

# WHAT IS IT LIKE TO BE A ROCK?

**Aaron Sloman**

**School of Computer Science,  
The University of Birmingham  
Birmingham B15 2TT  
England, UK**

**A.Sloman@cs.bham.ac.uk  
<http://www.cs.bham.ac.uk/~axs>**

Originally written Jan 1996

Added URLs 7 Jan 2009

## **Abstract**

This paper aims to replace deep sounding unanswerable, time-wasting pseudo-questions which are often posed in the context of attacking some version of the strong AI thesis, with deep, discovery-driving, real questions about the nature and content of internal states of intelligent agents of various kinds. In particular the question ‘What is it like to be an X?’ is often thought to identify a type of phenomenon for which no physical conditions can be sufficient, and which cannot be replicated in computer-based agents. This paper tries to separate out (a) aspects of the question that are important and provide part of the objective characterisation of the states, or capabilities of an agent, and which help to define the ontology that is to be implemented in modelling such an agent, from (b) aspects that are incoherent.

The paper supports a philosophical position that is anti-reductionist without being dualist or mystical.

## **Contents**

<b>1</b>	<b>Introduction</b>	<b>2</b>
<b>2</b>	<b>Strategy</b>	<b>2</b>
<b>3</b>	<b>Preliminary conclusion</b>	<b>8</b>
<b>4</b>	<b>Some philosophical objections</b>	<b>9</b>
<b>5</b>	<b>Do computers have a point of view?</b>	<b>9</b>
<b>6</b>	<b>More philosophical objections</b>	<b>10</b>
<b>7</b>	<b>Flipping qualia</b>	<b>11</b>
<b>8</b>	<b>Semantic traps</b>	<b>12</b>
<b>9</b>	<b>Incoherence is not the same as lack of meaning</b>	<b>13</b>
<b>10</b>	<b>Conclusion</b>	<b>13</b>

# 1 Introduction

Discussions of consciousness, including whether machines can be conscious, which animals are conscious, how consciousness evolved, what the function of consciousness is, etc. are often based on the assumption that we all know what we mean by the word “consciousness”. Moreover, it is assumed that what we mean can be identified by various words and phrases which are part of common usage, but which become elevated to pseudo technical terms for the purpose of bolstering philosophical or scientific discussions of consciousness. A much discussed example is this form of phrase: “What it is like to be an X?” (Nagel 1981). This recurs often, for example, in discussions in the comp.ai.philosophy newsgroup.

My own belief is that much discussion of consciousness is based on a highly inflated conception of the clarity of the questions being posed, and the objective of this paper is to deflate such discussions. I don’t deny that there are questions to be asked, but I claim that we need a new route to such questions, based on architectures for various kinds of intelligent agents, and analysis of the sorts of states and processes various architectures can support. Unfortunately, the temptation to ask pseudo-questions remains very strong. The purpose of this paper is to help to reduce that temptation. It will not work for all readers.

## 2 Strategy

My strategy is in part to consider this question

### **What’s it like to wonder what it’s like to be an X?**

I shall attempt to analyse what sorts of requirements there are for satisfactory answers, by considering a range of cases, from several viewpoints, and then end with a philosophical position which is both anti-reductionist and functionalist.

### **What is it like to be that rock over there?**

Well, I don’t know the whole answer, but, unlike the rock, I know a lot of it.

It is like being about a foot in diameter.

It’s like weighing a few pounds.

It is like being made mostly of silicon (I think).

It is like resting on a muddy patch of earth with a slight slope.

It’s sometimes like being pushed around, thrown up into the air, and falling to earth with a thud.

But never like knowing any of this is happening.

Some will object that I’ve distorted the question, for I’ve wrongly taken “What is it like to be X?” to ask only: “What is X like?” The former presupposes that X has a point of view and requests a description from X’s point of view, whereas the latter does not require X to have a point of view. Of course that merely shifts our problem: what is a point of view, and which sorts of things have them? Does a sunflower have a point of view?

## **What is it like to be a sunflower?**

I don't know as much about this, as about what it's like to be a rock.

It's like being able to grow bigger and being able to produce roots to find nourishment and support.

But not like being able to walk or run.

