To appear in International Journal of Machine Consciousness in 2010

Phenomenal and Access Consciousness and the "Hard" Problem: A View from the Designer Stance

Aaron Sloman
School of Computer Science,
University of Birmingham, UK
http://www.cs.bham.ac.uk/~axs/

DRAFT 18 Jan 2010
Revised May 14, 2010

NOTE: [a]

This paper is an attempt to summarise and justify critical comments I have been making over several decades about research on consciousness by philosophers, scientists and engineers. This includes (a) explaining why the concept of "phenomenal consciousness" (P-C), in the sense defined by Ned Block, is semantically flawed and unsuitable as a target for scientific research or machine modelling, whereas something like the concept of "access consciousness" (A-C) with which it is often contrasted refers to phenomena that can be described and explained within a future scientific theory, and (b) explaining why the "hard problem" is a bogus problem, because of its dependence on the P-C concept. It is compared with another bogus problem, "the 'hard' problem of spatial identity" introduced as part of a tutorial on semantically flawed concepts. Different types of semantic flaw and conceptual confusion not normally studied outside analytical philosophy are distinguished. The semantic flaws of the "zombie" argument, closely allied with the P-C concept are also explained. These topics are related both to the evolution of human and animal minds and brains and to requirements for human-like robots. The diversity of the phenomena related to the concept "consciousness" as ordinarily used makes it a *polymorphic* concept, partly analogous to concepts like "efficient", "sensitive", and "impediment" all of which need extra information to be provided before they can be applied to anything, and then the criteria of applicability differ. As a result there cannot be one explanation of consciousness, one set of neural associates of consciousness, one explanation for the evolution of consciousness, nor one machine model of consciousness. We need many of each. I present a way of making progress based on what McCarthy called "the designer stance", using facts about running virtual machines, without which current computers obviously could not work. I suggest the same is true of biological minds, because biological evolution long ago "discovered" a need for something like virtual machinery for self-monitoring and self-extending information processing systems, and produced far more sophisticated versions than human engineers have so far achieved.

[a] Available in two formats, PDF and HTML at
http://www.cs.bham.ac.uk/research/projects/cogaff/09.html#pach

Contents

1. Introduction

When my paper "An Alternative to Working on Machine Consciousness" (Sloman, 2010) (henceforth AWMC) was made available to *IJMC* as a target for commentary, it elicited commentaries that made me realise that I had been taking things for granted that are not widely known, and had been rather quickly drawing conclusions that not everyone finds obvious. This paper is an attempt to spell out the main points I had assumed, to expand some of the arguments and distinctions so as to make them easier to understand, and to draw some new conclusions. Whether it will convince doubters I don't know, but it should at least make clearer what I am claiming and what I am not. Although an attempt is made to summarise some work by others there is no claim to completeness. I have time to read only a tiny subset of the available literature and tend to ignore anything that is not freely available online. So there may be things already in print that contradict or improve on what I have written. I welcome pointers. This paper has a companion online tutorial presentation (Sloman, 2009d).

For many years I have been criticising attempts to define, explain, or model consciousness, on several grounds. In particular the noun as ordinarily used refers to many different phenomena (because the concept is "polymorphic" as explained below), so that there cannot be a unitary explanation of how "it" evolved, or how the brain produces "it", nor a time at which "it" first exists in a foetus, nor a machine model of "it". The different phenomena falling under the polymorphic umbrella can be investigated separately, and possibly modelled separately. Perhaps one day, after the richness of the phenomena has been adequately documented, it will prove possible to model the totality in a single working system with multiple interacting components. Producing such an integrated model will require massive multidisciplinary collaboration and an agreed conceptual framework for expressing both requirements and designs – none of which is visible on the horizon as far as I can tell. Some of the requirements for such a model and draft suggestions for meeting a subset of them were proposed in (Sloman & Chrisley, 2003), and will be expanded here.

In what follows I shall try (1) to clarify some of the things I have said or written in the past[b] which were not expressed clearly enough, (2) summarise some of the key features of the phenomena that need to be explained but are not normally described in sufficient detail, (3) explain various ways in which it is possible to use concepts that can be described as semantically flawed (and in some cases incoherent[c] even though they appear to be used to make meaningful statements or ask meaningful questions, (4) show why the technical concept of "phenomenal consciousness" (P-C defined in opposition to "access consciousness" A-C) is very seriously flawed, (5) show why the so-called "hard problem" described by Chalmers as a problem about P-C is a bogus problem, whereas the allegedly "easy" problem is much harder than often supposed, and actually decomposes into a family of related problems, because the ordinary notion of "consciousness" is polymorphic, and finally (6) show how generally unnoticed features of the concept of a (running) "virtual machine" that has been developed over the last six or seven decades in computer science and software engineering (not to be confused with the concept of "virtual reality") can play a key role in explaining how certain kinds of "emergent" non-physical states, processes, and machines, all with causal powers, are possible in a physical universe.

This notion of virtual machine can inspire new ways of investigating the products of biological evolution, which I suspect include types of virtual machine that are far more complex than any built so far by engineers, and probably use types of physical information manipulation that we don't yet understand. Those more sophisticated biological virtual machines may be able to provide platforms for constructing testable explanations of some of the varied phenomena motivating current research and theories about consciousness, described below.

[b] Starting with (Sloman, 1978, Ch 10).

[c] In this paper, instead of using the label "incoherent" as generally as I have done in the past I shall instead refer to different ways of being "semantic flawed", in the interests of greater precision.

1.1. *No facts are being denied, and nothing eliminated*

Some of the criticisms, e.g. when I say there is no unitary "it" to be explained, have been interpreted as denying the existence of consciousness, even though I repeatedly state that I am not criticising the ordinary uses of the words "conscious" and "consciousness". So I should make clear now that I am not pursuing an eliminativist agenda. For example, I do not deny the existence of contents of introspection and other mental phenomena (as illustrated in (Sloman, 2009d)), though I shall criticise some of the ways they have been described and some of the theories that have been formulated about them.

There are important phenomena to be studied and explained, but they are difficult to characterise accurately, which we need to do in order to specify what exactly needs to be explained. In other words, there is a non-trivial problem of specifying *requirements* for an adequate theory or model covering a wide range of mental phenomena. My impression is that most researchers grossly underestimate the complexity of the requirements. I have been collecting requirements for about four decades, but still find them hard to organise. I don't believe current research in AI (which potentially includes research using neural nets, analog computers, evolutionary computation, chemical computation, quantum computation, hybrid machines, or any other form of information processing that looks capable of being useful for achieving the long term goals of explaining, modelling and replicating the biological phenomena) is anywhere near to the construction of good working models matching the descriptions that I shall present below. The main reason is not lack of computing power, but lack of clarity and precision about what is required.

One of the features of the phenomena that need to be explained is their *diversity*. There are differences between the kinds of consciousness that can occur in different biological organisms, differences between what can occur in a single organism at various stages of development, or after various kinds of brain damage or disease, or treatment by an anaesthetist, or hypnotist. But even in a normal adult human I shall show that the ordinary concept of consciousness is polymorphic and systematically refers to different kinds of mental function that need distinct explanations, and which probably evolved and develop at different times. The diversity and polymorphism are illustrated in Sections 2.7 and 3.3.

Some thinkers who have noticed the polymorphism, ambiguity and the subtlety of the concept of consciousness used in everyday life, try to identify a special sub-case of the concept that can be precisely and unambiguously defined, and which describes phenomena that merit special attention in philosophy and/or science. I shall call this the *special case* strategy. Block's concept of P-C described below is an example. Apart from the fact that the selection of a special case can fail to do justice to the diversity of the phenomena that need to be studied, I shall argue that some of the best known attempts to characterise such a unique topic of research on consciousness have semantic flaws (or incoherence) of various kinds, distinguished below. I focus only on some of the most widely discussed proposals. So there may be other proposals that I have not encountered that avoid the problems.

I shall try to explain why the noun "consciousness" as used in everyday non-scientific contexts, or as used by philosophers and scientists, identifies no unique kind of state, process or mechanism, that could be a topic of scientific investigation, nor something that can be modelled by some particular consciousness mechanism. Despite this I do not object to talk of consciousness in many everyday and medical contexts, and I shall try to explain this apparent contradiction. There is no contradiction once the polymorphism and context-sensitivity of the ordinary usage are understood. I am not trying to offer a new definition of "consciousness" (or of "intelligence"). I have no interest in such definitions, though I think there are useful things to be said about how the adjective "conscious" and the corresponding noun are ordinarily used, including their polymorphism. This is connected with the fact that instead of identifying *one* topic for research and modelling, the adjective and noun point at a variety of disparate phenomena produced by

biological evolution and individual development that need to be investigated.

1.2. *Biological information-processing*

One of the assumptions made here and in AWMC, which is very widely shared, is that a key feature of biological organisms is processing of information. There now seem to be thousands of books and papers on very many different types of biological information processing, although a few individuals still seem to think that the word "information" refers to bit-patterns and that "information-processing" describes only what computers do.[d] In the ordinary non-technical sense of "information" all organisms, even the simplest, use information to control their behaviour, as argued by Steve Burbeck:[e]

> *A National Science Foundation workshop report points out that, "A series of discoveries over the past fifty years have illuminated the extraordinary capabilities of living cells to store and process information. We have learned that genes encoded digitally as nucleotide sequences serve as a kind of instruction manual for the chemical processes within the cell and constitute the hereditary information that is passed from parents to their offspring. Information storage and processing within the cell is more efficient by many orders of magnitude than electronic digital computation, with respect to both information density and energy consumption."*
>
> *Computing professionals would do well to understand the parallels too. All living organisms, from single cells in pond water to humans, survive by constantly processing information about threats and opportunities in the world around them. For example, single-cell E-coli bacteria have a sophisticated chemical sensor patch on one end that processes several different aspects of its environment and biases its movement toward attractant and away from repellent chemicals. At a cellular level, the information processing machinery of life is a complex network of thousands of genes and gene-expression control pathways that dynamically adapt the cell's function to its environment.*

One of the important aspects of these control pathways, both in computing systems and in biological information-processing systems is the use of what have unfortunately been called *virtual machines*, unfortunately because they and their contents are as *real* as socio-economic systems, crime, poverty, economic inflation, and other non-physical entities and processes that depend for their existence on a physical world. I shall say more about how virtual machinery relates to some introspectible contents in Sections 3.3 and 3.11. (There is more on this on Burbeck's web site: `http://evolutionofcomputing.org/Multicellular/Emergence.html`.)

1.3. *Information-processing architectures*

In order to construct explanations of the many phenomena associated with ordinary uses of the words "conscious" and "aware" that can be tested in working models, we shall need to be able to specify classes of information-processing architecture characteristic of different stages of development of human minds, and other animal minds, along with types of representation (i.e. types of information-bearer) and types of mechanism needed for use in such an architecture. Typically these are virtual machine architectures, not physical architectures.

Since humans (and some other species) are not born with a fully developed adult information processing architecture, one of the features of the required explanatory architecture, currently missing in AI, is an ability to construct itself and extend itself (though the early stages of development will use different mechanisms, mostly chemical information processing, required

[d] Anyone who has that opinion should try giving to google the phrase "information processing" in combination with one or more of the words: "human", "infant", "linguistic", "intelligence", "animal", "biological". I have attempted to give a schematic answer to the question "What is information?" in (Sloman, (to appear).

[e] `http://evolutionofcomputing.org/Multicellular/BiologicalInformationProcessing.html`

to build a brain). Some people think this could start from an empty, or nearly empty, information store, as Turing (who should have known better) suggested in (Turing, 1950):

> *Instead of trying to produce a programme to simulate the adult mind, why not rather try to produce one which simulates the child's? If this were then subjected to an appropriate course of education one would obtain the adult brain. Presumably the child brain is something like a notebook as one buys it from the stationer's. Rather little mechanism, and lots of blank sheets. (Mechanism and writing are from our point of view almost synonymous.) Our hope is that there is so little mechanism in the child brain that something like it can be easily programmed.*

For reasons discussed in (McCarthy, 2008), it is more likely that millions of years of evolution produced a great deal of information about the environment in the genome. This is most obvious in species born or hatched highly competent. Less obvious, more indirect, relations between the genome and development are discussed in (Chappell & Sloman, 2007).

