

# An Environment for Robot Learning

R. Andrew Russell

Intelligent Robotics Research Centre, Monash University, AUSTRALIA  
andy.russell@eng.monash.edu.au

## Abstract

This paper describes the progress of a project investigating development of unsupervised robotic learning at a most basic level. The project focuses on the transition between an organism whose genetically evolved competence is purely inherited and one with the added ability to learn from its environment. The framework of the project draws on Rolf Pfeifer's ideas about building complete autonomous systems that he calls "Fungus Eaters". For this project a self-contained environment has been constructed to act as an ecological niche for a mobile robot. It is anticipated that the robot will be able to improve its performance by learning from its interactions with this environment. This paper describes the mobile robot, the environment, and then outlines progress towards developing a method that will allow the robot to formulate a 'value system' that it can use to evaluate the benefit of its actions. All results presented in this paper were obtained by performing experiments using embodied robot systems.

## 1 Introduction

Robot learning is a very appealing area of research that has a number of potential benefits. A robot with the ability to learn would require less application-specific programming to customise it for performing a particular operation. If the environment changed then a learning robot may also be able to adapt appropriately without external guidance. In addition, the ability to learn is one of the key features of intelligence. Studying the creation of learning behaviour in robots may provide some insight into intelligent behaviour in general.

Ever since the creation of the first digital computers there has been an interest in using them to investigate machine learning [Turing, 1950]. As an application of machine learning the development of learning abilities in robotic systems adds extra challenges involving real-time constraints and sensing limitations. A good overview of the current state of robot learning research is given by Koren and Zelinski (Ward and Zelinski, 2000). From the animal behaviour perspective Scott [Scott, 1972] proposes the following definition of learning:

“The most general definition of learning is therefore the modification of behaviour by previous experience.”

Scott's definition matches well with the focus of this project. It implies that a basic level of competence already exists and that learning is the modification of this pre-existing competence. Evolutionary processes seem adequate for explaining how a large group of creatures could develop competence in dealing with their environment. However, at some point it must have become more beneficial for each individual to gain the ability to learn from its surroundings [Walter, 1963]. To investigate the acquisition of learning ability the robot ADAM and environment EDEN were developed. This robot/environment system is broadly based on Pfeifer's "Fungus Eaters" [Pfeifer, 1996]. Fungus Eaters are envisaged to be complete autonomous devices that have to deal with a number of interrelated and competing demands. In the examples given by Pfeifer these demands include collecting uranium ore, harvesting fungus that acts as an energy source and defending themselves against predators. In this project, to be successful the real world autonomous agent ADAM Robot has to balance interrelated constraints imposed by its ecological niche in the environment EDEN. In order to maintain strict control

of the learning process only the minimum of information was made available to the learning system. Sensor readings and the actuation levels of all of the robot's actuators were accessible to the learning system as well as the reactive nature of the underlying controller. The learning system was not given an evaluation function to assess the desirability of its the actions (no in built value principle). The ultimate aim of the project is to develop methods of unsupervised learning using observations of the performance of the robot's innate controller. In related work Noel Sharkey [Sharkey, 1998] has used observations of the performance of a robot with a hand coded innate behaviour to train a neural network controller. Michaud and Mataric [Michaud and Mataric, 1998] used observations of the performance of a mobile robot to select behaviour producing modules. In this paper, as well as describing the construction and function of ADAM Robot and EDEN, preliminary experiments are described where the robot uses observations of its own interactions with the environment to develop an evaluation function.

## 2 The robot environment EDEN

EDEN (an EDucational ENvironment) was built to provide a self-contained and structured environment for performing robot learning experiments. This 1.5m by 1.5m enclosure provides a ecological niche for a mobile robot.