It's like having information about which way up is and the direction of the sun, but not information about very much else, at least in the environment.

Nor can the sunflower do very much with the information that is available to it. For example, it's not like being able to use the position of the sun to decide that it's time for the children to go to school.

Neither does being like a sunflower include knowing very much about the differences between up and down, or the differences between the sun and the moon, or why it might be useful to keep facing the sun.

It has a viewpoint in the sense that a viewpoint is a location in the world which provides information about the world. Different information is obtainable from different viewpoints. But the question is not just about the physical or geometric properties of the viewpoint, but also about what information the sunflower (or whatever we are talking about) actually acquires and uses, and how it uses it. That's a topic for biologists to investigate. It may be difficult, and our knowledge at present is only partial, but it involves no intrinsic mystery.

## **What is it like to be a bat?**

That's probably much more fun - whizzing about at high speed, even in the dark.

A bit like being on a roller coaster, but much more in control. E.g. rapid changes of direction will cause rapid shifts in the magnitudes and directions of inertial forces to be detected and overcome.

It's a bit like being able to make and hear sounds too high pitched for human ears. But it's not much like being able to sing a Mozart soprano part or enjoy hearing one.

It's like hearing how far away something is, but not like hearing how big a room is by clapping hands.

It's sometimes like hanging upside down for long periods but without experiencing torture.

There's much more we can say about what it's like to be a bat than what it's like to be a sunflower, because a bat does far more, and also because it is far more like us, which tempts us to extrapolate and use our own descriptive categories.

On the other hand, what it's like to acquire, process, store, or use information as a bat does is not something we can hope to understand in any detail – for the bat does not use anything like our conceptual apparatus, as far as I know. This is a point Nagel makes, though he gives no explanation.

The explanation may be that the information processing media and mechanisms used by the bat have different structures and possible transformations from ours, and there are no ways of mapping its states and processes into ours without serious distortion of the structures and relationships.

There may be some small overlaps, e.g. to do with the bat's and our ability to cope with space, and time and motion, but these are embedded in very different webs of relationships to other things, for instance the shape of an attractive mate, or tempting morsel.

Its world is not our world, and there's no reason to believe that its categorisations of things, states,

events, processes, actions, or whatever replaces such things in its control mechanisms, will map in any straightforward way onto the categorisations we use.

In short, to misquote Wittgenstein, if the bat could tell us what it is like to be a bat we would not understand it.

There are at least two different reasons for this: first its requirements and its relations to the environment are very different from ours (we don't often fly through the air chasing and eating fluttering moths), and secondly its processing engine may be different from ours in subtle ways. For instance, it may be wholly incapable of being in as many different types of states, and it may not have the same variety of functionally distinct, coexisting, interacting components.

Its representational grammars are different, and hence their semantic capabilities are different. (Cf. Sloman 1994, 1996).

Some languages cannot be learnt if you don't have the right sort of engine to run them on. Obviously a simple information processing engine may be incapable of replicating semantic states of more complex ones. The converse is less obvious, but can also be true.

Despite the impossibility of *translating* bat information states into our own there is nothing, in principle, to stop us producing fairly detailed descriptions of the kinds of structures and transformations of such states. For example, we might discover that its sound-processing capabilities allow it to produce information states that can vary in 17 dimensions, or that its representations of the structures of other bats distinguish 37 bat components and use 5 types of relationships between those components.

A full theory of bat semantics might require is to extend our logical metalanguage, however: maybe bat cognition does not use objects, properties and relationships, but only interactions between attractors in phase spaces. (Cf. Cohen & Stewart 1994, Sloman 1995)

Of course, discovering such things may be extremely difficult. Moreover, although it would give us a partial answer to the question "What is it like to be a bat?" it will not enable us to *experience* bat-like states. You cannot necessarily experience, or even imagine experiencing, everything you can describe, for instance, being a sixteen dimensional dancer in a forest of nineteen dimensional shapes.

This is perhaps an explication of the possibility Nagel raises towards the end of his article, of an objective description of subjective states. Note that such objective descriptions need not be descriptions in the language of physics. They are more likely to be descriptions in the language of linguists and computer scientists, who talk about the properties of information structures independently of their implementation, even though the implementation is always physical.