Trying to characterise in a *generic* way the space of possible information-processing architectures and mechanisms and the space of sets of requirements that they can satisfy can help with the longer term project of specifying *particular* architectures and sets of requirements, a much harder task. Doing the more generic task, or at least starting it, is a requirement for scientists and engineers to agree, at least provisionally, on a set of concepts and a language for expressing their proposals and criteria for comparative evaluation. Of course, such a language will develop and change as it is used in the process of building, testing and altering working systems. But we have already had several decades of exploration and should make use of what has already been learnt about alternative architectures. Unfortunately, many AI researchers proposing architectures for intelligent systems ignore such preliminaries and simply present their own designs, using their own ad-hoc notations, diagrammatic conventions and concepts.

This myopia is often manifested in the use of the phrase "an architecture" or "the architecture" in titles. This is a poor way to do serious science. I have tried to remedy this over many years by stressing the importance of explorations of alternatives in design space, by attempting to specify dimensions in which architectures and sets of requirements can vary, and by providing a toolkit for exploring alternative architectures (Sloman, 1999; Sloman & Logan, 1999). Unfortunately attempting to specify the spaces within which design options exist, as was done in sections 7 and 8 of AWMC, is sometimes misread as attempting to propose a specific architecture, which is what most researchers do, including, for example, many researchers studying consciousness or emotions.

Alternative approaches making use of brain scanners and other probes, are based on the assumption that studying brain mechanisms without having a deep theory of the functions they support can reveal the information processing architecture: a sort of "de-compiling" process. I suspect this will turn out to be nothing more than a dream.

1.4. *Immodest criticisms*

The criticisms made here and in AWMC of research on consciousness are not specific to research on *machine consciousness*. For many years, since writing Chapter 10 of (Sloman, 1978), I have been (immodestly) criticising a wide variety of researchers attempting to treat the noun "consciousness" as a label denoting a unitary object of scientific study or philosophical puzzlement, concerning which questions can be posed such as: How did "it" evolve? When does a foetus have "it"? Which neural mechanisms produce "it"? What is "its" function? Which animals have "it"? How can "it" be modelled? I shall later compare this with treating "disease", "danger" or "efficiency" as a label denoting a unitary object to be modelled.

It may seem immodest to criticise such a large and varied collection of researchers, but the criticisms are based on arguments and evidence, for example the sleepwalking example in (Sloman & Chrisley, 2003) and AWMC, which indicates contradictions in widely-held beliefs

about consciousness. So far, to my knowledge, refutations showing flaws in the arguments have not appeared. Instead, the arguments have simply been ignored except by two commentators on AWMC who seemed to endorse the arguments (Legg, 2004; McDermott, 2007).[f]

Moreover, the criticisms are not just my criticisms. For example, in (Jack & Shallice, 2001) we find *"The theoretical and methodological difficulties facing a science of consciousness run deep. From time to time, a precarious consensus may emerge and cause these difficulties to fade from view. At such times, there is a temptation to forge ahead with experimental and theoretical work - to take advantage of the temporary suspension of critical impediments. Yet, there are eminently practical reasons for attending to the difficulties. Unless they are dealt with explicitly, they are likely to resurface, throwing much previous work into doubt"*. Further, Block's 1995 paper introducing the much referenced distinction between phenomenal and access consciousness, discussed further below, includes criticisms of concepts used by philosophers, some of which are the same as my criticisms except that he uses the term "mongrel concept". However, I'll try to show below that his own concept of p-consciousness cannot be used for serious purposes of science or engineering and is used to pose bogus philosophical problems. That opinion is also not original or unique to me.

2. Some philosophical background

Discussions of the mind-body relationship and the nature of minds have a very long history, which I cannot survey in detail here. Instead I shall summarise a small subset of the background relevant to the main concerns of AWMC and this paper, starting with a very compressed introduction to conceptual analysis. For more detail on that topic see chapter 4 of (Sloman, 1978) online here: `http://www.cs.bham.ac.uk/research/projects/cogaff/crp/`, along with works of recent analytical philosophers, e.g. J.L. Austin, G. Frege, B. Russell, G. Ryle, P.F. Strawson, and others. After the introduction to conceptual analysis, including distinguishing a number of semantic flaws that can afflict serious thinkers, I summarise and criticise a specific subset of what has been said about consciousness, especially what some people call "phenomenal consciousness" following (Block, 1995), although the core idea is much older.

The terminology I introduce, regarding kinds of semantic flaw, and polymorphic concepts, is not universally used by philosophers. However the notion of a polymorphic concept will be familiar to analytic philosophers. After explaining confusions arising out of failing to notice that some concepts (including "conscious" and its cognates) are polymorphic, in Section 2.2, I shall go on, in Section 2.3, to describe five kinds of semantic flaws: 1. self-defeating semantics, 2. extrapolation failure, 3. dogmatic extrapolation, 4. semantic disengagement, and 5. lack of coherence.

2.1. *Polymorphic concepts*

The idea of polymorphism of a concept, is partly analogous to the notion of polymorphism in biology, and goes back at least about 60 years to the philosopher Gilbert Ryle, and possibly others before him. A closely related concept is also familiar in computer science (possibly introduced by a philosophically educated computer scientist?). Polymorphism is sometimes referred to as "overloading" in connection with object-oriented programming. Different sorts of polymorphism in programming languages have been distinguished. The notion of "parametric polymorphism" is an important subclass, explained below.

A concept has "higher-order" functionality insofar as it does not determine criteria for its applicability until it is provided with another concept, or predicate, or function, as one of its arguments, and possibly also some contextual information. For example, if you learn that my

[f] As this article went to press I found that an attempt had been made in (Velmans, 2009) to address some of the contradictions.

latest acquisition is highly efficient you cannot tell anything about its behaviour unless you have additional information, e.g. that it is a lawn-mower, a dish-washer, a cure for coughs, etc. The extra information may be provided explicitly, e.g. as subject or object of a verb, as an adverbial phrase, or in some other way. However it is often implicit and left to the knowledge and intelligence of the hearer. We shall see that this is one way that polymorphic concepts can cause confusion.

Depending on what extra information is provided, what the concept applies to can vary a great deal, and precisely what is meant can vary: hence the label "polymorphic". In particular, I shall try to explain how the concept "conscious of" is polymorphic, because in "X is conscious of Y" what is meant can vary widely both according to what Y is and also according to different kinds of X, e.g. an adult human, an infant, another type of animal, a family, a nation, a chemical plant control system, a future robot, etc.

A concept, relation, or function has parametric polymorphism if it can be applied to entities of different types, and the precise specification of the concept, relation or function, e.g. the truth conditions of a concept, depends in a principled way on the type of entity. For example, in some programming languages what the symbol "+" refers to will depend, in a systematic way, on whether it is applied to integers, decimal numbers, vectors, strings, lists of symbols, or something else. (Burbeck[g] uses the word in a different though related way, namely by defining a message or signal (e.g. in organisms or computing systems) as polymorphic if its meaning is partly determined by the recipient rather than the sender alone. This is compatible with the notion of implicit arguments, except that it allows the sender not to know which argument is to be used.)

Parametric polymorphism can be contrasted with "ad hoc" polymorphism, where instead of the conditions of applicability depending in a principled way on the arguments of functions, predicates or relations, they result from a sequence of extensions that some language users found convenient, rather than being determined by the previous meaning. What Wittgenstein called a "family resemblance" concept in (Wittgenstein, 1953) typically has this character. It is sometimes a result of what I call "dogmatic extrapolation" below in Section 2.3. Parametric polymorphism is typical of the kind of abstraction common in mathematics, whereas ad hoc polymorphism is typical of metaphorical extensions of usage. I suspect that "conscious of" has both kinds of polymorphism, but that is a topic requiring further investigation.

Describing a concept as polymorphic is not a criticism: polymorphic concepts are important in our language, as illustrated by ordinary uses of "conscious" and "consciousness" and many other examples discussed below. However this complexity can lead to confusion and error when it is misunderstood, or goes unnoticed. A (hopefully) uncontroversial example of a polymorphic concept is "build". It is possible to build a sandcastle, a brick wall, a house, a company, a reputation, or a nation, and the criteria for deciding whether the statement "X built Y", is true will vary enormously depending on what Y is, and what X has to do to build Y will vary enormously. Less obviously, similar comments can be made about "X is conscious of Y", in the ordinary use of the word "conscious" (at least in English, though I assume many other languages have similar words). In English there are several related verbs, adjectives, and adverbs that have this kind of polymorphism, including "aware", "experience", "notice", "attend to", "discover", and "perceive", for example. Even the appallingly overused phrase "what it is like to be X" is polymorphic in this sense.

In what follows, I shall make many statements about ordinary usage that are based on years of observation, and can be checked by searching for examples of their use in many technical and non-technical texts. (I have recently done this using google.) I shall not provide evidence here, apart from a small number of examples needed to aid communication.

[g] http://evolutionofcomputing.org/Multicellular/Messaging.html

2.2. *Confusions caused by polymorphic concepts*

Philosophical and linguistic puzzles can arise if thinkers use concepts that appear to be problematic because they use over-simplified theories about what concepts (or meanings) are. For example, there have been puzzles about what we could mean by words like "heap" or "pile". One stone, clearly doesn't make a "heap". If you don't have a heap, you cannot produce one by adding one stone. So it can be argued that it is impossible to make a heap of stones if you don't start with one! This problem (referred to as the "sorites paradox") goes back to ancient Greece.[h]

There is a (generally unnoticed) solution to this which is to treat words like "heap" and "pile" as expressing higher order (polymorphic) concepts that require an implicitly specified goal and a context and determine criteria to be used. So if your goal is to hold down a tarpaulin that is blowing in the wind, the request in that context to put a pile of stones on each corner determines how many stones make a pile, namely enough to hold that tarpaulin down in that wind. If you can't see over a wall, you might decide to get a pile of bricks to stand on. How many bricks are enough will depend on your height and the height of the wall, and perhaps other considerations such as stability. Here the polymorphism is linked to an *implicit* argument provided by the context, which if it were made explicit might take a form using the preposition "for", or "to" (infinitive) e.g. in "a heap of Xs for Y", or "a heap of Xs to do Z". Often the extra parameters are not made explicit because they are assumed to be obvious from the context, in accordance with one of Grice's well-known maxims of communication namely, when talking, people should provide all and only the information required for the communication to be successful (the "Maxim of Quantity").

Other familiar polymorphic concepts are "large", "short", "good", and "bad". It is a familiar point that most uses of these words presuppose what is often called a "comparison class", unlike the *comparative* versions of the adjectives, e.g. "larger", "shorter", "better", etc. So a tall mouse will be much shorter than a short baboon. The comparative works across classes, but not the "base" adjective. (The comparatives seem to be semantically and epistemologically more fundamental, as explained in (Sloman, 1969, 1970).)

"Efficient" is an example where there is usually an explicit argument though sometimes a context is also referred to, implicitly. E.g. attempts to formulate a definition of "X is efficient" in terms of properties that suffice to make X efficient will fail, because the criteria for efficiency differ according to what kind of functionality is being assessed, e.g. something can be an efficient lawnmower but not an efficient harvester, or an efficient delivery van without being an efficient racing car. The criteria for calling something efficient are usually clear from the context, e.g. when people are talking about efficiency of lawnmowers, car engines, animal hearts, fitness regimes, algorithms, or government organisations.

However there is no ONE property they are all referring to. Efficiency is essentially a complex *relation* (to some goal, or function) rather than a *property*. If someone claimed to be building a working model of *efficiency* to explain all examples of efficiency, we should be very suspicious. "Efficient" expresses a higher order predicate that takes a lower order predicate or goal as one of its arguments, and possibly additional contextual information (e.g. time of year, type of physical location, type of person operating a machine, available funds, legal or other constraints, etc.)

If this is correct, then producing a working model of efficiency would be an incoherent (semantically flawed) goal, unlike the goal of producing an efficient lawn scarifier, or an efficient treatment for a particular disease. Additional kinds of semantically flawed concepts or goals are discussed below, in 2.3.

