

# Can Learning take us past Programs to Consciousness?

**John H Andreae**

Department of Electrical & Computer Engineering  
University of Canterbury, Christchurch, New Zealand  
email: [andreae@elec.canterbury.ac.nz](mailto:andreae@elec.canterbury.ac.nz)

[This is the text of a previously unpublished, invited address delivered at the Fifth Biannual Conference on Artificial Neural Networks and Expert Systems, ANNES'2001, held at the University of Otago, Dunedin, N.Z. 22-24 November 2001. A one page summary of the address appeared in the Proceedings of the Conference]

Keywords: Consciousness, Learning, Robot.  
Domains: Artificial intelligence, Cognition.

**Abstract:** A human-like, intelligent, conscious robot is the ultimate goal for many researchers and has been mine for 40 years. Artificial Intelligence (AI) and Connectionism seem to be getting nowhere, while Neuroscience forges ahead with consciousness in its sights. Here, I argue (again!) that AI and Connectionism must abandon top-level programs and encodingism to get back into the race for consciousness.

## Contents

1. Are Neuroscientists within reach of Consciousness?
2. Designing Learning Robots since 1961
3. No Top-Level Program
4. My Recipe
5. Rules
6. Encodingism
7. Squash-Pop Microworld
8. Mother and Infant
9. Results of the Experiment
10. Top-Level Program Still Lurking
11. A Plan
12. Visual Consciousness
13. Simple Example of Constructed Vision
14. Talking-to-Ourselves Consciousness
15. The Problem of Experience
16. Thought Experiment
17. Hard Problem Still Relevant
18. Appendix: Memory and Consciousness

References

## **1. Are Neuroscientists within reach of Consciousness?**

This year's volume 79 of the journal *Cognition* provided readers "with a perspective on the latest contributions of cognitive psychology, neuropsychology, and brain imaging to our understanding of consciousness." Reviewing the articles in the volume, the philosopher Daniel Dennett [1] says "As the Decade of the Brain (declared by President Bush in 1990) comes to a close, we are beginning to discern how the human brain achieves consciousness." He agrees with Dehaene and Nacacche [2], who wrote the article introducing the volume, in seeing "convergence coming from quite different quarters on a version of the *global neuronal workspace model*" of consciousness. It is clear that neuroscientists feel that they are within reach of explaining consciousness.

## **2. Designing Learning Robots since 1961**

My aim since 1961 has been to design a robot that could learn like a human. My first design, called STeLLA and published in 1963 [3], learned to find its way to goals, made plans and learned tasks such as steering a car on a cambered slippery road. My current design, called PURR-PUSS (PP for short), replaced STeLLA in 1972. STeLLA can be seen as a connectionist system as it was an extension of the Perceptron and depended on probabilistic or fuzzy weights. PP, on the other hand, has a strong computational central core and only the surrounding motivational system is probabilistic. I see PP as a half-way house between the techniques of Artificial Intelligence (AI) or Expert Systems on the one hand, and Connectionism or Neural Networks on the other.

I am not going to describe PP to you in any detail, but I will use some of its features to illustrate the points I wish to make about designing a conscious, learning robot.

## **3. No Top-Level Program**

1972 was an important year for me because of the birth of my system PP, but it was also a significant year for the field of Artificial Intelligence. I have in my library two books that were published in that year. One was Terry Winograd's [4] description of his famous SHRDLU system, perhaps the epitome of classical Artificial Intelligence. The other was the philosopher Hubert Dreyfus' [5] influential book "What Computers Can't Do" that attacked AI at its foundations. Dreyfus concluded that human intelligence could not be programmed and that intelligence required a body as well as a brain. It is ironic that Sir James Lighthill broke up the well-thought-of British AI team at Edinburgh with the gibe that they should be doing something useful like

SHRDLU. Just 14 years later Winograd [6] had moved over to Dreyfus' side, arguing in a book with Fernando Flores that "one cannot program computers to be intelligent and that we need to look in different directions for the design of powerful computer technology."

A similar message can be taken from John Searle's [7] Chinese Room thought experiment, in which a person in the room, who doesn't understand Chinese, is producing answers to questions in Chinese by following a fixed set of rules. The Chinese Room has been attacked from many angles, but it was intended to show that human language understanding cannot be carried out by a fixed set of rules, that is by a computer program.

Some would say that it is obvious, or commonsense, that no program could be written to account for all the creative potential of the human mind.

A way out is suggested by Roger Penrose in his popular books "The Emperor's New Mind" [8] and "Shadows of the Mind" [9]. Penrose, following Lucas [10] before him, used Gödel's famous theorem in mathematics to argue that human intelligence couldn't be an algorithm, that is a computer program. Penrose thinks that the answer lies in quantum mechanics, but we don't need to go to such extremes to find a computer that isn't driven by a program. My recipe may not be the only answer or the best, but it certainly shows an easier way out than quantum mechanics.