### **What's it like to be a (normal) new-born human infant?**

It's like being simultaneously too weak, too uncoordinated and too ill-informed to be able to cook your own supper.

But it's not like knowing you are all those things, for the infant (probably) knows nothing about being weak, coordinated or informed, or about cooking or supper.

It's a bit like being able to see, to hear, to feel hunger and pain, though via processes that do not make use of typical adult human concepts and which we are therefore currently unable to describe in detail. If a neonate could talk, we'd understand only a little more than if the bat spoke. (As puzzled parents will confirm.)

One of the striking features of the human mind is that it changes. Many of the changes may seem gradual at the time, but the cumulative effects can include major discontinuities, including the creation of new information structures (e.g. learning about quantum mechanics) and new control architectures (e.g. becoming able to postpone gratification). We know very little about such developments, not even how much is pre-programmed genetically and how much driven by interaction with the environment. Certainly the environment plays an important part, since not everyone learns to speak Chinese, or read music, or grasp quantum physics.

So there's enough diversity in adult human information states to make full translation between them impossible. The difference between what I can know of what it's like to be a bat, and what it's like to be you is only a difference of degree.

Not all the development is growth or improvement, alas.

### **What's it like to be in the advanced stages of Alzheimer's disease?**

This something about which I know very little apart from the sort of thing picked up by watching a sad and moving television documentary.

For example, in some cases it's apparently like wanting to dust a window-ledge without remembering that you have dusted it already a minute or two before.

### **What's it like to be autistic?**

I recently (November or December 1995) watched the second half of a spellbinding television program (UK TV Channel 4 I think) about an autistic woman who had become very articulate and had written a book. It's very clear that what it's like to be her is somewhat different from what it's like to be a 'normal' person.

Sloman (1989)<sup>1</sup> referring to (Self 1977), conjectures that the spectacular drawing ability of Nadia, an autistic child, might be a result of abnormal concentration of processing resources on low level image analysis and interpretation because higher level integrative and interpretative mechanisms which would dominate normal vision were non-functioning for some reason.

If so, processes that in normal brains work to minimise or cut short low level analysis and interpretation, for the sake of speedy high level recognition and decision-making, are not available in some autistic brains. So the low level processing dominates, and the higher level more abstract and holistic interpretations therefore do not occur, or are simplified.

The autistic woman tried to describe her experiences shopping in a supermarket, and various camera tricks attempted to give the viewer a taste of what it was like: lots of small scale, low level, rapidly changing detail, from which it was hard to synthesise an overview. Of course, a normal person would get an overview such as I've just described, but that is not the same as getting an integrated visual overview. Camera tricks probably don't actually convey what it's like to be autistic: they merely *change* the high level synoptic characterisation, without suppressing it altogether.

So here's another case where we may be able to go a long way towards describing the nature of a (partly) alien form of experience, while being incapable of having or imagining that experience. Description is always easier than replication.

A person who is not autistic could get quite a lot of information about what it's like to be autistic

---

<sup>1</sup>Online here <http://www.cs.bham.ac.uk/research/projects/cogaff/81-95.html#7>

by watching the film. In fact I found that her unusual use of language and explanatory constructs and illustrative models (moving toy animals around on a table top) temporarily changed me in a strange way while I was listening to her – a reaction to the extraordinary mixture of familiar and unfamiliar features of both her way of speaking and the content of what she said. But most of what I got was description, not replication, of her experiences.

## **What is it like to be a seer?**

You don't need to be like X to know a lot about what it's like. A congenitally blind person can know a lot about what it's like to see.

It's like getting information about the surfaces of things – their textures, orientations, locations in space – but without touching them.

Also it's more parallel than the serial exploration of the outside of an elephant with your hands. It's more like simultaneously exploring the whole shape of a cup with several fingers on two hands, but not including the 'far' surfaces.

Much of the information obtained by visual and tactile perception is the same, though there are subtle differences in its availability for various uses: e.g. rapid comparison of two faces, or searching for a family resemblance. If I could only see my bow arm moving, and not also feel it, the sound of my violin would be even worse.

Seeing colours is often very like detecting the textures on surfaces - there are colour and texture regions, colour and texture boundaries, colour and texture gradations, colour and texture composition, 2-D shapes such as letters or polygons made by regions of a common texture or colour, etc.