So a concept that appears to lack coherence, because it seems to be ambiguous, may really be polymorphic, i.e. capable of taking different forms determined by the context of use. It is

[h] `http://plato.stanford.edu/entries/sorites-paradox/`

possible that one of the ideas Block had in mind when discussing "mongrel" concepts and different sorts of ambiguity in (Block, 1995) was polymorphism. Polymorphic concepts are very useful, and that applies to the colloquial uses of "aware", "conscious", and their derivatives, and also their typical medical uses. They become sources of deep muddle when used in formulating scientific or philosophical questions and theories without allowing for their polymorphism.

2.3. *Kinds of semantically flawed concepts*

It has long been known to philosophers that forms of words that look as if they express some definite meaning may be incapable of referring to or describing anything. In some cases that involves an explicit **contradiction**, as in "a triangle with five corners", which cannot describe anything if "triangle" is defined to refer to a planar polygonal object with exactly three corners. "The largest prime number is not divisible by 95" may appear to be true by definition of "prime", but a well known, very beautiful, ancient piece of mathematics demonstrating that there are infinitely many primes, shows that the subject of the sentence cannot refer to anything.

Self-defeating semantics: A more subtle kind of semantic flaw, which could be called **self-defeating semantics**, involves assembling meaningful words so as to describe something that undermines the conditions for those words to be able to refer to something. For example, "the time at the centre of the earth" must be incoherent if "the time at X" is defined in terms of the altitude of the sun above the horizon at location X, and "the centre of the earth" species a location that cannot have a horizon because only points on the surface of a planet can.[i] Likewise asking whether the whole universe is moving north at 3 miles per hour has self-defeating semantics if motion can exist only relative to something else, and if the whole universe by definition includes everything. The history of philosophy has many examples of more or less complex and subtle examples of self-defeating semantics. For example, some notions of "free will" and of "God" can be shown to have such flaws.[j] I shall later illustrate this kind of semantic flaw with a concept of identity of enduring spatial location.

Extrapolation failure: A related kind of semantic flaw, which I'll call **extrapolation failure**, can afflict attempts to extend the use of a concept beyond its normal range of circumstances of use. Attempts to apply an old concept, in new contexts where normal assumptions are violated, can fail because the normal conditions of use need to be extended. What I am calling "extrapolation failure" typically occurs where there are different options for extension, none of which is determined as the *right* extension by the previous use.

E.g. If I choose a point P on the surface of Mars and ask "What time is it at P now" there is an extrapolation failure, because our normal use of "time at location X" is defined only for locations on or close to the surface of the earth, where time-zones have been specified. That use does not determine whether we should say of the point on Mars (a) there is no time it is now at P because P is too far from earth, (b) the time depends on the angular elevation of the sun at P, (c) the time depends on which timezone on earth is intersected by the line from P to the centre of the earth, or something else. (Further unclarity arises from the finite speed of light.)

Some of the cases where there are alternative possible extrapolations of usage, but none of them is determined as the *right* one by previous usage, are sometimes described as "cluster concepts". F. Waismann used the label "open texture" for similar cases. Assuming there is a unique correct extension when there isn't one can lead to unjustified new uses of old concepts, a sort of **dogmatic extrapolation**. Examples might be using a notion like "emotion", or

[i]This is a modified version of Wittgenstein's example "What time is it on the sun?". Other variants are presented later.

[j]Discussed further in
`http://www.cs.bham.ac.uk/research/projects/cogaff/misc/varieties-of-atheism.html`

"consciousness" to describe a machine whose states have some features that are similar to the states of humans thus described, but which lack others, e.g. describing a machine as happy simply because it produces happy-looking behaviour, or describing it as conscious just because it includes a mechanism that matches somebody's theory of consciousness, while lacking most of the other characteristics of humans or animals described as conscious. Dogmatic extrapolation may not matter when nothing hangs on it: e.g. if there is an agreed convention to use an old word in a new way in new contexts. This kind of dogmatic extrapolation is sometimes useful in mathematics: e.g. extending arithmetical operators, such as multiplication and exponentiation to negative numbers for the first time. Sometimes, what starts as dogmatic extrapolation can become a new accepted usage of a polymorphic concept.

Semantic disengagement: Another kind of semantic flaw, might be described as **emptiness**, or perhaps **semantic disengagement**, by analogy with a gear wheel that does not engage with any other gear wheel or sprocket chain, etc., whose turning is therefore of no consequence. If someone were to invent a concept of "aethereal-alter", allegedly associated with every physical object, occupying exactly the same space as the object, moving whenever the object does, but impossible to detect or measure except by observing its object, and whose states, properties, or changes of size or location have no detectable effects, then the notion of your or my aethereal-alter will be an example of a concept that is useless because disengaged. Everything its instances might possibly be used to explain is already explained using other concepts. This is the kind of flaw I'll point out in the concept of "p-consciousness" (or P-C, for short). Block more or less explicitly defined P-C to be uselessly disengaged. When I discuss the notion of a virtual machine running in a physical machine, I shall try to show why that is not a disengaged concept, and also why a certain sort of concept of content of introspection is not disengaged, but engages in rich and complex ways with other parts of the system, as allowed by Block's definition of "access-consciousness" (A-C). That sort of introspection, and those introspective contents, may one day be replicated in a suitably designed machine.

Lack of coherence: This is use of a word by different people, or by the same person in different contexts, to refer to different things, without being used in a systematically polymorphic way. This may simply be a case of harmless ambiguity, like the ambiguity of the words "bank", "pound", and "record", where the context usually makes clear what is being said and everyone is aware of the ambiguity. **Pernicious** lack of coherence occurs when researchers think they are identifying some unique phenomenon to be explained, without noticing that they use their words to refer to different things. In that case the claim to be doing research on "it", or to be modelling "it" is systematically ambiguous and should not be made by scientists or philosophers as if they were uniquely identifying a research problem. The word "consciousness", as used by many researchers trying to study or model what they refer to using the word, often has this pernicious lack of coherence: there is no well-defined, generally agreed, subject of study.

If the lack of coherence between researchers goes unnoticed, then claims to have explained or modelled or accounted for the evolution or development of consciousness will have no clear content and could be seriously misleading, even to the authors. I am not alone in claiming that, and some of the commentators on AWMC agreed. In 1995, Block used the phrases "mongrel concept" and "cluster concept", though in different contexts (see Section 3 below). Others have also talked about "cluster concepts". Minsky in (Minsky, 2006) refers to "suitcase concepts". Some of the divergences particularly between those interested in sensory-motor theories of consciousness and those interested in global workspace theories are described in (Degenaar & Keijzer, 2009). (The authors also propose a synthesis of the two approaches.)

Sometimes thinkers try to remedy the semantic problem by defining a precise kind of consciousness as their subject of study, adopting the *special case* strategy (defined in Section 1.1, without realising that they have introduced another kind of semantic flaw, for example self-

defeating semantics, or semantic disengagement – defining a concept that operates ineffectually, failing to engage with anything of substance.

The ordinary, pre-theoretical, use of the adjective "conscious" does not have those flaws, although it expresses a concept that is polymorphic in the sense defined in Section 2.1. The notion of access consciousness, A-C, will also turn out to be polymorphic. In contrast, attempts to define P-C in opposition to A-C have produced a new concept that is semantically disengaged. I shall try to illustrate how that can happen with a simpler example, the concept of an enduring region of space (RoS), before returning to the semantic flaws of the P-C concept.

2.4. *The "hard problem" of spatial identity*

One of the standard strategies for attempting to show that a concept is not flawed is to present examples. Of course, just producing the examples and identifying them proves nothing, since that leaves open what the objects presented are examples of. Every entity in the universe instantiates many concepts, so that pointing at something does not identify any particular concept. Pointing at several examples may help to eliminate some misunderstandings by ruling out concepts that apply only to a subset of the examples, but still does not identify a concept uniquely since any set of objects will have more than one thing in common.[k]

For these reasons, so-called "ostensive definitions" are useless as definitions, unless supplemented with additional information about the concept being defined,[l] although providing examples of important new concepts can be helpful as part of a larger process of communication, as every teacher knows. Conceptual confusions leading to the use of semantically flawed concepts can arise from thinking that what is meant by a word or phrase can be explained by pointing at examples while introspecting. I don't claim that introspection is impossible, or that it is not possible to attend to and think about contents of introspection – several examples are presented in (Sloman, 2009d). However, simply pointing at contents of introspection in order to define a concept has specific problems mentioned later. I'll illustrate the general problems with a simpler example involving pointing at a region of space.

Someone might wish to talk about regions of space (RoS) as things that can endure across time.[m] He could explain what he means by "a RoS" by pointing in front of him and saying "An example of a RoS is *here*" (possibly cupping a region in his hands). Clearly he is pointing at, and has successfully identified a region of space, and it is possible to present several examples. There may be ambiguity about a region's precise boundaries but we could eliminate or at least reduce the ambiguity by grasping an object of the required shape and size, holding it in the required location and saying "I mean the RoS occupied by this cup, *now*".

Such a person might think that having identified the RoS, he can later ask where it is. After all, the RoS existed earlier, and there is no reason to think it was destroyed, or even could be destroyed, so it must still exist and be somewhere. Where?

The assumption was once tempting to people who had never heard of Einstein (e.g. to Newton `http://plato.stanford.edu/entries/newton-stm/`, though not Leibniz) that "here", said while pointing, identifies a portion of space whose identity across time is somehow intrinsic to itself and independent of all its relationships to other space occupants, like walls, floors, trees, people, planets, or galaxies. So it may seem that you can ask five minutes later, "Where is IT now?", or "Is this the same RoS?" said while pointing again. But there is no way of answering the question. It was once thought that if not space, something that filled all

[k]This problem is familiar to philosophers. A little more information and some references can be found here `http://en.wikipedia.org/wiki/Ostensive_definition`

[l]This point is ignored by AI researchers who think a robot with no initial information about how to divide up the world, or why it is being divided up, can be taught concepts by being shown examples paired with labels. The point also undermines symbol-grounding theory, as Immanuel Kant had done long ago.

[m]I use the word "region" in a general sense that includes 3-D volumes.

space, a kind of stuff called "aether", could have portions that endure, so that we can ask how fast an object moves through it. But the negative result of the Michelsen-Morley experiment attempting to measure such a speed undermined reasons to believe in such stuff. Nevertheless, it may seem to some that we can still ask how fast something is moving *through space*, if not through aether.

I hereby label the problem of re-identifying a RoS *"The hard problem of spatial identity"*. Alas, the question "Where is that RoS now?" cannot be answered because there is no such thing as an enduring RoS whose identity is independent of all its spatial relationships to other space occupants (as Leibniz argued, against Newton). Someone who feels there *must be* such a thing is either engaged in unwitting self-deception, or possibly trying to revive the theory of "aether" that was undermined by the Michelsen-Morley experiment, and then demolished by Einstein's theories. There are many similar unwitting self-deceptions in the history of philosophy, and I shall try below to explain why belief in the existence of P-C (as defined by Block, in opposition to A-C), and the belief that there is an associated "hard problem" are examples of such unwitting self-deception, though very much more complex examples than the spatial identity problem. (I don't claim to be saying anything new, however.)

Note that I am *not* saying that there is no such thing as an enduring region of space! However, whether my foot is still in the same place now as five minutes ago depends on whether "same place" refers to the location on my body, the location in my office, the location in the solar system, the location in our galaxy, etc. Someone in a fast moving train being asked where his foot was five minutes earlier is likely to specify a location relative to the contents of the train rather than a location relative to the railway track, or the solar system. In a murder investigation both locations (e.g. of a gun fired in a moving train) might be relevant. Normally, a question about the location of a region of space originally identified in the past (or in the future, e.g. an expected point of collision) will explicitly or implicitly make use of a framework of reference defined by a collection of spatially located objects. Our common verbal expressions do not make this explicit (e.g. when we say that something is or is not moving, or has or has not moved), and are therefore systematically incomplete, but the incompleteness does not matter when everyone concerned knows which frame of reference is in question (even if some are not conscious that they know it and make use of it!).

Any attempt to identify a notion of re-identifiable location without such a framework, e.g. asking whether the whole universe has moved, uses a semantically self-defeating concept.

