Evolving Robot Consciousness: The Easy Problems and the Rest

Inman Harvey
School of Cognitive and Computing Sciences
University of Sussex
Brighton BN1 9QH, UK

1 Introduction

Car manufacturers need robots that reliably and mindlessly repeat sequences of actions in a well-organised environment. For many other purposes autonomous robots are needed that will behave appropriately in a disorganised environment, that will react adaptively when faced with circumstances that they have never faced before.

The design of autonomous robots has an intimate relationship with the study of autonomous animals and humans — robots provide a convenient puppet show for illustrating current myths about cognition. Like it or not, any approach to the design of autonomous robots is underpinned by some philosophical position in the designer. Whereas a philosophical position normally has to survive in debate, in a project of building situated robots one's philosophical position affects design decisions and is then tested in the real world — "doing philosophy of mind with a screwdriver".

In this paper I shall first follow other authors in distinguishing various uses of the word 'consciousness'. Using Chalmers' characterisation of 'the easy problems of consciousness' (Chalmers, 1995) I shall show how the evolutionary approach to robotics handles them. But then the main focus of the paper will be what Chalmers calls 'the hard problem' — which I will suggest is an easy non-problem.

2 The easy problems and the hard problem

There is currently a fashion for asserting that 'consciousness' has become a respectable topic for scientists, as well as philosophers, to discuss. A number of scientists, many of them eminent in their own fields within which there is a general consensus on the usage of technical terms, have blithely assumed that there is a similar consensus in discussion of consciousness. Oblivious of the multiple meanings available for the word 'consciousness', they frequently talk past each other and their audience.

Philosophers have for the most part been aware of this potential for confusion (though that does not mean that they have always avoided it). Chalmers (1995) warns of these multiple meanings, and proposes a basic distinction between those types of consciousness which offer (relatively) 'easy' problems for the scientist, and in contrast the 'hard one'. I will be largely agreeing with his analysis of the easy problems, and will tackle them first.

Chalmers defines the easy problems as "those that seem directly susceptible to the standard methods of cognitive science, whereby a phenomenon is explained in terms of computational or neural mechanisms". He chooses a different list from mine, but he covers the same ground. My broad categories I will characterise in terms of degrees of consciousness while driving my car home from work past a particular bend in the road:

- consciousness^1: though I have no recollection of the journey, I got home safely, so I cannot have fallen asleep or blacked out.
- consciousness^2: though I have no recollection of the roadside advertisement recently erected at the bend, later that evening I choose the new brand of beer that was advertised there, so it did indeed affect my later behaviour.
- consciousness^3: I notice the advertisement, and on arriving home I can recollect and describe it.

Here I have described these scenarios in the first person, but we can check whether somebody else is conscious$^{1-3}$ by simple tests:

1. Did they react to the environment?
2. Was their later behaviour changed?
3. Can they report verbally on what they experienced?

All three of these classes of consciousness are 'easy' in Chalmers' sense: in principle we expect no mystery or magic in the underlying neural mechanisms in animals or humans. The complexity and the details may be difficult, and Chalmers suggests that it might take a century or two of work to uncover them, but conceptually we have no difficulties. The first two types of consciousness we attribute to non-human animals, and we can and do currently make robots that exhibit them. Indeed we have even simpler mechanisms. When we test a newly installed burglar alarm for its sensitivity to an intruder's movement, we are testing for consciousness^1. A soft-drinks vending machine which accepts a sequence of coins and alters its internal state in doing so passes the test for consciousness^2.

Consciousness^3 in contrast seems currently constrained to humans, on any definition of linguistic competence that excludes a trained parrot or a telephone answering machine. Nevertheless, as long as one takes the criterion for such consciousness^3 in a third party to be the issuing of appropriate words in an appropriately wide range of contexts (in continued interaction with other language-users) then I agree with Chalmers. I go along with the basic credo of a cognitive scientist that ultimately we will be able to demonstrate underlying mechanisms that generate such behaviours. Where I differ from most of my colleagues is in my expectation that we will never *comprehend* how such mechanisms operate *as a whole*, even when we can *create* or *display* them, and comprehend any small part of them. This limitation is not because of any deep mystery, but simply due to the limitations of us poor humans in understanding complex systems; below I will discuss how an evolutionary approach allows emulation without comprehension.