What it's like to see colours is partly like getting all that surface information very rapidly and at a distance, and also having it presented to you in a spatially structured way (i.e. not as a long list of sentences).

What does that leave out? Well, quite a lot, and I could describe some of it in terms that a sighted reader would understand but a blind person might find difficult.

On the other hand, we should remember that a congenitally blind person whose eyes have never worked may still have a lot of the brain mechanisms used by the rest of us in seeing colours, developed in a shared evolutionary history. Perhaps such a person can use some sort of abductive inference from hearing other people talk about colours and other spatial properties, and this might recreate similar structures to those used by sighted people in comprehending talk about colours. The important point is that it's an *empirical* question what a congenitally blind person is capable of understanding. And that depends on what information processing capabilities are in his brain.

Blind people given a stylus with a hot pointed tip for drawing on plastic can produce pictures with some of the structure, e.g. the topology, of pictures drawn by sighted adults. However, angles and other relatively global metrical properties are different. That's also true of pictures drawn by young children whose vision is perfectly normal.

It might be true of a robot whose visual system is designed to give it information about 3-D structures in the environment, not about the 2-D structure of the optic array at a viewpoint. The fact that that 2-D structure is there in its visual system and is used by low level visual procedures does not mean that the robot can consciously access it or use the information to produce an external 2-D drawing. What it's like to be that robot could include inducing 3-D structure from various intermediate 2-D shapes, without knowing anything about the 2-D shapes, just as we are ignorant

of the vast majority of what's in our own minds.

What it's like to be able to see is not necessarily what it's like to be able to draw what one sees. The information processing requirements are very different. Similarly, being able to see need not involve the ability later on to recreate accurately a detailed internal replica of the experience. When we remember seeing things normally only a small subset of the original state is reproduced (though some people, like Nadia, can store and reconstruct more details than others).

### **What is it like to be a woman?**

No doubt there's a lot I'll never know. I know what it's like to find a woman sexually attractive, and on the basis of that I think I know at least some of what it's like to find a man sexually attractive, but not all. So perhaps I know more about what it's like to be a lesbian than a 'straight' woman.

I don't have direct experience of feeling 'broody', i.e. desperately wanting to produce a child, but I know quite a lot about what it's like, from talking to some who have. I think I know what it's like to be consciously or unconsciously discriminated against or patronised, in a male dominated society, though I have not had that sort of experience myself.

Besides common differences between infants and adults, between brain damaged or senile individuals and those who are still functioning normally, there may also be deep and subtle differences in cognitive functioning between adults, based on *genetically* determined differences in information processing mechanisms in their brains.

Whether this is so or not, and whether it explains such things as the rarity of musical or artistic or scientific genius, it is clear that there are some differences that are related to gender, and for good reasons: without them the gene complexes constituting human beings would be much less successful at replicating themselves. (Some would argue that at present they are far too successful either for their own good or the good of other organisms, and a change is long overdue.) Thus an intelligent robot could not simultaneously be exactly like a woman and like a man.

### **What is it like to be a robot?**

Obviously it's going to depend a lot on the type of robot.

A robot that has only a tiny fraction of our bodily sensors will never have the full and rich experience of riding in a roller coaster, feeling all the centrifugal and gravitational forces, feeling the rush of wind on face and in hair, hearing all the screams, feeling the closeness of the excited and terrified child clinging to it, or the dryness of the mouth that can go with one's own terror.

But it might know a lot about what it's like for human roller coaster riders, by using what it knows about the whole situation and about human sensory capabilities, motivation, emotional states, etc. In other words, it may have a great deal of knowledge about the information processing capabilities, the forms of representation, and the cognitive functions of human beings, even if it is not able to replicate most of them within itself.

So what it's like to be a sophisticated robot with a body very different from ours might include knowing quite a lot about what it's like to be a human being, though not everything. The robot could well know a lot more about what it's like to be a human being than what it's like to be a bat, and for exactly the same reasons as our bat-like knowledge is limited:

- (a) lack of information
- (b) wrong conceptual apparatus to be able to replicate bat-like information processing.