2.5. *Polymorphic identity*

The concept of identity of regions of space that we normally use is *polymorphic* insofar as different frameworks of reference can be provided as an explicit or implicit argument to the relation (binary predicate) "same place" to instantiate it in particular contexts of use. Whether a particular statement using the concept is true or false will depend on the extra argument. So the concept is *polymorphic*, in the sense explained above: how to decide where a previously identified RoS is now, depends in systematic ways on extra information required to specify the question. Without the extra information the question has a serious semantic flaw: it cannot be answered – because there is no correct answer!

2.6. *One-way causation and epiphenomenal fairies*

Newton's regions of space that are real, and endure with an intrinsic identity do not have any causal powers or causal relations. Rather they provide a framework in which other things can exist and interact causally, unlike the space of Einstein's general theory of relativity whose curvature can produce accelerations in bodies with mass. In contrast the concept of "aethereal-alter" (A-A) introduced in Section 2.3 allows one-way causation: moving an object changes the location of its A-A. Depending on how the theory is elaborated, doing other things to

the object may produce other changes in its A-A. Because the causation is one-way, the A-A is "epiphenomenal". It allegedly exists and is influenced, but cannot influence anything, and therefore its existences and changes of state can never be detected by anything else.

Once we allow any such concept to be used we can go on multiplying them: e.g. every object has a bunch of invisible, intangible, fairies of different colours associated with it and things done to the object can affect the fairies, but nothing they do, and none of their changes of state can affect other things, or ever be detected by anyone or anything else. Of course, if such fairies are also philosophers who think about existence, and they hear our comments they will be silently, and ineffectually, willing us to acknowledge their existence, but all in vain, since whatever we believe about them will not be affected by their willing or anything else they do.

There is no logical argument that can prove that such things do not exist, but pondering their possibility seems to be completely pointless, except perhaps as a form of entertainment. And for that entertainment to exist, the fairies and alter-aliases do not need to exist, for the entertainment is produced by other means. Moreover, once you start thinking about such undetectable entities you can easily fill all your waking hours contemplating ever more new kinds. We here have a clear example of *semantic disengagement*, defined in 2.3.

The fairies are unlike regions of space, because we need those in our ontology, including ones that are not yet occupied by anything, since we often have to select those to pass through or put things in. RoSs, even empty ones, play a key role in the affordances we perceive, think about and use. However, we don't need them to have intrinsically enduring identities: their identity is polymorphic, relative to a framework of reference. So they are a clear step away from semantically disengaged fairies enjoying the flowers at the bottom of the garden and willing them to grow because they don't understand their own causal disconnection.

2.7. *Why "conscious" is polymorphic*

The ordinary uses of "conscious" are far more complex and messy than the ordinary uses of "same place", or "efficient", but there are similarities, especially insofar as in ordinary (non-scientific) language what being conscious means is usually clear from the context, e.g. "I wasn't conscious of wearing my sweater inside out". As that example illustrates, in ordinary language consciousness is normally *of* something, and is therefore inherently relational. Just as the goal or function determines the features or mechanisms required for efficiency, and further contextual information can modify the requirements, so also the object of consciousness determines what being conscious involves, and other things can modify the requirements for being conscious of it. In this ordinary relational sense, being conscious of X means roughly having some information about X produced by X and being in a position to get more information about X or to continue receiving information about X.

What sort of process that is, what kinds of mechanism are required, and what sort of formalism or form of representation is useful, will all vary widely according to what X is, and also on what kind of individual is conscious of X. Compare being conscious: of a toothache, of a difference between this toothache and one experienced a month earlier, of the beautiful weather, of being about to lose your balance, of the fragility of an ornament, of the beauty of a piece of music, of enjoying a joke, of the danger of trying to jump across a deep chasm, of having forgotten to lock your front door when you went out, of being watched by someone, of being unable to remember someone's name, of being unsure whether you have found a new proof of a theorem, of being disliked by your colleagues, of being puzzled by a noise or an argument, and many more. Humans can be conscious of all of these things, or unconscious/unaware of them, but there are very great differences in the states and processes referred to. Analysing all those examples, and spelling out their similarities, differences, types of access to the object of consciousness, forms of representation required, mechanisms required for acquiring and using the information, architectures within which such mechanisms and others can cooperate, and

the resulting capabilities and uses, would require a paper far longer than this one, but should be done by anyone attempting to produce a good theory or model of human consciousness.

Perhaps someone whose work I have not encountered has already done that. Some of the details are in (Ryle, 1949), and probably much else can be found in books on phenomenology, though probably without discussion of, or understanding of, information processing mechanisms and architectures. Ryle even regarded such science as irrelevant, though his work could be construed as assembling *requirements* that any proposed design, or any good explanatory scientific theory, should meet.

The diversity of requirements can be illustrated by the following example. Any attempt to define a very general sense of being conscious covering all of the ordinary uses, leads to a notion of consciousness that does not require being awake, since it would have to include examples of consciousness that span several days! If someone says "I have been conscious of your antagonism towards me for the last few weeks" that does not imply that the speaker has been awake for a few weeks. (Exactly what it does imply is left as an exercise in conceptual analysis for the reader.) But it would be seriously misleading to say that consciousness of antagonism was present only at certain times of the day: that sort of consciousness is quite different from being conscious of something moving towards you, or being conscious that you are conscious of something moving towards you.

For these reasons, trying to produce a working model of consciousness in general is as misguided as trying to produce a model of efficiency in general. However, there is nothing wrong with trying to produce a working model of a mind, in all its richness (apart from the extreme difficulty of the task – and the need to take great care over the *requirements* to be satisfied by such a model). There is also nothing wrong with identifying portions of the task and trying to make progress by building models, as long as the work on parts is not presented as work on the whole, by using labels such as "a model of consciousness", "a theory of consciousness", as opposed to for example "a model of consciousness of X with/without self-awareness" or "a model of introspections of states and processes of type X", etc.

A possible source of confusion may be the fact that there is a related use of "conscious" to label a state (consciousness, as opposed to unconsciousness) in which one is capable of being conscious of a variety of unspecified things, but is not necessarily conscious of anything in particular. Another label for that is "being awake". But there is no corresponding notion of being efficient without being efficient about something in particular. The fact that "consciousness" has this non-specific use, might tempt an unwary researcher into trying to characterise what that sort of consciousness is, and trying to explain how "it" evolved, at what stage of development a foetus has "it", what mechanisms produce "it", and from there to build a working model of "it", whereas nobody would dream of trying to do the same for efficiency. (I hope.) Of course there is nothing wrong with trying to understand the (many) differences between being asleep and being awake, or trying to understand intermediate cases like sleep-walking, absent-mindedness and hypnotic states, or the varieties of causes of loss of consciousness, such as a blow to the head, fainting with horror, falling asleep, being anaesthetised (in various ways), or even being totally engrossed in a task that shuts out everything else.[n]

2.8. *Ordinary non-scientific uses of "conscious" and "consciousness"*

It is possible for readers to be confused by my claiming that ordinary non-technical uses (including medical uses) of the noun "consciousness" and the adjective, are fine, while criticising scientific and philosophical uses or attempts to build a working model of consciousness.

[n]Luc Beaudoin has pointed out to me that besides the problems of polymorphism, some words reflect language-specific or culture-specific concepts. E.g. in some languages "consciousness" and "conscience" are apparently not distinguished as sharply as in English.

In AWMC, I somewhat obscurely claimed both that much that has been said and written about consciousness in science and AI is muddled or incoherent (perhaps "semantically flawed" would have been a more suitable term), while at the same time allowing that those criticisms do not apply to ordinary non-scientific uses (including medical uses) of the words "conscious" and "consciousness" because what they mean is normally clear from the context. I hope it is now clear what was meant by that. The words "conscious" and "consciousness" are polymorphic in ordinary usage and the polymorphism generally does not prevent successful communication, because the missing arguments are unwittingly inferred from the context, so that nobody thinks being conscious of a toothache is like being conscious of being unpopular among colleagues. It is only when people start theorising that they make false assumptions and things go wrong.

So there is no contradiction because I have been contrasting facts about use of language in scientific and non-scientific contexts. The misuse of an ordinary word in a scientific context is not uncommon. I have elsewhere criticised similar misuses of the words "emotion", and "self", even though both words have acceptable uses in ordinary language. I have discussed some of the confusions linked to the word "self" in (Sloman, 2008b).

I do not know of any attempt to specify or replicate polymorphism in models of consciousness: that would require building a collection of different models, or a single complex system with different mechanisms of consciousness that are invoked as needed. There have been attempts to model specific aspects of human consciousness (or alleged aspects), for instance attempts to build robots with sensory-motor capabilities describable as being conscious of some of their environment and of what they are doing (because of feedback loops), and other attempts to build machines that have some sort of limited capacity global workspace that can be described as being conscious of varying aspects of tasks they are engaged in. There are also attempts to build simulations of aspects of animal brains in the hope that replicating their dynamics will suffice to justify claims about consciousness. Often such projects are pursued in isolation, though a somewhat abstract proposal for merging the mechanisms of two of the theories is presented in (Degenaar & Keijzer, 2009). We need a deeper, more comprehensive study of the varieties of competence that need to be supported by a single architecture that can support multiple capabilities, including its own development.

Some philosophers, aware of the diversity of uses of "conscious" have tried to define more technical notions about which they then proceed to raise questions or offer theories. This attempt to avoid polymorphism can lead to various kinds of incoherence, or flawed semantics, as I shall try to explain. In the process I shall have to say more than is normally said about some of the causal relationships of particular examples of access consciousness, including discussing some of the phenomenology of visual perception and its relationships with Gibson's notion of "affordance".

The introspective contents in such cases have a rich web of causal relationships that would have to be replicated in machines with that kind of consciousness. When we understand that richness we can also understand how the concept of phenomenal consciousness defined in terms of its causal disconnection, is a bit like intrinsic (non-relational) identity of enduring spatial regions, and also a bit like the fairies and aethereal-alters. P-C is a semantically disengaged concept. This can be contrasted with the concept of states, processes, and active components of a virtual machine which do fit into a causal web. That's exactly the sort of thing we need for the many varieties of access-consciousness, as I'll try to show.

3. What is phenomenal consciousness (P-C)?

The idea that there is something puzzling or mysterious about consciousness is very old, and was referred to as "the explanatory gap" by T.H.Huxley. The phenomena generating this puzzle have at various times been referred to as "sense data", "sensibilia", "qualia", "ideas", "contents of consciousness", "the theatre of consciousness", "what it is like to be something",

"experience", "feelings", "the feeling of ...", and more recently "phenomenal (p-) consciousness", whose motivation I shall try to explain before criticising it. However, it will turn out that the all the examples of P-C are construed as ghostly shadows of instances of A-C, but that when we have a good theory of A-C the concept of P-C can be abandoned, since what it adds is vacuous and a source of confusion.

3.1. *How can the concept P-C be specified?*

Because humans have experiences and are able to introspect and pay attention to that fact, it is usually not thought to be necessary to say very much to define the concept of P-C: people instantly think they recognise what is being referred to. But they may be deceiving themselves, partly because they are unaware of the dangers of definition by pointing, described above in 2.4, and partly because of the unnoticed variety of phenomena associated with the polymorphic concept of "consciousness", or "awareness", even when it is restricted to introspected states. The concept was described as a "mongrel" by Block, who wrote (1995) "The concept of consciousness is a hybrid or better, a mongrel concept: the word 'consciousness' connotes a number of different concepts and denotes a number of different phenomena", which is what I have also been arguing for many years, since chapter 10 of (Sloman, 1978). This is what justifies the claim that uses of the word "consciousness" among scientists who ignore all this exhibit a lack of coherence.

However, some philosophers and scientists have attempted to identify a much more precisely defined concept, by narrowing down the notion of consciousness they are interested in, with or without awareness of the traps like trying to use "here" to identify an enduring intrinsically identified region of space (RoS), described in 2.4.

3.2. *The special case strategy.*

The *special case* strategy, defined in Section 1.1 was used in (Block, 1995) to avoid the "mongrel" (polymorphic) concept by defining the labels "phenomenal" and "access" for two related concepts of consciousness were. Those labels have been used by many other researchers, though it is not clear that all of them use his words in accordance with his specification.