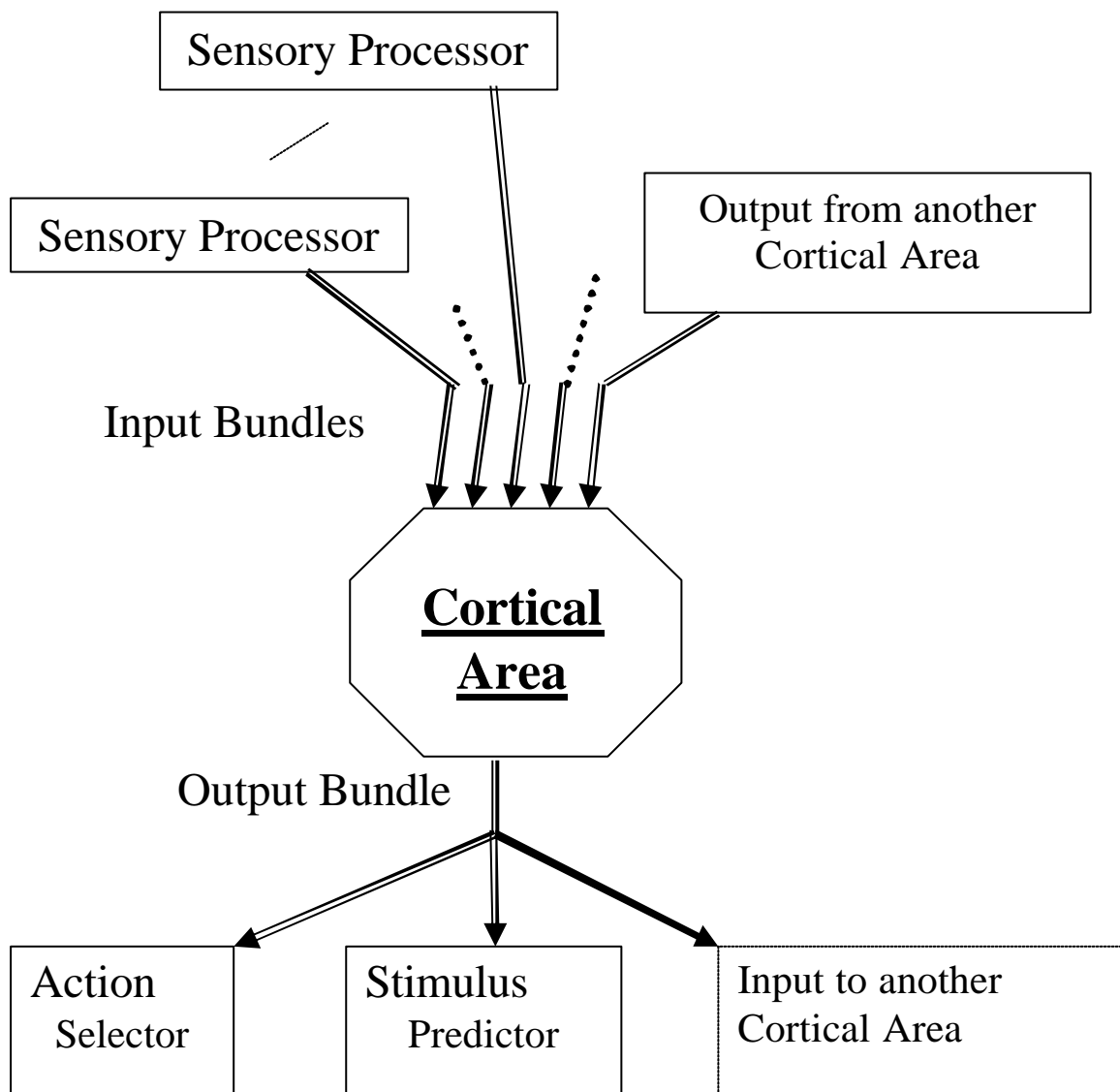
#### **4. My Recipe**

The first step of my recipe is to recognize that a computer brain, like a human brain, will have many programs. There are many innate (pre-programmed) processes in the human brain. For example, there is evidence that face recognition, the processing of emotions, parts of language processing and early stages of visual processing are among the innate processes. It doesn't matter if there are programs in the brain, as long as there is *no top-level program*. There must be no program, algorithm or fixed set of rules in the driving seat.

The second step of my recipe is to recognize that computers --- and brains --- contain both programs *and data*. If we don't have a program in the driving seat, then we must have data in the driving seat. Perhaps the data is a fixed list of rules that a low-level program is carrying out. No! That won't do. That is just a program again. We must ensure that the list of rules is changing all the time and also that the low-level program plus the changing list of rules are not equivalent to a fixed top-level program with variable data.

The final step of my recipe is to put the computer ‘brain’ in a robot body which can move about in the real world and to arrange for the changing list of rules to be generated in real time by the interaction of the computer brain through its body with the open world. The robot becomes its changing list of rules. Many people since Dreyfus have emphasised the need for a body acting in the real world. I have spelt out how this can be done in two books. [11, 12].

To give you an idea of how a computer brain can be driven from a changing collection of rules, here is a very brief and superficial account of PP.



**Figure 1.** Cortical Areas of PURR-PUSS

Think of the PP ‘brain’ as a cortical sheet with a number of labelled areas, which we will here call *Cortical Areas*. See Figure 1. Each Cortical Area

has bundles of inputs from different sensory processors or from other Cortical Areas; it has an output bundle carrying an action or stimulus prediction or an input to another Cortical Area.

## **5. Rules**

The connections to and from a Cortical Area determine what rules it can learn, each rule being of the form:

**IF** condition (= context), **THEN** output.

or, in greater detail,

**IF** these input events (one from each input bundle) occur together,  
**THEN** this event of the output bundle is likely to follow.

The input events can be delayed versions of stimulus or action events, so a rule condition is a spatio-temporal sample of event space. Rules can represent dynamic behaviour in multi-dimensional space. If the rule condition is satisfied, then the output will either contribute to the selection of an action or it will contribute to the prediction of the next stimulus. When the memory for rules fills up, the rules that haven't been used for the longest time are discarded. Needless to say, the collection of rules is in a constant flux of change. As long as this change is being caused by the interaction of the robot with the *real, open, unpredictable world*, this collection of rules executed by a low-level program remains *open-ended, nonalgorithmic* and therefore *not a program*.

## **6. Encodingism**

My next topic is encodingism and I will illustrate its avoidance with an experiment I have been carrying out this year. I get the term “encodingism” from Bickhard & Terveen [13]. First I should give you examples of its use in AI and Neural Networks.