It's often claimed (by Nagel, and many others) that a robot might simulate many of the behavioural capabilities of a human being without having anything remotely like the conscious states of a human being, or a bat, i.e. without there being anything that it is like to be the robot.

But this ignores the fact that a robot of the sort in question could not function without sophisticated information processing mechanisms, and we can then ask questions about the internal states and processes concerned with the information processing level (Sloman 1994) which are *not* questions about its physical state. We can ask what information it has about various objects in the environment, or about itself, or about its own internal states (McCarthy 1995).

Some robot states we may be unable to replicate in ourselves, as with bat states. But we may nevertheless be able to describe them, that is describe their structures, their semantic contents, their transformations, and the uses to which they are put. Any such information processing system will necessarily have as much of a viewpoint as you or I, or a bat. Whether we can experience the same viewpoint or not, we can talk about it and the role it plays in enabling the system to function. There is no reason to suppose that such a properly functioning robot could be a 'zombie', defined as something that merely produces behaviour without having any experiences, viewpoint, or something it's like to be. People who claim they can imagine such a thing are deluding themselves, like people who think they can imagine a method of accurately trisecting any angle using only ruler and compasses. They lack adequate training (in this case engineering training): such a robot could not work.

For a good engineer to imagine a zombie robot she have to turn off what she knows about what enables the robot to process perceptual information, to acquire new goals, to select between goals, to be inclined to continue with some activities and desist from others, to detect achievement of its goals, or cope with unexpected and sudden external disruption.

I am making strong claims about the high level information-processing ontology of any system with human-like intelligence. But I need to be careful: it may be theoretically possible to have a huge lookup table in which all possible sequences of sensory input have been previously stored and appropriate external responses provided. If this were physically possible (which it isn't for most human-like capabilities), then such a machine would not require a human like internal ontology. It would be much more like a rock than like a bat. If you want to imagine a zombie, you had better design one that could work, as I've just done.

(It follows from the above that feeding a design through a heavily optimising compiler that unfolds loops and conditionals and flattens subroutine calls, could produce a zombie implementation, indistinguishable externally from a sentient, thoughtful robot. In this sense implementation can matter.)

### **3 Preliminary conclusion**

I have tried to show, by considering different cases, that there are many things we can know about what it is like to be an X, and things we cannot, and the reasons are different in different cases. There is no one deep philosophical reason. However, some aspects of what it is like to be an X cannot always be replicated in a Y, where Y has a different information processing architecture. When understood right that's a theorem of information science, not a philosophical problem!

## 4 Some philosophical objections

Many philosophers will agree with all that. But, some philosophers (and some non-philosophers) will protest, none of this adds up to telling me what it is like to be an X, FROM THE INSIDE.

It's not at all clear what sense the capitalised words add, or whether there is any coherent sense that they add. They do make some sense, which I'll now explain. But some people will feel I've still left something out. I'll offer them therapy later.

I can certainly talk about what it is like to be you, located where you are, seeing things that only you can see, feeling things which I don't (e.g. because you are sitting in a comfortable armchair and I am squatting on the floor), and knowing things about your own state of mind which I can at best guess.

If what it's like to be you 'from the inside' means something about what it's like to have your view of the world, including your view of the current state of your own mind and body, then I can't know exactly what it's like to be you because I lack much of the information.

Some of it I may be able to work out. I infer that you see part of that wall which a large pillar obscures from me, and I can also work out that you know whether my remark really made you angry or whether you are just teasing me. You know, but I don't, but I know that you do.

There's no huge difference between knowing that there's something outside you to which you have access I lack and knowing that there's something inside you to which you have access which I lack. Both cases involve differences of access to information.

Of course, I can remove one of the differences by walking round to where you are, and then I'll see the previously invisible part of the wall.

But I can't walk round to a place where I'll have your view of your current emotional state.

Moreover, if you are colour blind and I am not, or vice versa, I won't even have exactly your view of the fresco.

Maybe one day, a pair of special helmets, linked by optic fibres, may overcome this obstacle to sharing your viewpoint, just as my walking closer to you, or using appropriately placed mirrors and cameras can overcome some of the lesser obstacles.

Whether such a helmet is possible is an empirical issue: there's nothing philosophically deep about it. (Not as deep perhaps as the difficulty of knowing what's going on inside subatomic particles, or whether 'inside' even makes sense in their case: now there's a hard problem if you want one.)