This list does not yet include what I will here call consciousness* and Chalmers (1995) characterises as:

> The really hard problem of consciousness is the problem of *experience*. When we think and perceive, there is a whir of information-processing, but there is also a subjective aspect.

3 Zombies, Computation and Dynamical Systems

That philosophical favourite, the zombie, can pass all the tests for consciousness$^{1-3}$, yet lacks 'phenomenal consciousness' or 'qualia', or 'conscious experience' — my 'consciousness*'. It can react to red traffic lights, even utter the words that describe yesterday's red sunset, yet it does not *experience* the sensation of red that I have, that I assume you have.

I cannot distinguish a zombie from you by its behaviour; this has the oft-forgotten corollary that I cannot distinguish you from a zombie by your behaviour. This does not stop me from in practice treating all humans who display signs of consciousness$^{1-3}$ as being more than zombies, as having consciousness*. I also find myself doing likewise with a fair number of animals, and even the occasional machine when I am being slapdash: "that printer always chooses the day of a deadline to break down, it seems to enjoy being awkward". One goal of Artificial Intelligence (AI) is to produce machines which emulate human performance of all kinds; not just the intelligence of humans, but ultimately the consciousness also. This could naively be broken down into two tasks:

1. Produce a zombie machine with the right behaviours — the easy problems.
2. Add the extra ingredient that gives it consciousness*, that makes it more than a zombie — the hard problem.

The first task is that of the roboticist. In the past AI practitioners might have said AI-oriented computer scientist rather than roboticist, given the prevailing fashion of the 1960s–80s of equating cognition with computation. Many, such as Penrose (1989), whose knowledge of AI is secondhand do not appreciate that for some time the new thrust in AI has been towards situated, embodied cognition, and a recognition that formal computation theory relates solely to the constraints and possibilities of machines (or people) carrying out algorithmic procedures. For those who do not accept the computational perspective on cognition, the worries advanced by Penrose are meaningless and irrelevant.

The astronomer, and her computer, perform computational algorithms in order to predict the next eclipse of the moon; the sun, moon and earth do not carry out such procedures as they drift through space. The cook follows the algorithm (recipe) for mixing a cake, but the ingredients do not do so as they rise in the oven. Likewise if I was capable of writing a computer program which predicted the actions of a small creature, this does not mean that the creature itself, or its neurons or its brain, was consulting some equivalent program in 'deciding what to do'.

Formal computations are to do with solving problems such as 'when is the eclipse?'. But this is an astronomer's problem, not a problem that the solar system faces and has to solve. Likewise, predicting the next movement of a creature is an animal behaviourist's problem, not one that the creature faces. However, the rise of computer power in solving problems naturally, though regrettably, led AI to the view that cognition equalled the solving of problems, the calculation of appropriate outputs for a given set of inputs. The brain, on this view, was surely some kind of computer. What was the problem that the neural program had to solve? — the inputs must be sensory, but what were the outputs?

Whereas a roboticist would talk in terms of motor outputs, the more cerebral academics of the infant AI community tended to think of plans, or representations, as the proper outputs to study. They treated the brain as the manager who does not get his own hands dirty, but rather issues commands based on high-level analysis and calculated strategy. The manager sits in his command post receiving a multitude of possibly garbled messages from a myriad sensors and tries to work out what is going on. Proponents of this view tend not to admit explicitly, indeed they often deny vehemently that they think in terms of a homunculus in some inner chamber of the brain, but they have inherited a Cartesian split between mind and brain and in the final analysis they rely on such a metaphor.