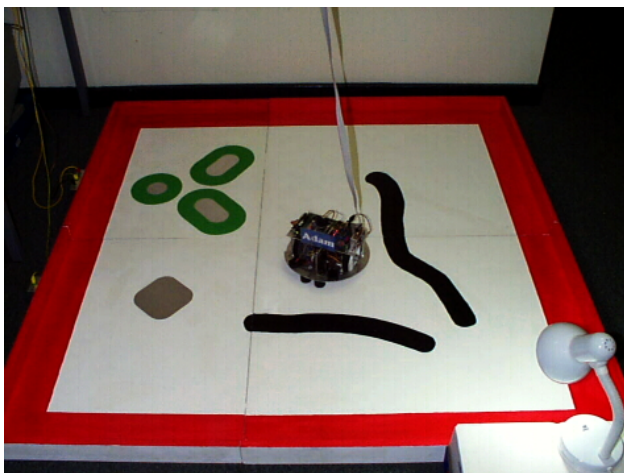


Figure 1 A photograph of ADAM Robot in its world EDEN.

Within EDEN there are four 'flowers' that the robot can feed from to gain energy, and an illuminated corner of EDEN provides a place for the robot to bask and reduce its energy consumption. Collision with the perimeter

wall causes the robot to lose energy. Green colour markings on the floor indicate proximity to three of the four flowers and a red border denotes the perimeter wall. Black floor markings pass between the flowers and the basking area. A photograph of EDEN is shown in Figure 1. This world was designed to provide motivation for the robot to move about and explore. There are also sufficient sensory cues for the robot to improve its performance in terms of maintaining an adequate energy level and energy usage. ADAM Robot was designed to be able to interact effectively with EDEN.

## 3 A robot to explore EDEN

Physically ADAM Robot (the ADaptive Mobile Robot) has a circular chassis 24cm in diameter with two side-by-side driven wheels and balanced by a plastic roller to provide a third point of contact with the floor. The robot can turn on-the-spot or move straight backwards or forwards. ADAM is only provided with local sensors and cannot perceive the whole of EDEN at one time. From ADAM's point-of-view EDEN is partially observable. Colour coded areas on the floor provide clues about the robot's proximity to walls and flowers. Twin floor colour sensors can discriminate white, red, green, and black areas. A schematic diagram showing the major subsystems in ADAM and EDEN is given in Figure 2.

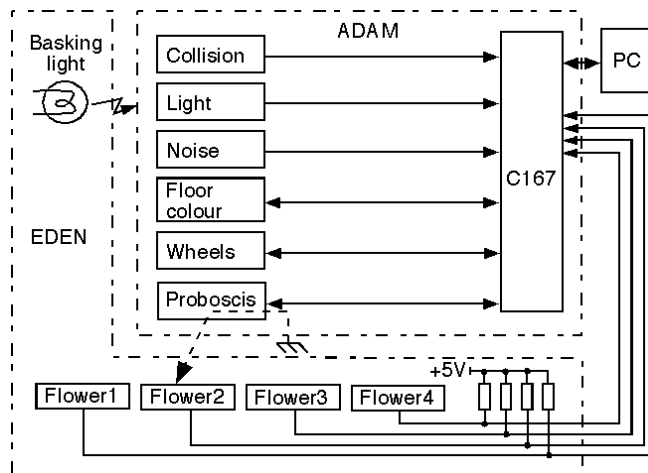


Figure 2 A schematic diagram of the ADAM/EDEN system.

To measure floor colour each sensor is equipped with a red and a green LED together with a phototransistor. The phototransistor response is recorded when the floor is illuminated with red and green light. Figure 3 shows how the relative response during the two measurements can be

used to discriminate floor colour. Note that a fall in the numbers plotted in Figure 3 corresponds to an increased response.