SOAR [14] is a classical AI system which claims to be an architecture for general intelligence. When given a problem to work on, it is given a problem space complete with symbols representing the various objects and attributes it has to work with. The meanings of these symbols are in the *head of the user, not in SOAR itself*. SOAR can acquire relationships between these symbols, but the only way for it to learn the meanings of the symbols would be through a complex name-learning process, as we do. The same is true of PP, but in its case we *do not pretend that PP understands* the symbols we give it.

A classical use of encodingism in Neural Networks is Rumelhart and McClelland's "On Learning the Past Tense of English Verbs" [15]. Again the researchers attach meanings to the coding of inputs they feed into the networks and to the outputs that the networks produce. The networks simulate a process but the meanings of the codes used are ascribed by the researchers. All is well *as long as no one says* that the network is learning the past tense of English verbs!

A third example would be NETtalk which is a neural network that is said to read English text. It is given strings of English text and it outputs phonemes into a speech synthesiser. A fourth example would be the whole field of Expert Systems. Expert Systems are based on encodingism, but few people would count them as intelligent. It is the job of a knowledge engineer to give an expert system symbols and rules for representing its knowledge.

Rodney Brooks' [16] robots *avoid* encodingism by having layers of competences in what he calls a subsumption architecture. He insists that an intelligent system's representations must be *grounded in the physical world*. That is, symbols and meanings must originate in the interaction between the intelligent system and the world.

There does seem to be a ground swell of opinion towards robot systems in which knowledge is learned by interaction with the world through the body.

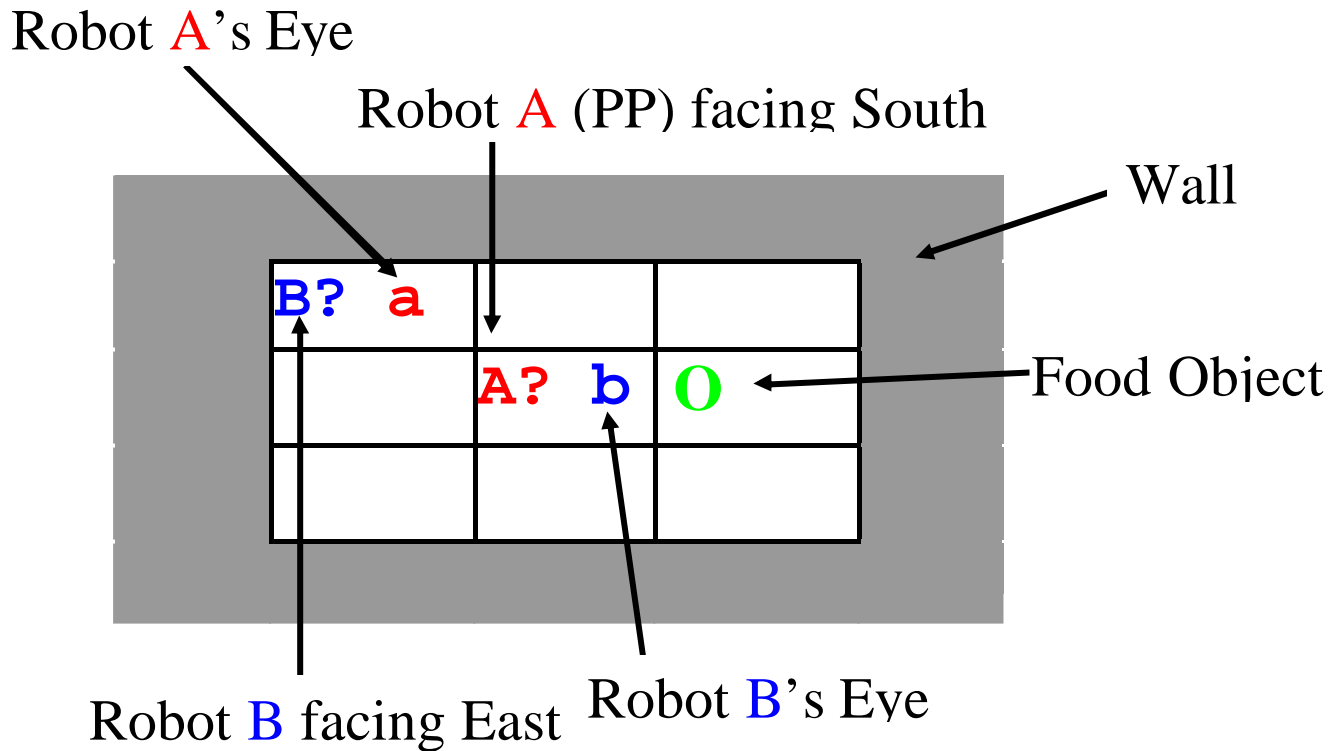
## **7. Squash-Pop Microworld**

My current experiment is a preliminary study in robot awareness. Like all my work with PP, it tries to avoid encodingism. The object of the experiment is to have a robot learn the significance of some Sounds without the designer or experimenter giving PP that significance on a plate. This description of my experiment in the Squash-Pop Microworld is very sketchy because you won't need more than the main ideas.

Two robots, A and B, move on a 3 x 3 squared board surrounded by wall squares. (See Figure 2.) Robot A has a PP brain; robot B can be programmed as we wish. Robot A has an eye 'a', while robot B has an eye 'b'. Each eye can be moved North, East, South or West.

One of two objects, food 'O' or drink 'Q', is on the board at any time. A robot can move Forward into the square in front of it, or rotate 90° Left in its square or rotate 90° Right in its square.

## Squash-Pop Microworld



**Figure 2.** The Squash-Pop Microworld.

Food and drink objects are consumed by a robot squashing them in a corner. It takes two robots to complete a squash, one pushing and the other stopping the object from being squeezed sideways. When a food or drink object is squashed, a new object pops up in a random empty square. After 20 squashes, a food object reappears as a drink object, a drink object as a food object.