An alternative view has gained favour in the last decade, though its origins date back at least to the early cybernetics movement. One version of this is the Dynamical Systems view of cognition:

> . . . animals are endowed with nervous systems whose dynamics are such that, when coupled with the dynamics of their bodies and environments, these animals can engage in the patterns of behavior necessary for their survival. (Beer & Gallagher 1992: 91)

At this stage we downgrade the significance of *intelligence* for AI in favour of the concept of *adaptive behaviour*. Intelligence is now just one form of adaptive behaviour amongst many; the ability to reason logically about chess problems may be adaptive in particular refined circles, but the ability to cross the road safely is more widely adaptive. We should note the traditional priorities of AI: the computationalists' emphasis on reasoning led them to assume that everyday behaviour of sensorimotor coordination must be built on top of a reasoning system. Sensors and motors, in their view, are 'merely' tools for information-gathering and plan-execution on behalf of the central executive where the real work is done. Many proponents of an alternative view, including myself, would want to turn this on its head: logical reasoning is built on top of linguistic behaviour, which is built on prior sensorimotor abilities. These prior abilities are the fruit of billions of years of evolution, and language has only been around for the last few tens of thousands of years.

A dynamical system is formally any system with a finite number of state variables that can change over time; the rate of change of any one such variable depends on the current values of any or all of the variables in a regular fashion. These regularities are typically summed up in a set of differential equations. A Watt governor for a steam engine is a paradigmatic dynamical system (van Gelder, 1992), and we can treat the nervous system plus body of a creature (or robot) as one also. The behaviour of a dynamical system such as the governor depends also on the current value of its external inputs (from the steam engine) which enter the relevant differential equations as parameters. In a complementary way, the output of the governor acts as a parameter on the equations which describe the steam engine itself as a dynamical system. One thing that is very rapidly learnt from hands-on experience is that two such independent dynamical systems, when coupled together into (e.g.) steam-engine-plus-governor treated now as a single dynamical system, often behave in a counterintuitive fashion not obviously related to the uncoupled behaviours.

Treating an agent — creature, human or robot — as a dynamical system coupled with its environment through sensors and motors, inputs and outputs, leads to a metaphor of agents being *perturbed* in their dynamics through this coupling, in contrast to the former picture of such agents *computing* appropriate outputs from their inputs. The view of cognition entailed by this attitude fits in with Varela's characterisation of cognition as *embodied action*:

> By using the term *embodied* we mean to highlight two points: first, that cognition depends upon the kinds of experience that come from having a body with various sensorimotor capacities, and second, that these individual sensorimotor capacities are themselves embedded in a more encompassing biological, psychological and cultural context. By using the term *action* we mean to emphasize once again that sensory and motor processes, perception and action, are fundamentally inseparable in lived cognition. Indeed, the two are not merely contingently linked in individuals; they have also evolved together. (Varela et al., 1991: 172–173)

4 Evolutionary Robotics and Behaviourism

Moving from natural agents to artificial robots, the design problem that a robot builder faces is now one of creating the internal dynamics of the robot, and the dynamics of its coupling, its sensorimotor interactions with its environment, such that the robot exhibits the desired behaviour in the right context. Designing such dynamical systems presents problems unfamiliar to those who are used to the computational approach to cognition.

A primary difference is that dynamics involves time, real time. Whereas a computation of an output from an input is the same computation whether it takes a second or a minute, the dynamics of a creature or robot has to be matched in timescale to that of its environment. A second difference is that the traditional design heuristic of *divide and conquer* cannot be applied in the same way. It is not clear how the dynamics of a control system should be carved up into smaller tractable pieces; and the design of any one small component depends on an understanding of how it interacts in real time with the other components, such interaction possibly being mediated via the environment. This is true for behavioural decomposition of control systems (Brooks, 1991) as well as functional decomposition. However, Brooks' subsumption architecture approach offers a different design heuristic: first build simple complete robots with behaviours simple enough to understand, and then incrementally add new behaviours of increasing complexity or variety, one at a time, which subsume the previous ones. Before the designer adds a new control system component in an attempt to generate a new behaviour, the robot is fully tested and debugged for its earlier behaviours; then the new component is added so as to keep to a comprehensible and tractable minimum its effects on earlier parts.

This approach is explicitly described as being inspired by natural evolution; but despite the design heuristics it seems that there is a practical limit to the complexity that a human designer can handle in this way. Natural Darwinian evolution has no such limits, hence the more recent moves towards the artificial evolution of robot control systems (Harvey et al., 1997).