Block notes that "p-consciousness" cannot be explicitly defined, but he gives examples and a partial definition, and then adds "I take P-conscious properties to be distinct from any cognitive, intentional, or functional property." Since some forms of consciousness, which he describes as "access-consciousness" (A-C) do have cognitive, intentional and functional properties, and play a causal role in our minds, he needs to identify them and show how P-C is distinct from them. Block's notion of P-C corresponds very closely to older ideas about the nature of mind and the contents of consciousness, especially sensory consciousness. What I believe was original about his work was the idea of A-C, which was introduced merely as a contrast to P-C. I shall try to show how A-C has greater significance than is generally appreciated, though it is polymorphic in ways that Block may not have considered. I.e. "access-consciousness" is also a mongrel! Part of its significance is that it can be used to show the incoherence or emptiness of the concept P-C, in something like the way that the *useful* concept of enduring spatial location identified relative to some collection of space occupants can be contrasted with the *semantically disengaged* notion of a spatial region with an intrinsic enduring identity.

Block claims that it is possible to have A-C without P-C and P-C without A-C. Without going into all the details necessary for specialist readers, I shall try to show that pressing this requirement renders the P-C concept seriously semantically flawed, because it is either self-defeating or disengaged. I shall stick with his 1995 paper, since that is most readily accessible, though he continued refining the ideas after it was published. His specification of A-C, against which P-C has to be understood is:

A state is access-conscious (A-conscious) if, in virtue of one's having the state,

> *a representation of its content is (1) inferentially promiscuous...., i.e. poised to be used as a premise in reasoning, and (2) poised for [rational] control of action and (3) poised for rational control of speech. (I will speak of both states and their contents as A-conscious.) These three conditions are together sufficient, but not all necessary. I regard (3) as not necessary (and not independent of the others), since I want to allow non-linguistic animals, e.g. chimps, to have A-conscious (access-conscious) states. I see A-consciousness as a cluster concept, in which (3)–roughly, reportability–is the element of the cluster with the smallest weight, though (3) is often the best practical guide to A-consciousness.*

From the standpoint of a designer of working information-processing architectures we can criticise the details, but nevertheless accept the spirit of this definition of A-C. The notion is not semantically flawed insofar as it refers to familiar (though highly polymorphic) states of mind that may or may not involve introspection, but which can be the contents of introspection, such as various kinds of seeing, being aware of various things that are not objects of perception, noticing, intending, understanding, enjoying, disliking, being surprised, and many more, all of which are capable of being caused and of being causes, though the examples given by philosophers tend to be selected from a small subset of introspective contents related to perceptual experiences, and possibly affective states, such as moods, desires or emotions. Some more detailed examples are discussed in (Sloman, 2009d).

Because there are numerous and very diverse examples of the intended notion of A-C it is polymorphic, just as the ordinary notion of consciousness is. But that is not a semantic flaw (as implied by the presumably pejorative label "mongrel"). The polymorphism is also compatible with all the special cases being suitable subjects for scientific investigation and for computational modelling, which I have argued elsewhere (e.g. (Sloman & Chrisley, 2003)) requires careful design to enable virtual machine entities, states and processes to be objects of various kinds of self-monitoring, self-categorisation, and self control in an information processing architecture. I shall elaborate on that below.

In contrast, the key feature that makes the notion of P-C semantically flawed, as we'll see, is the causal disconnection built into its definition, which can be compared with the relational disconnection of enduring identity of absolute spatial regions and the causal disconnection of aethereal-alters and fairies introduced in 2.6.

Block added further details in (Block, 1999) "A state is access-conscious if it is poised for global control; a state is reflectively conscious if it is accompanied by another state that is about it; and a state, for example, a pain, is self-reflectively conscious if it promotes an accompanying thought to the effect that I, myself, have that pain." Obviously such constructs can be embellished further. Moreover, they are all very abstract and capable of being instantiated in very different ways, which would require different information processing mechanisms to implement them. They can also have different kinds of causal consequences, depending on which detailed phenomena are referred to.

The main point is that P-C and A-C are allegedly both produced by things happening in the environment and in the brain, i.e. they are usually effects of causally efficacious processes and mechanisms, by definition, whereas only A-C has a functional role and can be causally effective. For my purposes the details of how Block describes the functional roles of A-C instances do not matter: the main point is that having those functional roles would be impossible if instances of A-C were incapable of being causes. So they are intermediate entities in a web of causation, involving the environment, brain mechanisms, other mental states and processes, and possibly actions. In order fully to specify what instances of A-C can be, we need to describe both their internal structure, their changes of internal structure, and their intricate causal connections with other things, as I shall partially illustrate below. This is also done in some sensory-motor theories of consciousness (e.g. (O'Regan & Noë, 2001)) but sometimes with an agenda of denying

any role for internal states and processes, which I think can be shown to be seriously misguided (as discussed in (Degenaar & Keijzer, 2009)).

For P-C, everything is intended to be the same as for A-C except that there is no functional role. That implies that instances of P-C have no effects, for if they did those effects would be capable of playing a functional role, unless all their effects feed into a causal web which, as a whole, is disconnected from all the functional mechanisms of mind. From a computational viewpoint, such a thing is possible, since a complex virtual machine can include components that are largely disconnected from other components, and which most or all of the time cannot influence any externally visible behaviour. Any experienced programmer should be able to design such a system. So this is not a completely nonsensical notion. However, insofar as such a virtual machine makes use of physical memory, processing mechanisms, and communication channels for its implementation, it is always possible to extend such a system by making some of its internal structures shared, so that it does have effects on other virtual machines and ultimately on external behaviour. So this is not quite like the semantically disengaged concept of definitionally disconnected aethereal-alters and their accompanying fairies introduced above in 2.6.

That interpretation of the concept of P-C as including contents that happen to be causally disconnected but which are capable of becoming more causally connected by changes in the information processing implementation, does not seem to fit the intended interpretation of what Block and other philosophers have written. I suspect he intended the contents of P-C to be defined as *incapable* under any circumstances of having a functional role. In that case we do seem to be back to flawed semantics, in this case semantic disengagement of the kind that permits endless proliferation of similar entities, like the fairies (ineffectually) willing flowers to grow.

Block attempts to be more precise than I have been: he talks about a functional role in inference, control of action and speech, though for now we can, like him, discount the importance of speech if we wish our theory to be applicable to other animals and pre-verbal humans, who, for some reason, he does not mention. His emphasis on *rationality*, and *use in inference* raises the question whether something could still be an example of P-C if it affected thinking, decision making or action, without doing so rationally, and using mechanisms other than those involved in "rational" inference, as in the case of phobias, or the varieties of learning that go on when children, unlike other species, apparently pick up new vocabulary and grammatical rules and other generalisations from what they experience, but without reasoning about them: they just *do* it because they are products of a certain kind of evolution. (What mechanisms make this learning happen is still a topic of controversy.)

There is also a puzzle as to how anyone can relate to occurrences of P-C by thinking about what they are, how they can be explained, whether they can have causal roles, and so on, if their thoughts and wonderings are not caused by the P-C examples. Surely if the instances were truly functionally disconnected there would not be any of these effects, and all those papers and books defending claims that they exist would never have been written.

However, if P-C instances had no effects that would not make it impossible for philosophers to invent the idea, and wonder whether there were instances, in exactly the same way as we can invent the idea of aethereal-alters and fairies, and with the same semantic flaw of emptiness, or disengagement (apart from the kind of semantic engagement as a form of entertainment that fairy stories can have).

My impression is that Block and other philosophers who worry about the explanatory gap, the hard problem, and other old philosophical problems about qualia, sense-data and the like, intend the concept of what they are talking about (e.g. P-C) not to allow any of those rational or non-rational causal routes to functionality. But in that case we can simply discard the concept as having at most entertainment value (like a fairy tale), especially if we can come up with a better theory to account for the phenomena. The notion of access-consciousness provides the

seed of a better theory, but a very complicated theory, linking A-C to kinds of information processing systems that we do not yet know how to build but which look like being reachable in principle on the basis of things we have learnt in the last half century. However, much more is needed to complete the task. I shall try to indicate some of what is needed to deal with A-C, and then return to why the hard problem is bogus, before sorting out some loose ends and finishing off.

3.3. *Let's do some phenomenology*

A-C, as I understand the concept, covers many types of consciousness that can be the contents of introspection, as illustrated in Section 2.7. These are more or less closely linked to actions and sensory contents, and they have different life-spans, different causal consequences, and different roles in our decision making, social interactions, control of movement, plan execution, learning, motive formation, moods, emotions, attitudes, etc. I shall focus below on examples related to visual experiences of acting in a 3-D environment, but must stress that that is a special case. There are other very different introspectible contents of consciousness that need different information-processing mechanisms: e.g. consciousness of what you are hearing, consciousness of where you are, consciousness of what day of the week it is, consciousness that a plan you are considering is risky, consciousness of having forgotten something, having something "on the tip of your tongue", and many more. We need to understand all of those cases, but for now I shall merely note that they exist, and need to be accounted for in an adequate theory, along with many other cases.

There are some theories, e.g. "global workspace theory" presented in (Baars, 1988, 1997) and partially modelled in (Shanahan, 2006), among others, that claim to be sufficiently general to cover all the cases, including many different functions of consciousness, but working models of that theory will probably remain toy demonstrations for a long time, unable to approach human competences. (They may nevertheless be useful in explicating Baars' theory and its implications.) In particular, the current state of machine vision is still quite pathetic compared with human and animal vision, so the forms of visual consciousness capable of being modelled are very limited. Making progress will require creating new forms of representation, new architectures, and new links between vision and a host of other capabilities, including affordance perception, abilities to control actions, mathematical reasoning, communication with others, social interaction, and many more.

Huge efforts have gone into visual recognition systems: but recognition is a small part of seeing, since there are many things you can see and interact with (e.g. poking, pushing, climbing over, smashing, moving out of the way, asking about, etc.) that you don't recognise. What is still missing includes perception and understanding of 3-D structure, perception of processes of many kinds, including animal actions, perception of types of material, perception of what I call "proto-affordances" (including possibilities for, and constraints on, possible physical processes) (Sloman, 2008a), and also aesthetic appreciation of visual scenes. Different again is perception of epistemic affordances (what kind of information is and is not available in the environment, and opportunities to alter epistemic affordances by performing physical actions) (Sloman, 2008a). Whenever those competences are exercised, some aspects of what is going on can become the focus of attention, and thereby contribute to current states of consciousness.

3.4. *Two-stream muddles*

An aspect of visual perception that has caused much confusion is that there are at least two very different functions of perception of process, namely (a) acquiring re-usable *descriptive* information about what is going on which can be stored and used for many different purposes on different time scales (e.g. "my right hand is moving towards that cup and the forefinger will soon be ready to go through the hole formed by the handle") and (b) acquiring and using

transient *control* information including alignments, discrepancies, directions of approach, and rates of change, as part of the process of controlling the fine details of action. Information of type (b) (discussed in (Sloman, 1978, 1982, 1989) and later work) is often transient, and used only for a particular control function, or during skill acquisition, though in some cases it can also be stored and used later for other purposes, e.g. comparisons useful for learning, or summaries of control sequences that might be needed for diagnosis in case things go wrong.

The differences between descriptive or factual information and control information required for visual servoing, are loosely related to two information routes through primate brains (so-called ventral and dorsal streams), though visual information actually goes in more than two directions. I suspect there are more pathways than have so far been identified. These two routes are often very badly mis-described as differences between "what" information and "where" information. This is muddled for several reasons, partly because it completely mis-describes the control functions, and partly because seeing what something is typically involves seeing where its parts are and how they are related. The mistake was noticed and partially corrected, but using poor terminology, in (Goodale & Milner, 1992), but over a decade later many researchers are still using the silly "what"/"where" distinction.