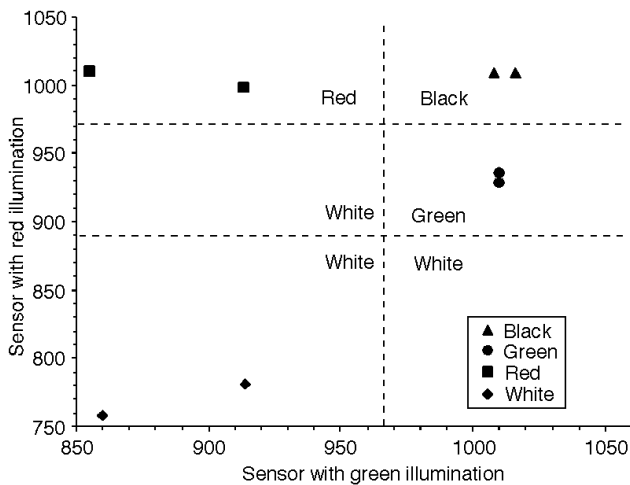


Figure 3 The response of the floor colour sensor to different colour markings.

Collisions are detected by left and right forward facing bumpers. However, wheel rotation is monitored during every robot movement (including rotation) and anything that prevents the wheels rotating is reported as a wheel jam. Light falling vertically onto the robot is measured by a photoresistor and this information is used to detect the illuminated basking area. Loud noises are registered by a microphone and can act as a neutral stimulus during learning experiments. As well as being able to move ADAM has a servo controlled proboscis that it uses to feed from the flowers in EDEN. Each of the four flowers has its own electrical circuit that allows the identity of the flower to be determined during feeding. ADAM is provided with a basic level of competence for surviving in EDEN. The following production rules [Charniak and McDermott, 1985] describe the reactive controller that defines ADAM's inbuilt behaviour. Conflicts between the rules are resolved by giving precedence in the same order that the rules are listed below.

```

if both bumpers activated then back off and turn
                                clockwise
if left bumper activated then back off and turn
                                clockwise
if right bumper activated then back off and turn anti-
                                clockwise
if wheels jam then back off and turn
                                clockwise

```

```

if both floor colour sensors see green then feed and
                                                move forward
if light is bright and energy level is high then bask (stop)
otherwise move forward 3cm.

```

These rules allow ADAM to feed from the visible flowers (those outlined in green), bask in the illuminated area and to explore EDEN. However, EDEN provides many possible ways for ADAM to improve its performance.

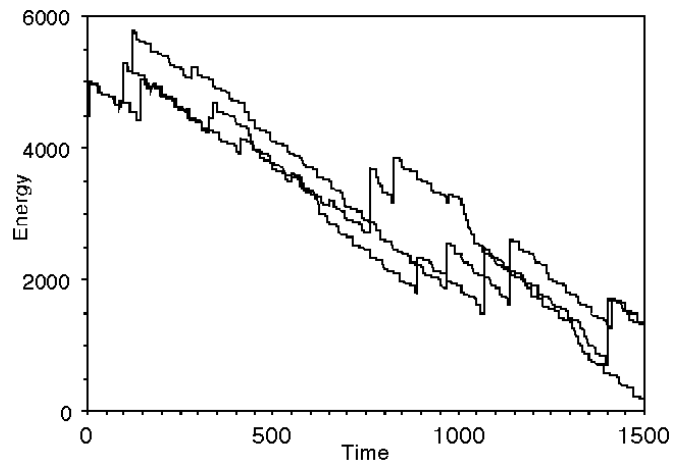


Figure 4 Results of three trials showing the robot energy level as it explores EDEN.

Figure 4 shows graphs of robot energy plotted against time step. The results of three trials are plotted, each taking 1500 time steps. ADAM starts the trials with 4000 units of energy and gains energy by feeding from the flowers. Each flower adds to its store of energy at the rate of 1 unit for each time step and feeding transfers the accumulated energy to the robot. Energy is consumed by ADAM at the rate of 2 units per time step that reduces to 1 unit in the basking area. Therefore, the strategy of continually feeding from a single flower incurs a steady energy loss of 1 energy unit per time step. Collisions with the walls also incur a loss of 100 units of energy. Although ADAM is able to negotiate its way around EDEN there is a steady loss of energy and therefore plenty of opportunity for improvement.