## 5 Do computers have a point of view?

Currently, computing systems are very different from ourselves. They don't have anything like our information processing architecture – for instance, their long term memories function in very different ways (the problems of acquiring, accessing and forgetting things are totally different). They (mostly) lack anything remotely like human motivation, and those that do have goal processing capabilities generally get their goals from someone else, whereas human beings contain multiple independent sources of motivation which (generally) do not subserve anyone else's goals. (Human motive generators do not always subserve the individual's goals, interests or needs, for instance if the generators were selected by evolution because of their contribution to promotion and survival of the gene pool. Our genes would laugh at us for some of the things they make us do, if genes could laugh.)

Because of these differences between computing systems and people (or other animals), our knowledge of what it's like to be a timeshared computing system is not all that different from our knowledge of what it's like to be a rock: it's mostly "third person" information about what's going on, and perhaps only a little less boring.

However, as AI systems get more and more sophisticated with more and more autonomy both in motivational mechanisms but also in development of conceptual apparatus controlling perceptual processes and determining the semantics of internal information stores, then there will be more and more questions we can ask about what it's like to be such a system 'from its viewpoint', e.g. what sorts of things can it see in various situations, what sorts of things does it want to do in various situations, what sorts of decisions does it face, what does it find easy or difficult, pleasant or unpleasant, etc.

I.e. there will be more and more about what it's like to be that sort of robot FROM THE INSIDE. Whatever that means.

Of course this will take a long time – maybe hundreds of years before they get to chimp-like ways of being and still longer before they reach human-like states.

When that happens, some of the robots will start wondering what it's like to be a person, or a bat. Or maybe even a rock.

## 6 More philosophical objections

"BUT, BUT, BUT" splutters the frustrated objector, growing ever more exasperated with me, "You still have not got around to what I was talking about: what it's REALLY like FROM THE INSIDE to be a bat or a person. All you are talking about is what it's like to have a variety of physical properties and states and processes and also some non-physical (though physically implemented) information processing capabilities and states and processes.

"That's not what I was referring to in talking about what it's like to be X. The things you are referring to are all things for which there might be objective evidence, e.g. evidence from X's relationship to things in the environment (determining which things are not currently visible to him), or evidence from what the designer knows about X's high level functional architecture (determining which kinds of visual processing, memory processes, goal generation, planning, reasoning, forgetting, etc. etc. can occur in X).

"I am talking about the INTRINSIC quality of what it's like to be X, which cannot be inferred from these things.

"E.g. two sub-aspects of what it's like to be X (in my sense) might be swapped without this having any externally detectable effects or causes, and without it making any difference to the functional capabilities to see, reason, plan take decisions, etc. For instance, what it's like to see the colour of the sky and what it's like to see the colour of grass might be swapped in X, and everything else could remain the same. X could tell us that something strange had happened, and that sky and grass each now looked the colour the other used to look. But he could not tell us what that colour was. He'd have no way of telling, for example, whether his new view of coloured things was the same as other people's or his old view, or neither."

Let's be clear about this: some of what's being said here is OK and some not. The interplay between the two is very subtle and it is difficult to separate them. But we must separate them to clarify requirements for the design of a human-like intelligent agent. Human like, non-zombie robots should be capable of experiencing Necker flips.

## 7 Flipping qualia

Consider what happens when you look at a Necker cube: suddenly it flips, and although the retinal image and visible 2-D structure are unchanged, the 3-D interpretation is different. Lines, or rather cube-edges, that once sloped down away from the viewer now slope up away from the viewer, and the vertical square cube face that was previously further away is now nearer. Perhaps there could be a sort of Necker flip of colours, with visible sub-regions switching how they look, just as parts of the cube switch (some aspects of) how they look?

Whether this is possible is an empirical question.

In the case of the cube nobody (so far) has claimed that the switch is undetectable from outside and indescribable to others. Indeed brain scientists may one day find out exactly what sorts of neural processes are involved in the flip from one state to another, and may even be able to create non-invasive mechanisms for detecting the occurrence of such a flip.