A better classification can be derived from an analysis of design requirements for a working system. When an animal is pushing an object and watching what it is doing, the contents of consciousness can simultaneously include, among other things (a) descriptive/factual information about what is changing in the 3-D environment, obtained through the eyes, (b) information about how the changes are being controlled, obtained from internal sources and (c) information about how well the actual movements produced relate to intended effects, where discrepancies can be used as part of a feedback control mechanism. For a more sophisticated animal the contents of consciousness during controlled action can include (d) the descriptive information that information of types (a), (b) and (c) is being acquired and used. In the case of normal adult humans, even more things can be going on, including aesthetic appreciation of what is seen and a simultaneous conversation with someone else. (Individuals can vary as to how many streams of information they can process simultaneously. Some of the differences become obvious when a person who is washing up dishes[o] is asked a question.)

It is illuminating to attend to the complexities of the special case of visual perception while acting in a complex 3-D environment. Humans are frequently in situations in which the visual contents of consciousness are both richly structured, with independently changeable components, and are also part of a causal web, tightly linked to physical changes, and capable of having rich functional roles.

J. J. Gibson identified many aspects of perceptual consciousness in (Gibson, 1966, 1979), that are not normally cited in analytical philosophical discussions, though they are directly relevant. Some of the phenomena are discussed under the label *"The sensory-motor theory of consciousness"* associated with the work of Kevin O'Regan and Alva Noë (e.g. (O'Regan & Noë, 2001)), among others, conveniently summarised in (Degenaar & Keijzer, 2009). Such thinkers draw attention only to *part* of what is required for animal vision, and consistent with what Gibson said about affordances and the importance of "active" sensing, but that describes only a subset of the requirements for a complex information-processing architecture, able to account for many ordinary types of consciousness.

I shall try to explain this by giving some examples of A-C that are very closely related to Gibson's ideas and to sensory-motor theories and then show why those ideas are too limited, by giving some pointers to additional requirements for a theory of introspective contents. Many of these are points that have been made by other thinkers, but researchers on consciousness who notice something interesting generally seem to focus on only a subset of the relevant phenomena. Because of the polymorphism there are many subsets! If I am guilty of that too, I am happy to

[o]Possibly a dying breed.

have my collection of requirements for an adequate theory extended.

3.5. *Behaviour controlling what is perceived*

One of the important differences between visual consciousness and auditory consciousness is the richness of the differences small actions can make to the detailed information acquired about the environment, which provide clues as to the details of contents of the environment. As explained below, in 3.6, in vision slight changes of viewpoint, saccades and alterations of orientation can simultaneously change many details of information acquired, as well as altering global features (such as bringing an entirely new object or process into view); whereas if you are listening, with eyes shut, to a string quartet, waves crashing on rocks, or the wind howling through trees, your own small movements don't have such an effect on what you can learn about the environment. (Blind users of a stick to tap surfaces, and those who have learnt to use clicking noises for echo-location are probably closer to the case of vision, but I don't have personal experience of that sort of expertise.)

So visual experience supports sensory-motor theories such as (O'Regan & Noë, 2001) more than ordinary auditory experiences do. When large movements are involved, both can have significant effects on sensory contents, though in different ways. Blocking or cupping your ears can change how things sound, just as shutting or nearly shutting your eyes can change how things look, but both of those are mostly cases of altering the amount of information or distorting the information.

Remembering what you saw or heard a short time ago, and reflecting on it provides yet another change of contents of introspection where the differences described above between visual and auditory contents no longer hold, though there are other differences of course. There is lots more to be said about different sensory modalities, including such things as smell, taste, vestibular sensing[p] (very important for motion detection and balance, as explained in (Berthoz, 2000)), and also various kinds of illusion and paradoxical contents. For instance, what is perceived in an Escher drawing (such as the "Waterfall") has a consistent and real 2-D content, many locally consistent 3-D contents, and a globally impossible 3-D content; where the inconsistency of content may or may not be detected by the perceiver. Even more odd in some ways are motion after-effects: after a pattern of dots has moved steadily from left to right for a while then stopped, motion from right to left is experienced without anything visible changing its location relative to anything else: motion without anything moving. This helps to show that the function of perception, at least in humans, is not to produce a model or replica of what is in the environment, but a more sophisticated and subtle collection of transformations of information to serve many different purposes, using different kinds of information-bearers, which are very different kinds of entity from the things they represent, as explained in (Sloman, (to appear).

A more thorough survey of types of contents of consciousness, illustrating the polymorphism of the concept, will have to be left to future research. However, it will be useful to add more detail to the account of vision, in order to explain why some people have wished to distinguish P-C from A-C.

3.6. *Examples of effects of motion*

Consider a person with normal (adult) human vision, who is awake with eyes open, in a well-lit environment, with walls, chairs, tables, windows, cups, bowls, cutlery, etc. made of different sorts of materials found in our culture. If such a person moves sideways very slightly, what he is conscious of in the environment (whether he notices that he is conscious of it or not) will

[p] `http://en.wikipedia.org/wiki/Vestibular_system`

change in very many small ways that depend in exquisitely detailed ways on what things are in the environment, and what the movements are.

For example: some portions of surfaces that were previously visible become occluded by nearer objects; some that were invisible become unoccluded; highlights and reflections from shiny surfaces move slightly, in ways that depend on where the light source is, the orientation of the surface, its curvature, whether the curvature is convex or concave, and the viewpoint; ratios of distances and gaps, and projected angles also change. If a light source is behind the viewer, edges of his shadow on the surfaces of some objects may move, or change their fuzziness, the precise changes depending on shape, orientation and distance of the shadowed surface.

If instead of moving sideways you move forwards or backwards, the many effects are different: there are optical flow effects involving expansion and contraction, and your shadow may grow smaller or bigger. Tilting your head has different intricate effects on your experience. Changes of similar general sorts, but with characteristic differences will occur if you are stationary but something in the environment moves sideways, up, or down, or tilts or rotates.

If you make a saccade or large head rotation, not only does that change what is and is not in view, it also changes what you can do with different subsets of the information: items that you could process in detail lose resolution whereas others become more detailed and specific. For instance a word that was visible but unreadable because it was in the visual periphery may become clearly readable after a head rotation, or a sideways movement. Other things may lose resolution – whether that is noticed or not.

A perceiver need not be aware of all the changes that occur in the contents of visual consciousness. For example, if you have one eye missing or covered with a patch there will be a gap in the information available about the environment because of the blind spot, and a movement can change what is missing. However neither the gap nor the change in what is missing will usually be noticed, except in very special cases where a small motion allows the disappearance of a distinct object to be detected. What this illustrates, as do phenomena of change blindness (O'Regan *et al.*, 1999), is that the mere fact that something is accessible, or that something changes, does not suffice to make a perceiver conscious of it. The information that is available may simply not be processed because the mechanisms required to detect it are not available or are not used. For example, an animal whose visual field is constantly changing may be constantly reacting to the current contents, but have no awareness that anything is changing. For awareness of change to occur it is necessary for previous contents to be stored and compared with new contents, as is done at a very low level by optical flow detectors. Some simple visual systems may have no such storage: their behaviour will be constantly reacting to changes but they will have no experience of anything changing, though all their experiences are constantly changing. From this viewpoint, change blindness is not puzzling: it is change awareness that needs to be explained by describing mechanisms able to detect and record changes and showing their role in a larger architecture. However, for organisms that have such mechanisms there can be a problem of explaining its limitations – as O'Regan has done. Chapter 10 of (Sloman, 1978) has further discussion of varieties of reasons for information not to be available to consciousness.

3.7. *Multi-layer perception*

Another feature of the contents of visual experience is that it can simultaneously include information at different levels of abstraction. For psycholinguists this has long been obvious with regard to spoken language: you simultaneously hear the acoustic signal, low level phonemes, syllables, words, phrases and whole communications such as questions, requests, statements, commands, etc. It is obvious that similar concurrent processing of information at different levels of abstraction also applies to reading text, except that the lower levels are different. A demonstration of this was the Popeye program described three decades ago in chapter 9 of

(Sloman, 1978) – though there were too many short cuts, including absence of perceived motion, for that to be any more than a demonstration of concept. The key point is that experienced "higher" levels of abstraction can include ontologies that are not definable in terms of the lower level ontologies: one of many reasons why symbol-grounding theory is false.

Looking at a wide variety of scenes through a narrow tube shows that the information available in a small region of the visual field is highly ambiguous as regards the distances, orientations, curvatures, and other features of visible surfaces. It seems that evolution has found ways of using the very rich and complex network of constraints relating details of what is in the environment, details of the contents of different portions of the optic array, details of the "projection" processes, details of changes produced by different kinds of motion, and information acquired previously about types of entity and types of process that can exist in the environment, to enable a very complex, not consciously accessible, parallel multi-level constraint-propagation mechanism to find a globally consistent interpretation of all the information fragments at high speed. Once formed, different layers in that multi-layer interpretation can be changed concurrently by movements of objects perceived and movements of the perceiver, because the interpretation is not a static structure but a dynamic one, as needed for perception of a dynamic world (Sloman, 2009a).

When the environment is of a familiar type, containing known classes of objects (e.g. a domestic scene, or a walk in a forest, or humans interacting socially or performing familiar actions), previously learnt constraints linking details of familiar structures and processes can be combined with the more general topological and geometrical constraints arising out of the 3-D structure of the environment to facilitate rapid convergence to a coherent interpretation. That is why even static scenes and still photographs, or pencil sketches with most of the details missing, and with no dynamism, can all be interpreted. How all that works is still mostly a mystery – to psychologists, neuroscientists, and AI researchers, though many of them ignore the big picture because it is possible to make what looks like progress on special cases. The existence of multiple, concurrently active perceptual layers using very different ontologies and forms of representation was referred to loosely as "multi-window" perception in the AWMC paper, and contrasted with "peephole" perception, which expresses a view of perception as simply reception and low level processing of sensory input. The requirement for multiple levels is neatly illustrated by well-known ambiguous figures such as the Necker cube and the duck rabbit: when the experience of looking at such a figure "flips" between the two views, the low level perception of marks on paper remains unchanged while a layer of interpretation changes in ways that can be described only by using an ontology that goes beyond 2-D image features. In the case of the Necker cube describing the flip requires an ontology referring to depth and orientation in 3-D. In the case of the duck-rabbit describing the flip requires an ontology referring to two species of animal, their body parts (eyes, ears, bill) and the direction in which they are facing. This is partly similar to flipping between two parse trees for an ambiguous sentence, but the visual case is far more complex.

The many published architecture diagrams that, unlike Figure 1 in AWMC, include a small box labelled "perception", or "sensing" are indications of the general tendency to ignore, or not notice, the more complex phenomena of multi-layer perceptual consciousness and perceptual function.

3.8. *Beyond James Gibson*

James Gibson, as far as I know, was the first to emphasise and explore the richness of changing visual percepts and their importance for obtaining information about the environment relevant to the perceiver's actions (Gibson, 1966, 1979), although some of the ideas go back to the gestalt psychologists and earlier, and some of the details have gone unnoticed even by Gibson. He described the information gained as being about positive and negative affordances for the

perceiver. I have argued (Sloman, 2008a) that the situation is far more complex than he noticed, since perception of "proto-affordances", i.e. possibilities for change and constraints on change, need not have anything to do with the perceiver's goals or actions. It is also possible to perceive "vicarious affordances", including dangers and opportunities for others, for instance one's children, or enemies.

Moreover, the ability to notice and reason about possibilities for and constraints on changes of configuration seems to be at the root of certain kinds of mathematical competences, for instance being able to create proofs in Euclidean geometry. As every mathematician knows, once that ability has developed it can even be deployed on purely imagined structures – which is why actually doing mathematics does not require physical embodiment in a machine with sensors and motors.

When you move there are, as noted previously, differences in the information available about the environment, whose description requires use of an ontology of 3-D objects and processes involving many different kinds of material and causal relationships. There are other subtle differences, which can be described as differences in *your experience*, whose description requires a very different ontology, though ordinary language is not rich enough to characterise those differences in complete detail. Examples include the famous changes in aspect ratios and angles that occur to the 2-D projection of a 3-D circle to give the *experience* of an elliptical shape, and the loss of rectangularity of the visible outer boundary of a rectangular table top as you move from looking directly down on it to viewing it from the side.

There are many other differences that are much harder to describe, because they involve changes in the appearances of objects with curved surfaces, where the experiences of perceiving gradually changing surface orientation or curvature are very complex and subtle, and form patterns of continuously but non-uniformly varying change spread across regions of the visual field.

