniques can reasonably be compared. A slightly different technique would obviously be required if the learning algorithm itself were designed to select the sensory features required to learn successfully. It is also worth noting that there is no clear division in this framework between exploratory and hypothesis testing experiments. It is assumed that the specific sequence depends largely on the domain under investigation.

9 Conclusion

The principal findings of this paper are as follows. First, internal estimates of performance are inadequate if there is any likelihood that the sensors upon which the reward function depends are unreliable. Second, appropriate internal reinforcement functions are hard to design, and performance can never be evaluated solely by these. Third, quantitative external analysis according to a similar performance metric can be used to detect discrepancies between the actual performance and the performance as predicted by the robot. Fourth, qualitative analysis is necessary in order to provide a rich enough description of the robot's behaviour to generate explanations of why the robot behaved in the way it did. These hypotheses can be used to modify the agent's sensors, sensory processing, or reward function.

From these findings we can draw the further conclusion that when using learning methods as part of the design process of a robot controller it is necessary to utilize several forms of evaluation as part of a coherent experimental framework. The evaluation and analysis tools we have described assist us by providing a simple and reliable means to isolate sources of error. In addition some of the problems in designing an appropriate experimental framework have been discussed.

Acknowledgements
Some of this work was carried out while John Hoar and Jeremy Wyatt were in receipt of EPSRC Research

[5]Though given the current unreliability of robots, and the consequent difficulties of conducting robot experiments, this is hard to achieve, and is a goal of which the authors have fallen short.

Studentships. Thanks to John Hallam, Andrew Haston, Tom Alexander and Douglas Howie.

References

[1] Yassine Faihe and Jean-Pierre Muller. Analysis and design of a robot's behaviour: Towards a methodology. In John Demiris Andreas Birk, editor, *Proceedings of Sixth European Workshop on Learning Robots*, 1997.

[2] Henry Hexmoor. Robolearn 97: An international workshop on evaluating robot learning. Technical report TR 97-03, Department of Computer Science, State University of New York at Buffalo, April 1997.

[3] J. Hoar. *Reinforcement learning applied to a real robot task.* Unpublished masters thesis, University of Edinburgh, Department of Artificial Intelligence, September 1996.

[4] Leslie Pack Kaelbling. *Learning in Embedded Systems.* PhD thesis, Dept of Computer Science, Stanford, 1990.

[5] H.H. Lund and O. Miglino. From simulated to real robots. In *Proceedings of IEEE 3rd International Conference on Evolutionary Computation*, IEEE Press, 1996.

[6] S. Mahadevan and J.H. Connell. *Automatic programming of behaviour-based robots using reinforcement learning.* Research Report RC 16359 (72625), IBM Research Division, July 1990.