In this work a genetic encoding is set up such that an artificial genotype, typically a string of 0s and 1s, specifies a control system for a robot. This is visualised and implemented as a dynamical system acting in real time; different genotypes will specify different control systems. A genotype may additionally specify characteristics of the robot 'body' and sensorimotor coupling with its environment. When we have settled on some particular encoding scheme, and we have some means of evaluating robots at the required task, we can apply artificial evolution to a population of genotypes over successive generations.

Typically the initial population consists of a number of randomly generated genotypes, corresponding to randomly designed control systems. These are instantiated in a real robot one at a time, and the robot behaviour that results when placed in a test environment is observed and evaluated. After the whole population has been scored, their scores can be compared; for an initial random population one can expect all the scores to be abysmal, but some (through chance) are less abysmal than others. A second generation can be derived from the first by preferentially selecting the genotypes of those with higher scores, and generating offspring which inherit genetic material from their parents; recombination and mutation is used in producing the offspring population which replaces the parents. The cycle of instantiation, evaluation, selection and reproduction then continues repeatedly, each time from a new population which should have improved over the average performance of its ancestors. Whereas the introduction of new variety through mutation is blind and driven by chance, the operation of selection at each stage gives direction to this evolutionary process.

This evolutionary algorithm comes from the same family as Genetic Algorithms and Genetic Programming, which have been used with success on thousands of problems. The technique applied to robotics has been experimental and limited to date. It has been demonstrated successfully on simple navigation problems, recognition of targets, and the use of minimal vision or sonar sensing in uncertain real world environments (Harvey et al., 1997; Thompson, 1995). One distinguishing feature of this approach using 'blind' evolution is that the resulting control system designs are largely opaque and incomprehensible to the human analyst. With some considerable effort simple control systems can be understood using the tools of dynamical systems theory (Husbands et al., 1995). However, it seems inevitable that, for the same reasons that it is difficult to design *complex* dynamical systems, it is also difficult to analyse them.

This is reflected in the methodology of Evolutionary Robotics which, once the framework has been established, concerns itself solely with the behaviour of robots: "if it walks like a duck and quacks like a duck, it is a duck". For this reason we have sometimes been accused of being 'the New Behaviourists'; but this emphasis on behaviour assumes that there are significant internal states[1], and in my view is compatible with the attribution of consciousness. A major conceptual advantage that Evolutionary Robotics has over classical AI approaches to robotics is that there is no longer a mystery about how one can 'get a robot to have needs and wants'. In the classical version the insertion of a value function `robot_avoid_obstacle` often leaves people uncomfortable as to whether it is the robot or the programmer who has the desires. In contrast, generations of evolutionary selection that tends to eliminate robots that crash into the obstacle produces individual robots that do indeed avoid it; and here it seems much more natural that it is indeed the *robot* which has the desire.

[1] Not 'significant' in the sense of representational — internal states are mentioned here to differentiate evolved dynamical control systems (which typically have plenty of internal state) from those control systems restricted to feedforward input/output mappings (typical of 'reactive robotics').

5 Back to Consciousness

Natural evolution has produced creatures, including humans, through millenia of trials, selection, and heredity with variation. Evolutionary Robotics similarly requires a multitude of trials of real robots within the real world situations they are required to face up to. These trials are explicitly behavioural tests, and on the basis of these we have clearly already produced robots that exhibit consciousness^1 and consciousness^2. For linguistic abilities, consciousness^3, I would agree with Chalmers that in principle this is a problem with no mystery, an 'easy problem', though I would expect more than his couple of centuries will be needed to crack it.

This reintroduction of the word 'consciousness' may have made the reader uneasy; but I am explicitly referring to the definitions of consciousness which can apply to a zombie or a machine. Now why do I attribute something extra to the humans I meet everyday, what is the magic ingredient consciousness*?

If I cannot distinguish between you and a zombie by your behaviour, yet I treat you as something more, then the difference is in my attitude. Given a creature, a human or a robot in front of me, I can adopt a number of different stances (Dennett, 1987). The mechanical stance is one I frequently adopt with machines, rarely with humans: with this perspective I treat the components as lifeless matter obeying physical laws, with not a trace of consciousness*. From a different perspective, however, I normally treat other humans as aware, intentional creatures that are conscious* like myself; I take this perspective occasionally also with machines, though the nature of machines that I come across generally means such a perspective is only short-lived, and soon reverts to a less personal one. As with the two perspectives of a Necker cube, it is impossible to hold both views simultaneously.