The robots have hunger and thirst drives, controlled by food and drink reserves. These reserves are increased by the squashing of food and drink objects. The food reserve is used up by Forward actions, representing work, while the drink reserve is used up by all actions, representing time. If a reserve is increased above a certain level, satiation sets in and the drive is switched off. Then, when the reserve drops down below another level the drive is switched on again.

Each robot has an eye that it can move over the board and walls. The square immediately under the eye gives a 'Fovea' stimulus. The four squares North, East, South and West of that square give the 'Wide' stimulus. If one robot can see the other robot in its Fovea or Wide stimuli, the 'Direction' stimulus gives the direction of that robot relative to its own direction. The

‘Proprioception’ stimulus is the position of the robot’s eye relative to its body. The ‘Touch’ stimulus tells the robot whether it has an object, robot, wall or nothing in the squares either side of its body and in front of its body. Each robot can make Sounds and can hear both its own and the other robot’s Sounds.

## **8. Mother and Infant**

Robot B’s relationship with robot A is intended to be like a mother and infant, with robot B as the mother and robot A as the infant. It is a fairly old idea in psychology that mother (B) sets up goals for the infant (A) to achieve, helps the infant to achieve them, and they then end up as goals of the infant. We can program robot B to act like a mother or carer.

1. Deterministic strategy which is easier for A to predict.
2. If A is silent, B makes a Sound corresponding to its move.
3. B does a body move corresponding to A’s Sound.
4. If A is not hungry or thirsty, B tells A what move to make.  
B says GOOD if A does 2 consecutive told moves.

**Figure 3.** Robot B has 4 Programs.

With robot B obeying its programs (Figure 3), robot A learns to control robot B by using Sounds learned from robot B, and it also learns to respond to robot B’s use of these Sounds to get a GOOD reward from robot B.

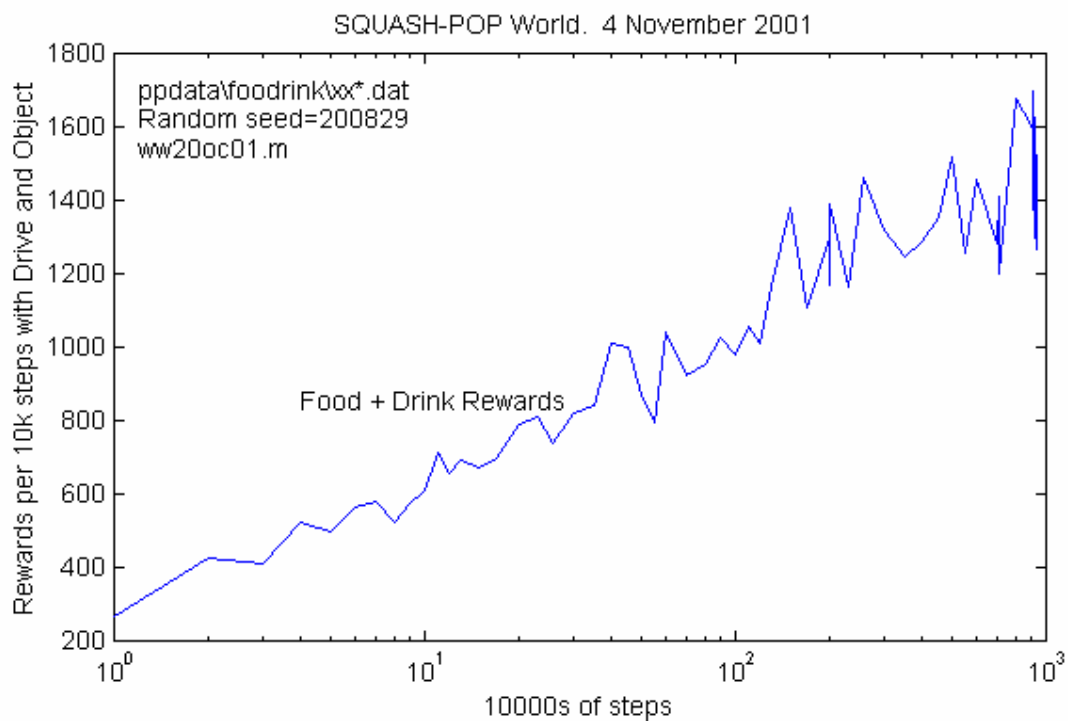
It is essential to remember that robot A, which has the PP brain, starts with *no understanding* of the symbols A, B, O, Q, Forward, Left, Right, Wall, Space, North, East, South and West. These are meaningful to us but not to PP. Any of them can be exchanged at the start for symbols that are meaningless to us as well as to it, but that would make it difficult for us to keep track of what was happening.

## **9. Results of the Experiment**

Figure 4 shows robot A learning to get food and drink and GOOD rewards. Figure 5 shows how the number of steps spent trying to get Food and Drink rewards decreases as robot A gets smarter. This leaves more time for robot A to