#### 4 Extracting a value system by self-observation.

Pfeifer [Pfeifer, 1996] claimed that for an autonomous robot to include a self-supervised learning mechanism it must have a means to judge what is good for it and what isn't. In this project the information available to the robot has been limited to the robot's sensory inputs and the state

of its actuators gathered over time. There is also an assumption that the robot's inbuilt controller is reactive [Gat, et al., 1994]. From this information the learning system must extract a 'value' system that indicates whether or not the robot has benefited from a particular action. The value system is expected to take the form of a change in the robot's state that is consistently associated with an improvement in the robot's situation. The challenge is to identify the state information that can be used to evaluate the performance of the robot.

#### 4.1 The robot state

At any instant in time the state of the robot's sensors and actuators constitutes the robot's state. Let  $S$  be the set of all possible robot sensor readings for all available sensors. The elements of the set  $S$  are of the form:

$s_j$   $s$  is the sensor {left colour, right colour, bump, illumination, sound}  
 $j$  is the sensor reading.

Sensor reading is selected from the following sets:

left colour	= {black, green, red, white}
right colour	= {black, green, red, white}
bump sensor	= {left, right, both, jam, none}
illumination	= {high, low}
sound	= {silent, noise}
energy level	= {high, low}

Let  $A$  be the set of all possible activation states for the available actuators. The elements of the set  $A$  are of the form:

$a_k$   $a$  is the actuator {wheels, proboscis}  
 $k$  is the activation state of the actuator.

Actuator activation state is selected from the following sets:

Wheels	= {forward, backward, turn clockwise, turn anti-clockwise, stop}
proboscis	= {feeding, not feeding}

For the robot time proceeds in discrete steps or instants. At any particular instant  $i$  the state of the robot  $R_i$  is an ordered pair:

$$R_i = (A_i, S_i)$$

where:

$A_i$  is a set of activation states of the robot's actuators at time instant  $i$   $A_i \subset A$

$S_i$  is a set of readings from the robot's sensors at time instant  $i$   $S_i \subset S$

The product set  $A \times S$  contains all possible robot states.

#### 4.2 Changes of actuator activation level

As noted previously ADAM's underlying control is reactive using little stored state information. State information is only stored when a sequence of actuator changes is triggered (eg. forward to backward to turn clockwise to forward - that would result from the left bumper being actuated). This implies that actuator outputs are directly related to the current sensor input unless a sequence of adjacent actuator changes has been triggered. In the steady state where the inputs and outputs do not change it is not possible to infer which of the sensor readings are responsible for the current actions of the robot. However, after a stable period, a change of actuator activation level  $\Delta A_n$  will be preceded by a change of sensory input  $\Delta S_{n-1}$  and at least one of these sensor changes must be responsible for the actuator change.

$$\Delta A_n \rightarrow \Delta S_{n-1}$$

where:

$$\Delta A_i = (A_{i-1}, A_i) \quad A_{i-1} \neq A_i$$

$$\Delta S_i = (S_{i-1}, S_i) \quad S_{i-1} \neq S_i$$

In order to distinguish actuator changes that result from sensor changes from actuator changes that are part of a sequence, only changes that occur at time step  $k$  after a period of  $n$  time steps with no state change will be considered.

$$\bigcup_{j=1}^{n-1} (A_{k-j} - A_{k-j-1}) \cup (S_{k-j} - S_{k-j-1}) = \emptyset$$

In order to identify a value system for the robot the following underlying assumptions are made:

- Each action selected by the robot's inbuilt controller benefits the robot (only a pathological organism would act to its own detriment).

- The benefit will be immediate (the possibility of delayed benefit will not be considered in the current investigation).
- The benefit conferred on the robot will cause some identifiable change in the robot state (the robot can detect the benefit).
- A similar state change (benefit) will be observed for a number of different actuator changes.

The identification procedure will involve recording the state changes that occur as ADAM moves around in EDEN. Any change in sensor reading that occurs regularly as a result of a number of different changes in actuator activation level will be identified as part of the robot's value system. Whenever this change in sensor reading is observed it will be assumed that the robot has benefited.