On a country drive when I notice some faltering in the power of my car's engine, my attention focuses on it as an object; in Heideggerian terms the car is no longer ready-to-hand, but becomes present-at-hand. If it has been temperamental recently, and I am used to its quirks, then I may well treat it not so much as an object but more like a person, and nurse it carefully with a soft touch on the accelerator. When it finally fails I open up the engine and take a mechanical stance, looking for broken wires or dripping fuel. Whichever of these three stances I take at any one time, like a view of a Necker cube it temporarily blots out any of the other possible perspectives.

6 An Attitude Problem

So if one accepts that the creation of robots with consciousness$^{1-3}$ offers merely 'easy' problems (stretching the sense of 'easy' for consciousness^{3}), the additional magic ingredient for consciousness* is merely a change of attitude in us, the observers. Such a change of attitude cannot be achieved arbitrarily; the right conditions of complexity of behaviour, of similarity to humans, are required first. If an alternative perspective is much easier or more beguiling, then it will be difficult to shift away from. The more we can understand of a system from the mechanical perspective, the less likely we are to attribute agency, personhood, consciousness* to it — which is why the ER approach that can produce comprehensible behaviour from incomprehensible mechanisms offers possibilities that the conventional design approach lacks.

Dennett (1996) in his response to Chalmers makes a move that has some resemblance to the one I have made here. Where Chalmers (1995) suggests that the extra ingredient consciousness* is a fundamental feature of the world, alongside mass, charge, and space-time, Dennett rightly pours scorn on this: consciousness* is not some new entity over and above all these subsidiary phenomena of consciousness$^{1-3}$. Dennett's response will leave Chalmers and many others including myself unsatisfied, however — he seems to be explicitly denying the phenomena, our experience of visual sensations, the redness of the rose, the smell of coffee, and reducing everything to behaviour.

I sympathise with this dissatisfaction which Dennett does nothing to acknowledge or resolve. And I can put forward a perspective from which this problem is not actually *resolved*, but rather *dissolved* in Wittgensteinian fashion.

I take a Relativist perspective, which contrary to the naive popular view does not imply solipsism, or subjectivism, or an anything-goes attitude to science. The history of science shows a number of advances, now generally accepted, that stem from a relativist perspective which (surprisingly) is associated with an objective stance toward our role as observers. The Copernican revolution abandoned our privileged position at the centre of the universe, and took the imaginative leap of wondering how the solar system would look viewed from the Sun or another planet. Scientific objectivity requires theories to be general, to hold true independently of our particular idiosyncratic perspective, and the relativism of Copernicus extended the realm of the objective. Darwin placed humans amongst the other living creatures of the universe, to be treated on the same footing. With Special Relativity, Einstein carried the Copernican revolution further, by considering the viewpoints of observers travelling near to the speed of light, and insisting that scientific objectivity required that their perspectives were equally privileged to ours. Quantum physics again brings the observer explicitly into view. As for mathematics, "I would even venture to say that the principle of mathematical induction is the relativity principle in number theory" — (Foerster, 1984). Cognitive science seems one of the last bastions to hold out against a Copernican, relativist revolution. Amongst the few to have been liberated were some of the early cyberneticists (Foerster, 1984) and more recent philosophies that owe something to them (Maturana and Varela, 1987).

Cognitive scientists must be careful above all not to confuse objects that are clear to them, that have an objective existence for them, with objects that have a meaningful existence for other agents. A roboticist learns very early on how difficult it is to make a robot recognise something that is crystal clear to us, such as an obstacle or a door. It makes sense for us to describe such an object as 'existing for that robot' if the physical, sensorimotor, coupling of the robot with that object results in robot behaviour that can be correlated with the presence of the object. By starting the previous sentence with "It makes sense for us to describe ... " I am acknowledging our own position here acting as scientists observing a world of cognitive agents such as robots or people; this objective stance means we place ourselves outside this world looking in as godlike creatures from outside. Our theories can be scientifically objective, which means that predictions should not be dependent on incidental factors such as the nationality or location or star-sign of the theorist; however we can only be objective about objects, not about our own subjectivity or consciousness*.