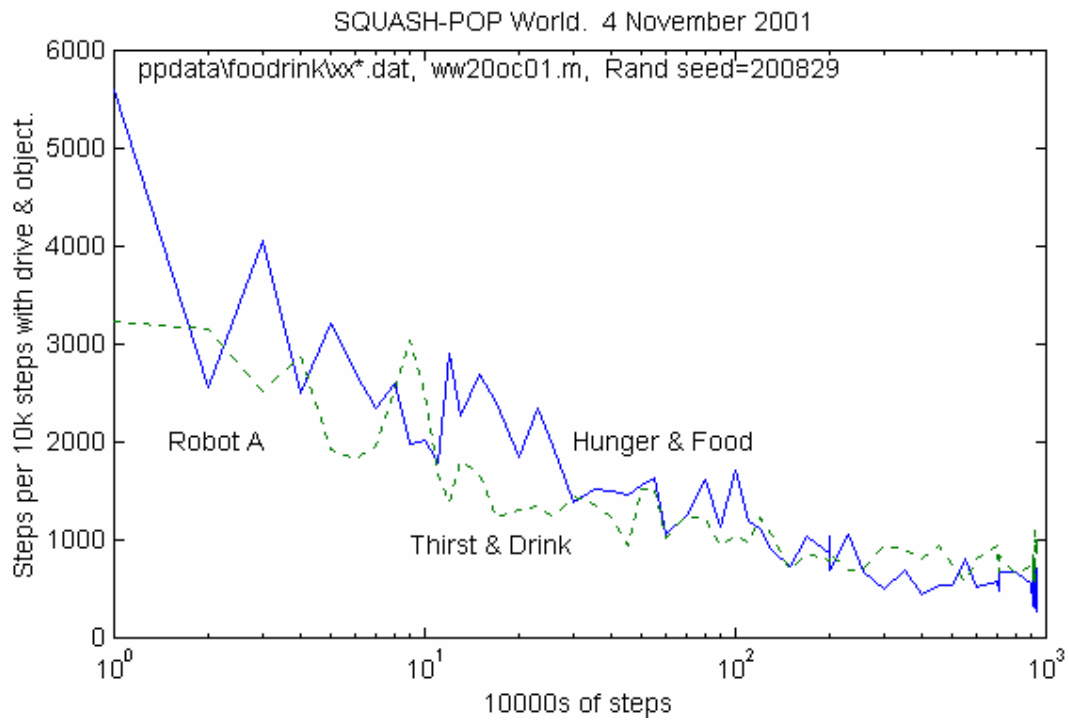
get GOOD rewards. Figure 6 shows how the ratio of GOOD rewards to told moves increases with time.

## Food and Drink Rewards



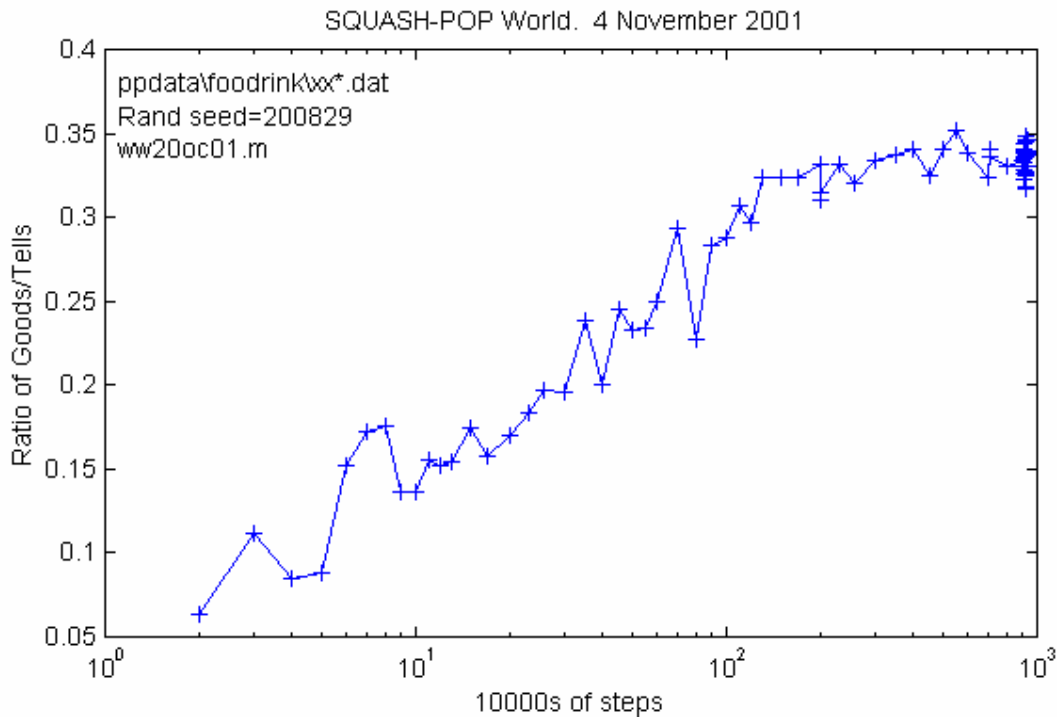
**Figure 4.** Robot A learns to get Food and Drink Rewards.

# Steps with Drive and Object



**Figure 5.** Number of steps in 10,000 used for getting Food and Drink.

# Ratio of GOODs to Tells



**Figure 6.** Ratio of GOOD Rewards obtained to Number of Moves Told.

## 10. Top-Level Program Still Lurking

Now it has to be admitted that the top-level program has **not** been avoided entirely in this experiment. PP has built-in knowledge of its reward signals. My excuse is that we humans also have built-in reward-responses to food and drink when we are hungry and thirsty, and we are also primed to respond to expressions of emotion in facial patterns and voice intonation.

However, if all behaviour is directed by built-in reward and punishment then there is a top-level program in the driving seat. Humans escape this top-level program by having *a natural curiosity*.

PP gains this ultimate liberty by having *novelty goals*. Every new rule learned by a Cortical Area becomes a goal until it is used again. Since new rules are new because of what is already in the collection of rules and because PP is its changing collection of rules, novelty goals can be said to be *self-set*. With novelty goals, PP finally escapes all traces of a fixed top-level program or algorithm. Interestingly, novelty goals make PP much more teachable.

Of course, PP is not free of limitations while simulated in the finite Squash-Pop microworld. That freedom will have to wait until someone with the money puts PP or its equivalent in a versatile robot body in the real world.