## 5 An experiment to identify a value system

For this experiment the actuator activation levels and resulting sensor readings were recorded for 1000 time steps as ADAM negotiated its way around EDEN. From the data were isolated changes in actuator activation level that occurred after at least 4 time steps with no change. After at least four time steps without sensor or actuator change there are four possible actuator changes that could occur -

- Forward to Backward (after a collision),
- Forward to Eat (when finding a flower),
- Forward to Bask (when entering the illuminated area) and
- Bask to Forward (when ADAM's energy supplies fall too low to remain basking).

Figure 5 shows the number of times a particular sensor output changes either immediately before or immediately after the actuator change. It is seen that sensor changes are specific to only one actuator change and therefore do not fulfil the criteria of being associated with a number of different actuator changes. Therefore, none of the discrete sensor readings fulfils the requirements for being part of the robot's value system. ADAM's state includes readings from two sensors that give a numerical value rather than a discrete result. The critical factor is how the change in sensor value is affected by the actuator change. This is indicated by the rate of change of sensor value  $\Delta^2 S_i$ :

$$\Delta^2 S_i = (S_{i-1} - S_{i-2}) - (S_i - S_{i-1})$$

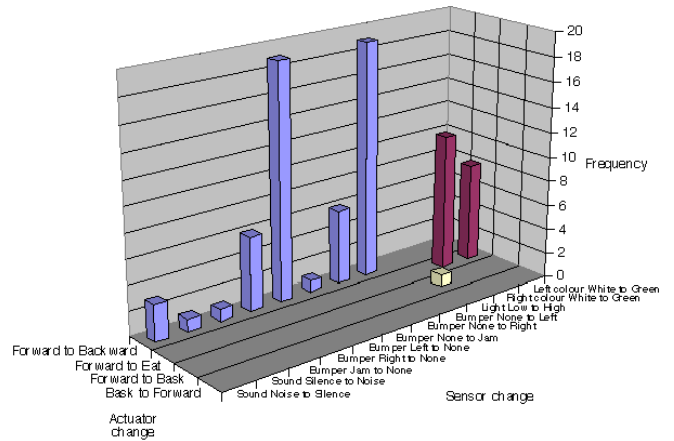


Figure 5 The frequency with which each sensor change is associated with a particular actuator change.

Figures 6 (light level) and 7 (energy level) show the number of times that the rate of change of the sensor value is positive (the positive bar) and negative (the negative bar) for each actuator change.

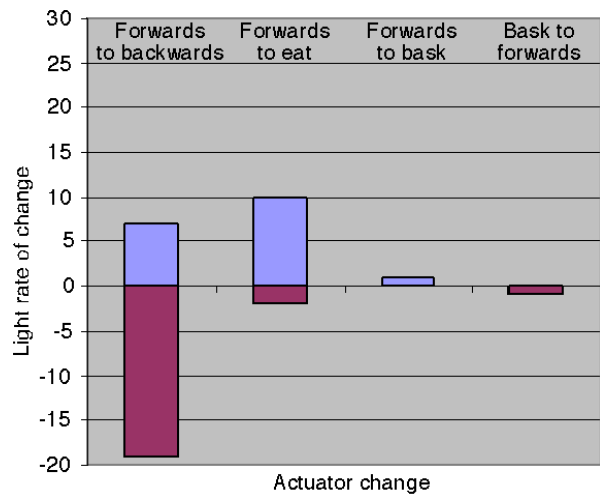


Figure 6 The frequency with which positive and negative rates of change in light intensity (negative total is the number of negative rates of change that occur) are associated with particular actuator changes.

It is seen that variations in the rate of change of light intensity occur for all of the actuator changes. However, both positive and negative values are present and therefore light intensity does not give a consistent indication. For energy there is a consistent positive rate of change associated with three of the four actuator changes and therefore energy would be a candidate for ADAM's value system.