These are phenomena that any artist has to learn to attend to in order to be able to draw or paint well, and they need not be correlated with linguistic capabilities, for instance in autistic children who are gifted artists. The variety and richness of such phenomena support J.L. Austin's remark: "Fact is richer than diction".

Besides the changes in the contents of perceptual consciousness produced by changes of viewpoint there are other changes, produced by such things as: saccades (mentioned earlier), interposing a refractive lens, loss of accommodation due to age, or drugs, exhaustion, switches of attention, and many more. The saccades and switches of attention can produce changes that are hard, or impossible to notice, including changes in available detail in different regions of the visual field. The existence of, and changes in the location of, the blind spot also cannot normally enter consciousness.

The explanation for the undetectability seems to be the (subconscious) construction, in low level sensory processing virtual machines, of enduring multilayer information structures only loosely related to the contents of retinal stimulation at any moment. A particular theory that is capable of explaining the invisibility of the blind-spot (though the theory's author, Arnold Trehub, did not notice this!) is in (Trehub, 1991). I doubt that that theory is correct in detail, but it reflects some aspects of what I called "multi-window" perception, and some of its features seem to be required by any theory of how (adult) human visual perception works, including preservation of information, at different levels of abstraction, about structures and processes in the environment, while details of retinal contents (or the optic array, or brain area V1) change, e.g. because of saccades.

This can, as Trehub points out, explain some discrepancies between the actual information available in the optic array and the information accessible through introspection. Some of those differences take the form of optical illusions about relative size (e.g. the Ebbinghaus illusion) which can occur while non-illusory variants of the information may be available for visual

servo-control of grasping, though precise details of the phenomena remain controversial.[q]

3.9. *Visual control information*

As explained in (Sloman, 1982) and Section 3.4, besides being used for producing *descriptive* information about the contents of the environment and the contents of the visual field, the visual information can also be used to produce *control* information, which is often more simply based on changing 2-D projections than on 3-D interpretations. How much of that control process is introspectible is not clear, although some is, since that is involved in knowing what you are trying to do. What is introspectible seems to be alterable through training, as many musicians and athletes have discovered. (Luc Beaudoin has reminded me that this change can go in both directions – some things become introspectible and others cease to be introspectible, as expertise develops.)

Other details of sensory processing whose accessibility can be altered by training are the ways in which low level features are processed to produce intermediate results, for example in learning to read for the first time, learning to read a new language, learning to see the differences in a pair of identical twins. Such learning can also alter the contents of auditory experience, e.g. learning to hear differences in features of linguistic utterances between speakers with different accents or non-normally noticed differences in low level features of one's own language, like the changes in pronunciation of the "d" in "good morning" and "good tea".

These examples provide reminders of the kinds of facts about introspectible experience that originally led philosophers to notice that, besides the entities in the environment that are objects of perception, there are entities, sometimes called "sense-data", or "qualia", or "sensibilia", that are not in the environment but can also be noticed, compared, recognised, described, etc. and which can change in different ways from their external correlates.

3.10. *Disconnected dynamism*

The discussion so far has been about how the contents of consciousness, including different layers of abstraction in what is experienced, are in some cases intricately related to structures and processes in the environment with "phase-locked" or "entrained" changes linking the two.

One of the features of humans, and I suspect some other intelligent species, is that some of these multifarious internal changes can be decoupled from sensory and motor signals, as suggested in (Craik, 1943). There are various reasons why this can enhance functionality, including the ability to plan several steps ahead, to construct hypotheses about what might have explained some observed situation, to construct theories about what unobserved entities could be causing observed phenomena, for posing questions, for formulating goals, and for learning by retrospective analysis of what did and did not happen. Once such mechanisms became available for particular functions their roles could have evolved and developed to provide many of the familiar aspects of human consciousness, including day-dreaming, doing mathematics with your eyes shut, playing games like chess, inventing stories, designing new kinds of machinery, thinking about the inaccessible contents of other minds, and many more.

And many of these purely internal processes and the structures they create and manipulate are capable of being attended to in an architecture with meta-management capabilities (Beaudoin, 1994). Such processes will produce contents of consciousness that may be partly like those of perceptual consciousness and partly different. The fact that they can exist without external causes has led to some of the details of philosophical theories of qualia and phenomenal consciousness, discussed later.

A consequence of the possibility of this sort of disconnection is that any theory that attempts to link consciousness and *all* its contents to sensory-motor processes must be wrong, even

[q]See `http://en.wikipedia.org/wiki/Ebbinghaus_illusion`

if it correctly describes a subset of the phenomena. Likewise a theory of consciousness that does not allow the contents of perceptual consciousness some of the time to be a multi-layer intricately changing web of causally connected information structures, partly controlled by the environment, and partly by the ontology and knowledge of the perceiver, must also be wrong.

3.11. *The causal relationships of contents of experience*

Past philosophers have argued that the introspectible changes that occur in us as our experience changes have a mysterious status, but we can now view them as changes in the contents of intermediate information structures in virtual machines in a multi-layered perceptual system, where the different layers use different ontologies, and perform different functions, and together reduce the problems of going from sensory signals to usable interpretations of 3-D structures and processes. Such information can be usable in many different ways, including controlling actions, triggering actions, suggesting new goals, revealing new obstacles or constraints, contributing to plan formation or evaluation, refuting or confirming theories or expectations, providing information someone else needs, reducing ambiguity in another part of the perceptual field, and many more.

Only some of those contents will be accessible to introspective mechanisms at any time, and which are accessible will depend on the self-monitoring capabilities of the whole system and of various sub-systems. These capabilities can change both dynamically during switches of tasks and over extended periods, through development and learning (e.g. learning to become an artist, or learning to hear more subtle features of language by studying linguistics, or learning to play a musical instrument more expressively, or in better coordination with others). Sometimes accessibility can be changed by a question someone else asks.[r] The changes, and the information contents, have objective existence and can be studied theoretically like other theoretical entities by scientists and (perhaps) replicated by engineers, but they also have a kind of "privacy" because they exist in virtual machines, and their contents are not accessible except from within the virtual machines. External measuring devices may detect physical or physiological implementation details, but that is not the same as detecting information contents.

Some aspects of the categorisation of internal introspectible states are "incommunicable", because the categories used arise from the internally created self-classification sub-systems using labels with causal indexicality, as explained in (Sloman & Chrisley, 2003). I suspect that the existence of additional connections during such self-training processes could account for inappropriate classifiers later being activated, so that, for example colour recognisers get triggered by non-coloured stimuli, because some additional links were active while the colour system was training itself: the result is synaesthesia.

3.12. *Internal/external ontological overlap*

Some aspects of the ontology used for describing internal structures and processes can be equally applicable to internal and to external phenomena, because they are instances of the same abstract concepts. For example, entities on a road may be ordered along the road, and can have an order of size, and events in a virtual machine may be ordered in time and also in intensity, or size in the virtual machine, for instance if experienced circles are nested. Other topological relations such as connectivity, containment, overlap, etc. can be equally applicable internally and externally. It follows that not all aspects of internal entities are describable only privately. (Such phenomena are sometimes misinterpreted as uses of metaphor rather than uses of an abstract widely applicable ontology.)

Long term changes in the whole system may be of different kinds, including changes in the forms of representation available, changes in the ontologies used for various subsystems,

[r] As illustrated in `http://www.cs.bham.ac.uk/research/projects/cogaff/misc/unconscious-seeing.html`

changes in the intermediate processing mechanisms (e.g. algorithms available) and changes in the architecture as new layers of processing develop, or new forms of information-processing develop, or new access routes between sub-systems develop. Some of these changes in forms of representation, mechanisms and architectures may occur across evolution rather than in individual development.

Moreover, rather than all introspection (i.e. self-monitoring of internal structures and processes) going through some single global control subsystem there are likely to be many subsystems monitoring other subsystems and using the results for purposes that include interpretation, inference, learning, control, planning, problem-solving, debugging and possibly other purposes, as suggested in the writings of Minsky (Minsky, 1987, 2006; M.Minsky, 2005). The internal entities we claim are contents of our introspection will be the ones that are accessible by subsystems that have relatively broad functions, including being used in communication with others, remembering for future internal use, and resolving conflicts of motivation, as suggested in chapter 10 of (Sloman, 1978), (Johnson-Laird, 1988) and many other publications. But there is no reason why there should not be other introspections that are not so globally accessible but can be provoked to communicate their contents in special circumstances, as illustrated by Figure 1 in (Sloman, 2008a) discussed in section 1.6 of that paper, and in Note r above.

An example some readers may recognise is being reminded of someone or something without knowing why you were reminded. Then later you may notice what it was. Presumably the subsystem in which the old association was triggered provided information about something being reactivated, for a different subsystem, but without reporting all the details. Later something may cause the details to be explicitly represented and transmitted.

3.13. *Infallibility and P-C*

Most of what I have been doing is taking what are either aspects of common experience or things philosophers have drawn attention to, and then showing how those are all to be expected in a well designed machine with human-like capabilities, because of the roles those aspects play in the functioning of the machine, or in some cases the side-effects of those roles.

However, once we get into philosophical mode and start talking about such things, without having been educated about AI and the designer stance, we may be sucked into accepting various obscure claims uncritically.

An example is the feature of contents of consciousness that is often described as infallibility. A truncated version of the philosophical claim might be that I can be mistaken about what colour or what size a box on the table is, but I cannot be mistaken about what I am conscious of, e.g. what colour it seems to me to be, or what size the box seems to me to be. This has led some thinkers to claim that a feature of P-C is infallibility which results from unmediated access to the information. However, I think that in such cases what is really a trivial tautology is being inflated into a metaphysical mystery. The trivial tautology is that you cannot be mistaken about how things seem to you, and it is a tautology because it is very similar to the tautology that every voltmeter is infallible about something, because it cannot be wrong about what it senses or calculates the voltage to be, even though it can be wrong about the voltage.

There is much else that philosophers and others have written about P-C in its various guises, as qualia, sense-data, introspective contents, or whatever, and a subset of it is, as I have indicated but not proved, true and explicable in terms of the functioning of very intelligent robots of the future. But there are other parts that are merely confusions, and the infallibility claim is one of them.

3.14. *P-C as a dispensable ghostly companion to A-C*

All the contents of experience that I have described have the ability to be causes and to have functional roles, including the roles described by Gibson. So that makes them really examples of A-C, and the idea of P-C can be abandoned, as promised in Section 3. However we are expected to believe that in addition to all these occurrences of A-C there are also parallel occurrences of P-C, the non-functional, and therefore causally inept contents of consciousness and states of being conscious of them that are supposed to define P-C. Block, Chalmers, Velmans, and many others have attempted to defend this P-C notion by alluding to the conceptual possibility of a machine or person having all the rich details of A-C yet lacking this something extra: this is the possibility generally referred to as a zombie. That argument depends on the claimed ability to imagine something containing all the rich and detailed visual information, changing in close coordination with movements of the viewer, without having P-C, i.e. "without there being anything it is like" to have and be able to attend to, use, remember, think about, and describe the information contents.

The claim that instances of P-C may or may not exist when all the above A-C phenomena exist, is very like the claim that undetectable fairies may or may not exist when all the visible biological growth occurs in the garden. The difference is that defenders of P-C believe that whereas the fairies mentioned above, and their uncontrollably multiplying counterparts are mere hypothesis, the contents of P-C are directly experienced by them, so they *know* that they exist, and such things really might be missing in a zombie, even if it claimed fervently that it too had direct access to P-C. What this amounts to is saying something like: "I know what I am referring to – it is *this*" said while pointing inwardly.

But that is like the defender of the enduring region of space (RoS) described in 2.4 saying he knows what he is talking about because it is what he referred to when he previously said "here". In both cases, something exists that we all agree on, but in both cases that which exists lacks some of the features some people are inclined to attribute to it: at any time a RoS does have relations to a host of space occupants, but it does not have an enduring identity over time independently of all its spatial relations to various frames of reference or spatial occupants.