## **11. A Plan**

Robot A makes and follows 2 or 3 plans every 100 steps, so the plan in Figure 7 is just one of hundreds.

While robot A is moving Forward, moving its eye North and saying “Right,” it is making a plan. Since the plan is then followed exactly including all predicted stimuli, I am showing here what happened in the *following of the plan*. You need to bear in mind that movements of robot B are in the plan only in so far as they appear in robot A’s stimuli or are implicit in robot A’s Sounds. You can see that robot B turns right when told. Now, robot B’s eye is not in the plan at all, because in this experiment robot A wasn’t able to see where robot B was looking. That is a feature of the microworld I have not yet exploited.

The second view is the situation robot A predicts with its first prediction of stimuli. It now chooses to move Forward, move its eye South and say “Left” as the first step of its plan. Notice that throughout this plan robot A keeps its eye on robot B and the Food object. I am showing what robot B did when robot A was *following the plan*.

The third and fourth views show the second and third steps of the plan.

In the fifth view, robot A is ready to move forward and squash the food object. With the plan made, the moves are carried out and everything goes according to plan.

In the experiment sketched briefly above, robot A with its PP brain learns to use Sounds to control the movements of robot B and learns to do the movements corresponding to those Sounds when they are made by robot B. There is, therefore, *a weak sense in which PP has learned the meaning of the Sounds*.

**Figure 7.** A Plan (on the next page)

|  |           |  |    |  |
|--|-----------|--|----|--|
|  |           |  |    |  |
|  | O         |  |    |  |
|  | B?<br>a b |  |    |  |
|  |           |  | A? |  |
|  |           |  |    |  |

Step 9003930

Robot A: move Forward, eye North  
say "Right"

Robot B rotate Right, (eye North)  
say nothing

|  |     |  |    |  |
|--|-----|--|----|--|
|  |     |  |    |  |
|  | Oab |  |    |  |
|  | B?  |  | A? |  |
|  |     |  |    |  |
|  |     |  |    |  |

Step 9003931 Plan-1

Robot A: move: Forward, eye South  
say "Left"

Robot B: rotate Left, (eye South)  
say nothing

|  |           |  |    |  |
|--|-----------|--|----|--|
|  |           |  |    |  |
|  | O         |  | A? |  |
|  | B?<br>a b |  |    |  |
|  |           |  |    |  |
|  |           |  |    |  |

Step 9003932 Plan-2

Robot A: rotate Left, eye North  
say "Right"

Robot B: rotate Right, (eye North)  
say nothing

|  |     |  |    |  |
|--|-----|--|----|--|
|  |     |  |    |  |
|  | Oab |  | A? |  |
|  | B?  |  |    |  |
|  |     |  |    |  |
|  |     |  |    |  |

Step 9003933 Plan-3

Robot A: move Forward, eye South  
say "Left"

Robot B: rotate Left, (eye East)  
say nothing

|  |     |     |  |  |
|--|-----|-----|--|--|
|  |     |     |  |  |
|  | O   | A?b |  |  |
|  | B?a |     |  |  |
|  |     |     |  |  |
|  |     |     |  |  |

Step 9003934 Plan-4

Robot A: moves Forward  
and squashes Food object.  
Rewarded.

In addition, there is also *a weak sense in which PP is aware of both itself and robot B* because it can see robot B and itself with its eye and can feel the other robot with its touch sense. PP's plans record its own and robot B's movements and the influence of its Sounds on those movements. Nevertheless, this is a long way from the kind of awareness and consciousness that we have. PP does not know anything about the symbols and their meaning. It has only learned simple functional relationships between them.

In the last part of this talk, I would like to examine the *prospects for giving a robot like PP real consciousness*.

## **12. Visual Consciousness**

The *global neuronal workspace model* of consciousness, which I mentioned at the beginning of this talk, is a kind of summary of discoveries in Neuroscience. Their evidence points to the need for a global organization of the brain rather than a local centre to account for awareness.

A new paper in the Behavioral and Brain Sciences Journal seems to have the right kind of explanation. Here is a very brief summary of what O'Regan and Noë [17] say about visual consciousness.

Seeing is an active process, in which we explore the world. Visual exploration uses *sensorimotor contingencies*, a mouthful I'll abbreviate to **SCs**. Vision involves active mastery of the SCs. SCs are the structure of rules governing sensory change produced by various motor actions. The SCs are learned and are available for recall. Vision involves active mastery of the SCs. Normal seeing ceases when eye movement is prevented.

The rules seem to be the same sort of rules as we postulate for the Cortical Areas of PP, rules that relate sensory change to motor actions.

Visual qualities, like the redness of red, are due to those SCs set up by our visual apparatus. Visual attributes of objects, such as their shape, are due to those SCs which have been determined by how the light entering our eyes changes as we move our viewpoint.

We become visually aware by tracking environmental features using SCs and integrating this with thought and action-guidance. This is how they explain *visual consciousness*.

We have the impression of seeing everything in front of us as a complete picture because “the slightest flick of the eye or attention allows any part of the visual scene to be processed at will.” It is easy to assume that the refrigerator light is always on.

Visual transients grab our attention giving us the impression of “having tabs on everything that might change.”

### **13. Simple Example of Constructed Vision**

Here is a simple example I have concocted to illustrate what O’Regan and Noë are talking about.

Imagine that you are at a flower show. You seem to be surrounded by colourful flowers. If you could fix your eyes so that you only saw what was landing on your retinas in one position of your eyes, then there would be a small central area in sharp colour surrounded by a more and more blurred and less and less coloured periphery.

In fact your eyes are not stationary even when you are looking at one flower. Saccades and other eye movements turn the scene into a uniformly sharp and colourful picture in front of you.