Moreover, it is more than likely that one day robot vision systems will be capable of such flips: in fact requirements for normal vision include the ability to handle locally ambiguous fragments of images which can have different interpretations fitting into different coherent global scenes. (What is seen as a leg in one context may be seen as an arm in another).

So the Necker flip in what it's like to see the pattern of lines as a cube is (a) part of the expected behaviour of functional components of a visual architecture, (b) capable of being explained in terms of underlying neural or computational mechanisms, (c) capable of being detected from outside (at least in principle), (d) capable of being described (in terms of changing geometric relations between parts of the cube), (e) likely to occur in visually sophisticated robots, under appropriate conditions.

Could there be flips in how colours appear too?

How you see the colour of a portion of a surface of an object can change according to the context, as shown by a variety of visual so-called "illusions", including for example the Kanizsa figures in which we see colour boundaries on a uniform surface. Another case is an array of black squares on a white ground, in which we see dark patches between the squares, which disappear when we look directly at them. Perhaps there are some experimental situations to be discovered at some time in the distant future, in which everything that now looks green suddenly looks red and vice versa, and this will be explained as part of a natural side effect of how a fully functioning visual system implemented in a certain way works.

Such a colour flip might have all the features (a) to (e) that the Necker cube flip has.

But none of that fits the intended philosophically puzzling situation: for there the alleged logically possible flip is supposed to be totally unrelated to anything functional, incapable of being explained in terms of underlying mechanisms, incapable of occurring in any robot, incapable of being detected from the outside, and incapable of being described.

At this stage we (or rather our worried friends) have begun to reach one of those ancient philosophical traps: words and phrases which sound as if they are saying something clear, and rich, and

fascinating, but which say nothing at all.

Or rather they say nothing capable of being the content of a true or false assertion or a question with right and wrong answers.

## 8 Semantic traps

Here are some old, well-known examples, of how one might fall into such traps:

“Clearly it’s (logically) possible for this cup to move a foot to the east, and then my spectacles to move a foot to the east, and then the table, and then the house, and then the earth and then the sun and then the rest of the solar system, and, and, and ... until everything in the whole universe has moved a foot to the east. (There might have to be some temporary changes in the laws of physics while these various steps occur, but that’s also logically possible.) And at the end of such a process EVERYTHING in the whole universe would have moved a foot to the east, but that new situation would be TOTALLY indistinguishable from the original state.

“So how can we be sure it hasn’t happened, and our memories tampered with so that we don’t remember any of our changes? And maybe it’s happening all the time, with everything slowly moving to the east, but no motion detectable because all the measuring instruments are also moving, and the laws of physics have been carefully adjusted to ensure that no experiment will reveal the motion. (Compare the Michaelson Morley experiment which failed to detect which way the earth was moving through the ‘aether’.)”

If you really think the hypothesis just described makes sense, and that there may be motion of the whole universe that’s totally undetectable, then you may as well stop reading, for I have no way of convincing you that you are deluding yourself. You need stronger therapy than I can offer.

It’s one of the features of being like a human that such delusions can be very tempting, and in some cases incurable. Similarly, if you are tempted to wonder whether it really is noon at the centre of the moon when the moon is directly above Greenwich and it’s noon at Greenwich, then that temptation may be incurable, no matter how much I try to convince you that the question is too ill-defined to have an answer, any more than the question whether the number nine is green or yellow.

The notion that the colour experiences you have when looking at grass and sky might suddenly be swapped in such a way that absolutely everything else, or everything else that someone else could observe, measure, control, etc. remains the same, is as coherent as the notion that this change is happening all the time, only you don’t notice because your memory is constantly being fixed so that you forget what the colours really were like.

And that’s as coherent as the notion that all sorts of pairs of experiences are CONTINUALLY being swapped in such a way as to be totally undetectable by you and in such a way as to preserve all functioning aspects of the system.

And that’s as coherent as the question whether the universe is constantly moving to the east, or the north, or the north north west, at three miles per hour, or three cm per hour, or at any other speed, but in such a way that the motion cannot be detected. See (Dennett 1991) for related arguments.

## 9 Incoherence is not the same as lack of meaning

Of course these words and phrases are not MEANINGLESS. They *resonate* with rich meanings. The semantics of the components combine to form a rich structure, which drives many of the processes of reasoning, question formation, wondering whether, and the like which are the very stuff of cognition in science and everyday life and philosophy.