Heinz Von Foerster explains why relativism does not lead to solipsism, and in doing so points to why we attribute to others the same consciousness*, qualia, that we experience. Other humans exist for me, and I live in a society where other humans have comparable physical and social relationships — "if you prick us, do we not bleed? ... and if you wrong us, shall we not revenge?". When I try to look at my own behaviour from an external, scientific position, I see remarkable similarities with other people's behaviour, including their interactions with third parties. As a relativist I take the Copernican stance of refusing a privileged 'objective' position — yet clearly the solipsistic position is uniquely privileged. The absurdity of solipsism was brought out by Bertrand Russell's solipsist correspondent who thought it such a sensible attitude that she wondered why there were not more people who agreed with it!

If we reject solipsism, this entails that we attribute consciousness* to others who behave in such a way that we take a personal, intentional stance towards them. This may still leave unresolved the worry of the person who asks: "but is the red that she experiences the same as the red that I experience, when we look at the same object?". Here I follow Wittgenstein in saying that this is a linguistic confusion, a mistaking of subjects for objects.

When I see a red sign, this red sign is an object that can be discussed scientifically. This is another way of saying that it exists for me, for you, and for other human observers of any nationality; though it does not exist for a bacterium or a mole. We construct these objects from our experience and through our acculturation as humans through education[2]. Just as our capacity for language is phylogenetically built upon our sensorimotor capacities, so our objects, our scientific concepts, are built out of our experience. But our phenomenal experience itself cannot be an objective thing that can be discussed or compared with other things. It is primary, in the sense that it is only through having phenomenal experience that we can create things, objective things that are secondary.

One version of the conundrum that puzzles people goes as follows: We agree that 4 billion years ago there was a lifeless planet with no conscious* beings, yet now there are; at some stage this magic ingredient consciousness* appeared, as a product of lifeless matter — how can this be? But if we carefully and consistently make a note of the observers involved in this scenario, the problem dissolves. The 'We' that 'agree' refers to us from the scientifically literate community of the late 20th century. For those of the mid-20th century the earth was only 2 billion years old, and who can now guess what current orthodoxy we will agree on in 50 years time? For us it is the case that 4 billion years ago there was lifeless rock, and there are now conscious* beings; for the sake of argument we can posit a time T when, for us, the first conscious* being appeared. The mystery arises only when we imagine ourselves being present at time T-*1* waiting for something — what? — to happen. But our assumption of no consciousness* before time T makes our imagined scenario — us as conscious observers *then* — illegitimate.

[2] It makes no sense to discuss (...for *us humans* to discuss ...) the existence of objects in the absence of humans. And (in an attempt to forestall the predictable objections) this view does *not* imply that we can just posit the existence of any arbitrary thing as our whim takes us.

7 Summary

I have started off my argument by agreeing with Chalmers' distinction between the easy problems of consciousness and the rest — except where Chalmers sees the rest as a hard problem, I see it as a linguistic non-problem. The easy problems, in the context of robotics, are those of generating the desired behaviours of a zombie machine which emulates the behaviours we see in animals or other humans.

Evolutionary Robotics gives us a methodology explicitly based on such behavioural criteria. As practised at Sussex we adopt a Dynamical Systems approach to the 'stuff' from which robot control systems are built (Harvey et al., 1997). This means that behaviour is derived from the way an organism is coupled with its environment. However, theories based solely on behavioural criteria leave out what Chalmers calls the hard problem of consciousness.

Chalmers' attempt at a solution is to assert that consciousness* must be a fundamental entity of the universe in the same way that mass or charge are. I agree with Dennett that there are no other entities over and above consciousness$^{1-3}$, but this still leaves for most people a sense of dissatisfaction — isn't Dennett denying phenomenal experience, implying that it does not exist?