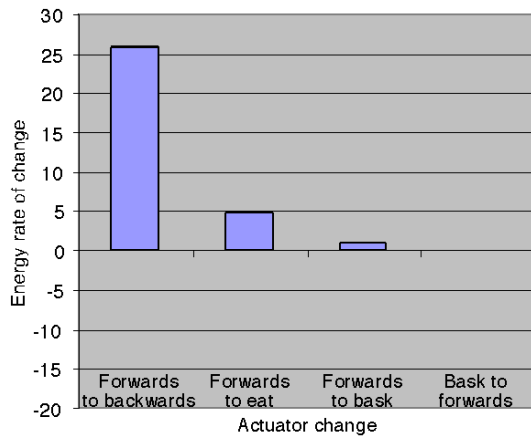


Figure 7 The frequency with which positive rates of change in energy are associated with particular actuator changes (there are no negative rates of change).

## 6 Conclusions

This paper has presented the preliminary stage of an investigation into robot learning. To provide a constrained experimental area for learning experiments a self-contained environment called EDEN has been constructed. EDEN was designed so that a robot would need to move around and avoid obstacles in order to function efficiently. ADAM Robot was provided with appropriate sensors and actuators to allow it to fit into the ecological niche provided by EDEN. The reactive controller built into ADAM gives it a basic competence to function in its environment. However, there is plenty of scope for it to learn and improve its performance.

Applying the principle of Occam's razor it was decided that any learning scheme should be given the bare minimum of information to work with. Indirect sources of information such as evaluation functions that could predispose the learning system to choose a particular solution were also ruled out. ADAM's learning scheme will have access to two pieces of information. The first is the time evolution of ADAM's sensor readings and actuator actuation levels as it negotiates its way around EDEN. The second is the fact that ADAM's underlying controller is reactive. Removing either of these pieces of information would seem to leave a situation where no learning would be possible.

A preliminary investigation seems to show that ADAM can develop a value system for evaluating the benefit of its actions using the limited information available. Future investigations will consider techniques

for improving ADAM's performance based on learning from self-observation.

## Acknowledgments

The author would like to acknowledge Infineon for providing the C167 microcontroller development board used in ADAM.

## References

[Charniak and McDermott, 1985] Eugene Charniak, and Drew McDermott, *Introduction to Artificial Intelligence*, Addison-Wesley, Reading Mass., 1985.

[Gat, et al., 1994] Erann Gat, et al. Behavior control for robotic exploration of planetary surfaces, *IEEE Transactions on Robotics and Automation*, Vol.10, No. 4, pp. 490-503, 1994.

[Michaud and Mataric, 1998] François Michaud and Maja J. Mataric, Learning from history for behavior-based mobile robots in non-stationary conditions, *Autonomous Robots*, Vol. 5, Nos 3/4, pp. 335-354, 1998.

[Pfeifer, 1996] Rolf Pfeifer, Building "fungus eaters"; design principles of autonomous agents, *Proceedings of the Fourth International Conference on Simulation of Adaptive Behavior*, The MIT Press, Cambridge, Mass, pp. 3-12, 1996.

[Scott, 1972] John P. Scott., *Animal Behavior, Second Edition*, The University of Chicago Press, Chicago, 1972.

[Sharkey, 1998] Noel E. Sharkey, Learning from innate behaviours: a quantitative evaluation of neural network controllers, *Autonomous Robots*, Vol. 5, pp. 317-334, 1998.

[Turing 1950] Alan M. Turing Computing machines and intelligence, *Mind*, Vol. 49, No. 236, pp.433-460, 1950.

[Walter, 1963] W. Grey Walter, W., *The Living Brain*, Penguin Books Ltd.,1963.

[Ward and Zelinski, 2000] Koren Ward, and Alex Zelinski, Acquiring mobile robot behaviours by learning trajectory velocities, *Autonomous Robots*, Vol.9, pp. 113-133, 2000.