But they may still fail to add up to anything with the properties required of a serious question or hypothesis, much as the fragments of the image of an impossible Penrose triangle all make sense, and globally they make a kind of sense, but fail to add up to a possible continuous, straight-edged, 3-D object. In the case of the triangle most people can see the impossibility fairly quickly. But when it comes to an impossible hexagon or an impossible 63 sided object, constructed in the same way, no human will be able to see the impossibility without very laborious checking. It will simply look like a complex 3-D object, especially if it is not presented as a regular polygon.

Similarly lots of people fail to see the incoherence of many of their philosophical questions and descriptions of allegedly possible scenarios.

Similar things underlie religions and many superstitions.

## 10 Conclusion

Asking what it's like to be an X, when taken seriously, so that we knuckle down and try to work out detailed answers, can be part of the process of uncovering some of the internal ontology, the virtual machine requirements, the abstract data-types and operations on them, which we need to think about if we are to design human like systems, whether as engineers building new slaves or playmates, or as scientists trying to understand by designing and implementing. So, far from being part of an anti-AI activity it is a necessary part of AI in the long run. (How to do it is another question, for another day.)

Of course, many philosophers will be unmoved by all this. They will accuse me of being a crude empiricist, a verificationist, an out of date positivist, a science worshipper, a zombie in disguise, a zombie-promoter, a heathen, a liar ("feigning anaesthesia", as one philosopher once put it), or worse.

And there's nothing I can do to prove them wrong. For when the human brain gets trapped into the state of believing that certain kinds of grammatically well formed sentences actually *mean* something to it, there's no rational argument that can change the situation. Sometimes long term philosophical therapy works, and sometimes it doesn't.

I say all this because I know what its like to think one understands what one is saying when one puts forward those arguments and asks those questions. I know what it's like to be there. I've been there.

## REFERENCES

J. Cohen and I. Stewart *The collapse of chaos*, Penguin Books, New York, 1994.

D. C. Dennett, *Consciousness Explained* Penguin Press, Allen Lane, 1991,

J. McCarthy, Making robots conscious of their mental states, *AAAI Spring Symposium on Repre-*

*sending Mental States and Mechanisms* Stanford, 1995, Accessible via  
<http://www-formal.stanford.edu/jmc/>

Thomas Nagel What is it like to be a bat, in *The mind's I: Fantasies and Reflections on Self and Soul* Eds D.R. Hofstadter and D.C.Dennett Penguin Books 1981, pp391–403 (Followed by commentary by D.R.Hofstadter, pp403–414.)

Selfe, Lorna *Nadia: a case of extraordinary drawing ability in an autistic child* London, Academic Press, 1977.

A. Sloman, 'On designing a visual system: Towards a Gibsonian computational model of vision' *Journal of Experimental and Theoretical AI* 1,4, 289-337 1989

<http://www.cs.bham.ac.uk/research/projects/cogaff/81-95.html#7>

(Also available as Cognitive Science Research paper 146, University of Sussex).

A. Sloman, 'The mind as a control system', in *Philosophy and the Cognitive Sciences*, (eds) C. Hookway and D. Peterson, Cambridge University Press, pp 69-110, 1993 (Supplement to *Philosophy*) <http://www.cs.bham.ac.uk/research/projects/cogaff/81-95.html#18>

A. Sloman, (1994) Semantics in an intelligent control system, in *Proc. British Academy and Royal Society Conference: Artificial Intelligence and The Mind: New Breakthroughs Or Dead Ends?* in *Philosophical Transactions of the Royal Society: Physical Sciences and Engineering*, Vol 349, 1689, pp 43-58 1994

<http://www.cs.bham.ac.uk/research/projects/cogaff/81-95.html#25>

A. Sloman, (1995) Musings on the roles of logical and non-logical representations in intelligence, in Janice Glasgow, Hari Narayanan, Chandrasekaran, (eds), *Diagrammatic Reasoning: Computational and Cognitive Perspectives*, MIT Press, 7–33

<http://www.cs.bham.ac.uk/research/projects/cogaff/81-95.html#33>