From a relativist phenomenological position I would assert that indeed we do (of course!) have phenomenal experience, but this is not a 'thing that exists' in the sense that matter and charge, indeed tables, and apples, exist. Phenomenal experience is primary, and through our experience we construct matter, charge, tables, apples as objects that allow us to make sense of the world. Though in many scientific disciplines one can get away without stating explicitly 'entity E exists from the perspective of those specific observers', as contrasted with 'entity E exists for us specific observers', the various Copernican revolutions have come about through rejecting (e.g.) absolute velocities in favour of relative velocities — 'velocities from the perspective of A or B'. If cognitive science follows the more basic sciences in accepting a relativist revolution, then the common philosophical puzzlements in relation to consciousness will just dissolve.

It follows that the creation of a robot which, for us, has the same forms of consciousness, even consciousness*, as you or me, does not have any 'difficult' hurdles to cross, in Chalmers' sense of 'difficult'[3]. The so-called 'easy' hurdles, however, will no doubt need many centuries, indeed millenia, of hard work.

[3] To speak of a robot or person *having* consciousness* is a potentially dangerous form of words, as it could be taken, misleadingly, to imply that consciousness* is a 'thing'.

We will, in practice, only want to attribute consciousness* to robots that we can see have their own concerns, and needs — so that objects exist-for-them. There are at least two possible reasons to suggest that an evolutionary robotics approach may be particularly appropriate to achieve this end. Firstly, it is easy to get needs and wants into such robots without explicitly programming them in, thus avoiding the GOFAI trap of the yawning chasm between an internal rule named by the programmer `robot_avoid_obstacle` and the robot actually *wanting* to avoid the obstacle. Secondly, the evolutionary approach will produce control systems that we cannot analyse — indeed, for me a major motivation for this method is that it allows us to produce systems more complex than our shallow understanding can cope with. It follows that a mechanistic understanding of such systems will not be available to us in practice, only in principle. Since that perspective on the Necker cube, that interpretation, is not available to us, we will in practice adopt the other natural interpretation at the behavioural level of description. It will be much easier for us to treat such robots as conscious*.

Acknowledgments

I thank the SERC and the University of Sussex for funding, and Shirley Kitts for philosophical orientation.

References

Beer, R. D. and Gallagher, J. C. (1992). Evolving dynamic neural networks for adaptive behavior. *Adaptive Behavior*, 1(1):91–122.

Brooks, R. (1991). Intelligence without representation. *Artificial Intelligence*, 47:139–159.

Chalmers, D. (1995). Facing up to the problem of consciousness. *Journal of Consciousness Studies*, 2(3):2–19.

Dennett, D. (1987). *The Intentional Stance.* MIT Press, Cambridge MA.

Dennett, D. (1996). Facing backwards on the problem of consciousness. *Journal of Consciousness Studies*, 3(1):4–6.

Foerster, H. V. (1984). *Observing Systems.* Intersystems Publications, Seaside, California 93955.

Harvey, I., Husbands, P., Cliff, D., Thompson, A., and Jakobi, N. (In Press, 1997). Evolutionary robotics: the Sussex approach. *Journal of Robotics and Autonomous Systems.*

Husbands, P., Harvey, I., and Cliff, D. (1995). Circle in the round: State space attractors for evolved sighted robots. *Journal of Robotics and Autonomous Systems. Special Issue on "The Biology and Technology of Intelligent Autonomous Agents"*, 15:83–106.

Maturana, H. and Varela, F. (1987). *The Tree of Knowledge: The Biological Roots of Human Understanding.* Shambhala Press, Boston.

Penrose, R. (1989). *The Emperor's New Mind.* Oxford University Press.

Thompson, A. (1995). Evolving electronic robot controllers that exploit hardware resources. In Morán, F., Moreno, A., Merelo, J.J. and Chacón, P. (eds.) *Proceedings of Third European Conference on Artificial Life.* Springer-Verlag, pp. 640–656.

van Gelder, T. (1992). What might cognition be if not computation. Technical Report 75, Indiana University Cognitive Sciences. Reprinted in *Journal of Philosophy* 92:345–381 (1995).

Varela, F., Thompson, E., and Rosch, E. (1991). *The Embodied Mind.* MIT Press.