O’Regan & Noë’s paper seems to be just what is needed to explain why consciousness needs a global organization. Theirs is a constructivist explanation.

Visual consciousness, together with our spatial awareness of sound and touch, is something we are *constructing* all the time.

Rodney Cotterill’s [18] explanation of consciousness as an active process, the construction of schemata, and “a mechanism which enables us to relate to our personal space” is very similar.

Visual consciousness is like a ballet, which works so long as everyone keeps to the choreography, but a single slip can collapse the whole display. The integrity of the organization is crucial.

My guess is that we will get used to this kind of an explanation just as we have got used to the explanation of electric and magnetic phenomena in terms of fields. Our minds don’t readily accept action-at-a-distance which the idea of a field softens, and they have as much difficulty in thinking about subjectivity.

Both are fundamental properties of the universe we live in and the best we can do is to find consistent explanations that can become part of our thinking.

#### **14. Talking-to-Ourselves Consciousness**

Visual consciousness is a major part of the consciousness of the not blind and probably of the higher mammals, but there is another equally, if not more, important part of our consciousness.

I am thinking of that inner narrative in which we constantly talk to ourselves. Let me call it verbal consciousness. There are three influential and compatible theories of consciousness which focus mainly on this aspect.

According to Rosenthal [19], conscious states must be accompanied by higher-order thoughts, and non-conscious states cannot be. For Bridgeman [20], consciousness is the operation and monitoring of plans. While, according to Weiskrantz [21], the ability to make a commentary on any particular event is what we mean by being conscious.

In each case, we have the requirement of *a higher level referring to a lower level*. Since in my learning robot, PP is continually making plans and trying to follow them, Bridgeman's version is the easiest for me to envisage.

So, how does one arrange for PP to monitor its own plans. There is a seamlessness about the way we do it and I have only one idea for its realization. *In one sentence* I can announce a plan and comment on it: "I must go down to the shops to buy some bread this morning because if I leave it until later the shops may be sold out." This suggests that when PP makes a plan, all actions in its plan should be imaginary *except speech*. The speech in the plan should be a continuation of PP's normal speech. (See Appendix.)

Whether this is the right way to approach verbal consciousness or not is of less importance here than that there appears to be a route.

We have already seen O'Regan and Noë's route to visual consciousness. Both appear to be compatible with the neuroscientists' global neuronal workspace model. The organizational requirement of O'Regan and Noë's active and constructive vision fits well with the global access requirement of the workspace, as does the making of plans.

Things are falling into place. *Consciousness is losing its mystery.*

## **15. The Problem of Experience**

Or is it? Listen to David Chalmers [22]:

“The really hard problem of consciousness is the problem of experience. There is *something it is like* to be a conscious organism. ... Why should physical processing give rise to a rich inner life at all? It seems objectively unreasonable that it should, and yet it does.”

Chalmers has argued that even if you find mechanisms for all the *easy problems* of consciousness (his list is given in Figure 8), you still have not solved the hard problem of experience.

- the ability to discriminate, categorize, and react to environmental stimuli;
- the integration of information by a cognitive system;
- the reportability of mental states;
- the ability of a system to access its own internal states;
- the focus of attention;
- the deliberate control of behaviour;
- the difference between wakefulness and sleep.

**Figure 8.** Chalmers’ Easy Problems.

Dan Dennett [23] says that is *absurd*. There is no hard problem beyond the easy problems.

Dennett argues that the understanding of consciousness now is similar to *understanding life before the discovery of DNA*. People could have argued then that even if you found mechanisms for reproduction and growth, there might still be entities that used these mechanisms but didn’t have life.

## **16. Thought Experiment**

I’ll come back to Chalmers in a moment, but first let me do a little thought experiment with PP. Suppose that PP has an eye, as in the Squash-Pop Microworld I described above, but the Fovea is no longer a single cell and the periphery is more extensive. The world is bigger and richer. There are many more Cortical Areas for different kinds of rules.

Movement of the robot through the world will activate other rules describing optical flow and the changing shapes of objects. This seems to be the kind of system that O'Regan and Noë are talking about. We begin to wonder whether it wouldn't be *like something to be such a robot*.

As things become more complicated the robot's vision might become more like ours. All the subtle relationships, which Dennett claims amount to having experience, would be incorporated in the adaptive rules. Maybe he is right and we should expect such a robot to have a visual experience similar to ours. *Maybe all that is needed for experience* is that there should be the appropriate functional relationships, that is, the SCs.

If functional relationships are all that is needed, then thought experiments in which the neurons of a human brain are replaced by functionally identical silicon circuits can take us from humans to robots, or at least to silicon isomorphs, without change of experience.

On this account, the brain (human or robot) is just a *very complex control system*, so the brain is not different from a suitably powerful computer, with signals coming from the various sensors in its body and going to various effectors (muscles and glands).

If now, for a short time, the robot goes into a *closed room*, then it should have the same sort of experiences that we would have in the closed room. However, both body and room would seem to be candidates for computer simulation. We can have a brain computer connected to a body-and-room computer. Finally we combine the lot in one massive computer. All that this overall computer is doing is manipulating binary words. There are no physical sensors or effectors, but the functional relationships or SCs are being represented perfectly.

I find it difficult to believe that *somewhere in those computations there is an entity experiencing something*.

## **17. Hard Problem Still Relevant**

So Chalmers' hard problem is still relevant for some of us.

My own escape-from-consciousness-in-simulations hypothesis is that the functional relationships which have to be satisfied by our experiences are like differential equations. Even when you have solved the equations, you still have to put in *boundary conditions*.

You have to tie down all those relative conditions to something absolute. This is done, perhaps, by the physics of our brains and bodies. When all the computations are being carried out appropriately, the physical nature of the body and world sets the absolute reference, or boundary condition, for the experience.

If this is the case, then one would expect some difference between the experience of a human and a real robot because of their different physical implementations. Better still, the isolated computer with brain, simulated body and simulated room could be *without any experiencing entity at all*.

So much has been written on consciousness that this is unlikely to be a new hypothesis. But is it *testable*?

We can wait until a real robot with language can describe to us what it feels, but even then, aren't those boundary conditions just that part of subjective consciousness which we can't share with others?

How could we ever decide whether I saw a particular red surface with the same redness as you did? Even when we have discussed all the possible relationships between a particular red surface and other surfaces with different hues, illuminations and textures, could we be sure that there isn't some difference in absolute reference between us?

Fortunately, this philosophical problem is *no* hindrance to the building of conscious robots.

It is the *functional relationships* that we have to get right. We can leave it to the physics to provide the remainder, if there is any remainder to provide.

## **Appendix: Memory and Consciousness**

The rules of the Cortical Areas in PP are of two kinds [12] *choice* and *replacement*. In a choice rule, the THEN part holds all the events that have followed the IF part in the past. In a replacement rule, the THEN part holds only the most recent event to have followed the IF part. These two kinds of rule hold two kinds of information, such as “varieties of fruit” (choice) and “what kind of fruit was in the fruit bowl when I last saw it” (replacement). Thus, choice memory holds all alternatives encountered, while replacement memory holds only the most recent example.

The memory associated with consciousness must be of the replacement type for two reasons. First, alternative perceptions and meanings cannot occur simultaneously in conscious thought [24]. For example, we cannot have alternative meanings of a word or alternative forms of a Necker Cube in consciousness at the same time. Secondly, only replacement memory can be used to construct a complex structure because its integrity and strength will depend upon each part being in the correct (i.e. unique) position. This will apply equally to the structure building of our visual consciousness and to the structure building for meaning and verbal consciousness.

In my initial experiments with the self-monitoring of plans, I see both kinds of memory being important. There will be a Long Term Choice Memory (LTCM) and a Medium Term Replacement Memory (MTRM). Plans will continue to be constructed with LTCM but this will be accompanied by speech-led monitoring and checking using MTRM. Two examples illustrate what I have in mind.

Example-1. My wife asks me to get an apple from the fruit bowl in the dining room. My mind travels mainly unconsciously by a plan through LTCM to the dining room with an apple in the fruit bowl. This plan provides context for my speech-led MTRM to make me say “But I only saw a couple of oranges in the bowl when I was there a little while ago.”

Example-2. PP makes a plan with LTCM in the Squash-Pop Microworld to squash the Food object in the corner, but the Sounds-led MTRM recalls seeing a Drink object and cancels the plan.

## **References**

1. Dennett, Daniel C. (2001): Are We Explaining Consciousness Yet? *Cognition*, **79** 221-237.
2. Dehaene, Stanislas & Naccache, Lionel (2001): Towards a Cognitive Neuroscience of Consciousness. *Cognition*, **79** 1-37.
3. Andreae, John H. (1963): STeLLA: A Scheme for a Learning Machine. *Proceedings 2<sup>nd</sup> IFAC Congress*, Basel.
4. Winograd, Terry (1972): *Understanding Natural Language*. Academic Press.
5. Dreyfus, Hubert L. (1972): *What Computers Can't Do*. Harper & Row.
6. Winograd, Terry & Flores, Fernando (1986): *Understanding Computers and Cognition*. Ablex Publishers.
7. Searle, John R. (1980) Minds, Brains and Programs. *Behavioral & Brain Sciences*, **3** (3) 417-457.

8. Penrose, Roger (1989) *The Emperor's New Mind*. Oxford University Press.
9. Penrose, Roger (1994) *Shadows of the Mind*. Oxford University Press.
10. Lucas, J.R. (1961) : Minds, Machines and Gödel. *Philosophy*, **36**, reprinted in A.R.Anderson (ed.) *Minds and Machines*. Prentice-Hall (1964).
11. Andreae, John H. (1977) *Thinking with the Teachable Machine*. Academic Press.
12. Andreae, John H. (1998) *Associative Learning for a Robot Intelligence*. Imperial College Press.
13. Bickhard, Mark H. & Terveen, Loren (1996): *Foundational Issues in AI and Computer Science*. Elsevier.
14. Norman, D.A. (1991): Approaches to the Study of Intelligence. *Artificial Intelligence*, **47** 327-346.
15. McClelland, J. L. & Rumelhart, D. E. (1986): *Parallel Distributed Processing*. MIT Press, **2** 216-271.
16. Brooks, Rodney A. (1991): Intelligence without Representation. *Artificial Intelligence*, **47** 139-159.
17. O'Regan, J. Kevin & Noë, Alva (2001): A Sensorimotor Account of Vision and Visual Consciousness. *Behavioral & Brain Sciences*, **24** (5). <http://www.bbsonline.org/Preprints/ORegan/>
18. Cotterill, Rodney (1998): *Enchanted Looms*. Cambridge University Press.
19. Rosenthal, D. (1986): Two Concepts of Consciousness. *Philosophical Studies*, **49** 329-359, discussed in Dennett (1991): *Consciousness Explained*. Penguin.
20. Bridgeman, Bruce (1992) On the Evolution of Consciousness and Language. *psycoloquy*.92.consciousness.1.bridgeman
21. Weiskrantz, Lawrence (1997): *Consciousness Lost and Found*. Oxford University Press.
22. Chalmers, David J. (1995): Facing Up to the Problem of Consciousness. *J. Consciousness Studies*, **2**(3) 200.
23. Dennett, Daniel C. (1996): Facing Backwards on the Problem of Consciousness. *Journal of Consciousness Studies*, **3**(1) 4-6.
24. Baars, Bernard J. (1997): *In the Theater of Consciousness*. Oxford University Press.