Similarly, the P-C is alleged to be like an A-C and to accompany an A-C, but it is also claimed to have an intrinsic nature independent of its causal relations both to things that precede and help to cause it, and to things that it can cause subsequently. In this case, the A-C accounts for all the phenomena, including the belief in P-C. Adding the P-C involves the kind of self-delusion involved in attributing enduring identity to a RoS, and the kind of unjustified redundancy involved in adding invisible fairies in gardens to theories of how plants grow.

3.15. *Why A-Cs are required, not P-Cs*

Moreover there are aspects of the phenomenology of perceptual experiences (especially visual experiences) that clearly require A-C and are incompatible with the causal/functional disconnection of P-C. Perception of processes and affordances provides many illustrations. I'll stick with processes: perception of a process, e.g. of something rotating, or moving from left to right involves not just a sequence of disconnected phenomenal contents, but an integrated (often not discrete) temporally extended set of states and relationships where what is experienced at one time helps to define what is experienced later as a change of a certain sort, and in cases where things like rigidity and momentum are involved the experience involves causal connections between happenings across time. If that is correct, defining P-C to exclude causal/functional roles has the consequence of excluding some of the most important contents of human consciousness, including not only in perception of everyday phenomena involving motion, but also experiences such as watching a ballet or hearing a piece of music. I conclude that given all the expectations heaped on the notion of P-C, which it is incapable of fulfilling because of the constraint of functional/causal disconnection, it is an example self-defeating

semantics, like the concept of the time at the centre of the earth, or the time on the sun.

A corollary is that insofar as Chalmers (implicitly) uses the concept of P-C to define what he calls "the hard problem", it must be a bogus problem, as described below in 3.20, just as the problem of intrinsic identity of a RoS across time is bogus.

This will not convince a firm believer: but my aim is not to convince anyone, but to explain what I was getting at when I talked about describing, explaining and modelling "everything else" in the target article (AWMC): at that point I meant everything else but P-C, but did not say so clearly enough: I foolishly assumed readers would recognise the target.

There are many, but diverse, ordinary uses of the words "conscious" and "consciousness" that are not semantically flawed, but since they express polymorphic concepts there is no unique entity or type of phenomenon they refer to that can be the target for scientific research or machine modelling. Because that use of the word "consciousness" covers diverse phenomena, trying to produce a machine model of consciousness in the everyday sense is analogous to trying to produce a machine model of efficiency, or tallness, or danger, whose labels also refer to different things in different contexts. Likewise trying to explain how consciousness evolved is like trying to produce one explanation of how efficiency evolved in organisms, disregarding differences between mechanical efficiency, thermal efficiency, efficiency of conversion of chemical energy, reproductive efficiency, and many more.

3.16. *Two-way physical-virtual dependence*

Although contents of virtual machinery in computing systems are not definable in terms of concepts of the physical sciences (e.g. concepts like "attacking", "winning" and "losing", used in describing many games, and the game-specific concepts such as "king", "pawn", and "checkmate"), the conceptual gap between the virtual and the physical does not imply that causal relations are impossible. The fact that computers are widely used not only to compute mathematical functions, and process numerical data, but also to control complex machinery on the basis of sensor information and behaviour policies (sometimes the result of previous learning by the machine), depends on the existence of causal links between virtual and physical phenomena, described in more detail in (Sloman, 2009b,c,d). (I am constantly amazed at how narrow-minded many philosophers are who regularly make use of such virtual machines, yet completely ignore them when discussing mind-brain relations.)

It is important to note that causation goes both ways: not only do the virtual machine phenomena (including everyday examples of consciousness) depend on physical machinery, it is also the case that the physical machines underlying those virtual machines cannot function as they do without the virtual machines that run on them and control them. produce physical effects, e.g. poverty causing theft, and jealousy Your body cannot get across a gap that's too wide for it to jump, without the use of mental processes that lead to the use of a long plank across the gap, just as your computer screen cannot display the spelling mistakes as you type your essay, without the virtual machine that forms the spelling checker subsystem in the larger virtual machine that is the word processor, running on the virtual machine that is the operating system, and interacting with hybrid machines such as devices and their firmware drivers. Biological mechanisms are probably different from this (and richer) in hundreds of ways that still need to be understood. Likewise socio-economic virtual machines can include states and processes that causing wars.

3.17. *Lack of coherence of the concept*

Not only is it the case that many individual researchers use the word "consciousness" in a semantically flawed way, there is also a great lack of coherence among researchers who profess to be studying consciousness insofar as they use the label "consciousness" in so many different ways (illustrated well in the special issue of the *Journal of Consciousness Studies* (16(5), 2009)

on "Defining Consciousness". It seems to me that most of the contributors to the journal don't even read most of what has been written by previous contributors.

Some consciousness researchers claim explicitly to be referring to what Block called "p-consciousness" whereas others feel it is not necessary to specify what they mean, except perhaps by giving a few examples, because they assume everyone knows the word. Or they use some other phrase that they assume everyone understands the same way, e.g. a form of words like "what it is like to be X". It would take too much space to document in detail the obscurity of that characterisation, and the diversity of meanings given to the noun "consciousness" when used by scientists to specify their research goals, but it is well known and often commented on (e.g. as motivation for the special issue of JCS). Researchers who ignore this lack of coherence risk investigating or attempting to model something for which there is no agreed label, while disregarding many of the phenomena to which the label has been attached by researchers with different interests.

3.18. *The semantic flaws of the P-C notion*

In some philosophers beliefs about consciousness are expressed using forms of expression such as "what it is like to be...", "phenomenal consciousness", etc. which assume that those forms of expression have identified a concept of P-C whose existence is not analysable in terms of patterns of causal relationships to other phenomena. I think that is unwitting self-deception (not uncommon in philosophy).

I am not saying there is no such thing as contents of consciousness in an ordinary everyday sense, such as someone becoming conscious that it is getting colder, or becoming aware that she has the uneasy feeling of being watched. Novels, plays, poems are full of descriptions of contents of consciousness of many different kinds. I am saying that there is no coherent concept of a P-C which can *only* be identified as the momentary contents of an internal act of introspection and whose existence is independent of causal and functional links to a host of other physical and mental phenomena, including dispositional mental phenomena, of the kind discussed by Ryle (in 1949) in connection with imagining a well known tune in your head.

I am also not denying that introspection is possible, nor denying that it is a source of knowledge, nor denying that the objects of introspection exist, just as I am not denying that regions of space exist when I deny the coherence of a notion of absolute identify of RoS. In the ordinary use of the word, introspection and its contents are causally connected with other things, and therefore would not be examples of phenomenal consciousness defined to be functionless.

In the case of the "hard problem" the claim is that the correlation between other phenomena and P-C is just a brute metaphysical fact, so that all we can do is look for correlations with other things, such as brain events and other mental events and processes, but we can't identify P-C in terms of relations to those other things (like the enduring portion of space). By definition of the hard problem: no matter which other things occur in an individual, P-C might logically be missing and the individual would then be a "mindless" zombie, but nobody would know, not even the zombie. Chalmers thinks there are laws of nature linking physical events and occurrences of various P-Cs, but acknowledges that investigating such things is difficult because each P-C is intrinsically private, exists only for an instant, and is accessible only through a mental state, then vanishes, though it may, for Chalmers, have effects, unlike epiphenomenal phenomenal consciousness and the type of qualia some philosophers discuss.

Other philosophers (metaphysical idealists) go further and say that the P-C can exist independently of any physical matter, a claim whose content will necessarily vary as physical theories about the nature of matter develop. This is not to be confused with the superficially similar "multiple-realisability" claim that virtual machines, and therefore also A-C experiences, do not require any particular sort of physical implementation, since they, like many virtual machines in computers, can be supported by different physical machinery. It is not clear what

evidence there could ever be for total matter-independent existence of mental phenomena. "Out of the body" experiences have been proposed as tests, but discussing them would be a digression here.

3.19. *Why science cannot answer questions about P-C*

Debates related to these issues have a very long history in philosophy. Debates about P-C are very different from debates about A-C. Attempts to answer questions about P-C by doing scientific research or computational modelling completely miss the point, e.g. if researchers think they are solving the "hard" problem, or modelling "phenomenal consciousness" in the philosophers' sense.

Philosophical concepts like P-C (and some uses of the word "qualia") are *designed*, by definition, to resist such solutions or such modelling, which is why the debates about them are intrinsically endless (and I would say pointless, except for their use in philosophical tutorials teaching students how to detect disguised nonsense), just like debates about where the previously identified point of space now is. Of course, the concepts can be re-defined so as to be amenable to scientific research (which is what Einstein did as regards spatial locations), but that's where different researchers go in different directions, while claiming to be addressing *the* problem of consciousness.

AI researchers, and also neural modellers, who think they are modelling phenomenal consciousness in Block's sense, or modelling "what it is like to be..." in the sense of Nagel, by some special new feature in their machine design, have missed the point about how the entity in question (P-C) is *defined* by the original discussants so as to be disconnected (both definitionally and causally) from everything else that we normally think of as happening in brains and minds, and therefore incapable of being modelled in the manner proposed.

Researchers and modellers who deny this must be using some *other* definition of P-C, or qualia, or phenomenal consciousness, or whatever. And some of them invent, or borrow, definitions that suit their preferred research projects. That would not matter if they knew they were doing that and acknowledged it publicly, instead of claiming to be solving a well-defined pre-existing problem.

3.20. *Chalmers and his hard problem*

Chalmers adds an extra wrinkle to the P-C theory by claiming that P-Cs exist, and are not physical, and there is no intelligible explanation of why they are associated with particular physical phenomena, so the only way to investigate his P-C is to find which ones co-occur with which brain states and processes.

But it seems that by definition this cannot be a subject for science as normally construed since only the experiencer can tell whether a P-C occurs and if there are different experiencers there is no way of telling whether there is anything in common between their P-Cs since by definition P-C is disconnected from anything measurable or observable: it can only be felt, or had, or experienced. It is very like the region of space that you can only identify by pointing and saying "here". A RoS can be re-identified relative to a framework of reference involving enduring objects, but the notion of enduring *intrinsic* identity for a RoS is self-defeating.

3.21. *How to create something close to p-consciousness in a machine*

Suppose I can specify kinds of virtual machine phenomena that arise necessarily in information processing systems with a particular sort of architecture. For example, consider a robot with a sophisticated multi-layer visual system, of the sort described in Section 3.7, that not only produces information about what is inferred to exist in the environment (e.g. "a rectangular table in front of me blocking my way to the doorway, but with a space on the right that I can get

through", etc.) but also creates and temporarily stores a host of intermediate data-structures related to various partial stages of the processing of the visual and other sensory data, including information-structures in the optic array, some of which have information about *appearances* of objects rather than the objects themselves.

Suppose the robot's virtual machine architecture includes a meta-management (or reflective) sub-system that is capable of inspecting and summarising the contents of some of the intermediate data-structures in the sensory system, e.g. some of the visual structures that record the partly processed richly detailed 2-D sensory input prior to the construction of 3-D interpretations and other interpretations in terms of affordances, causal connections, etc. Suppose also that it can switch between inspecting different aspects, of the intermediate sensory data, including changes in those intermediate information structures tightly correlated with the robot's movements. And suppose finally that in summarising and recording what's going on in its intermediate structures it uses descriptive features that it has developed itself for categorising those structures (as suggested in (Sloman & Chrisley, 2003)). Then such a design would produce many phenomena that match very closely the phenomena that philosophers have used to explain what sense-data or qualia or the P-C contents are.

If in addition the sub-system can detect the similarities and differences between both perceptual information that is about independently existing objects in the environment (e.g. rectangular tables) and the robot's internal information about the experiences those objects produce (e.g. changing non-rectangular 2-D appearances of a rectangular 3-D table-top), then the robot would have a basis for inventing the philosophical distinction between "primary" and "secondary" qualities of things (like John Locke), as well as rediscovering problems about them.

For instance the robot might notice and record existence of 2-D intermediate structures with acute and obtuse angles corresponding to the 3-D surface of the table with right angles, where the 2-D angles in the intermediate structures (and also the length ratios) change as the robot changes its position or as someone rotates the table without altering the shape of the table. Likewise there would be extended features that vary across the surface, some caused by grain in the wood, some by shadows, some by reflections and highlights, some by different coloured paint or varnish which the whole visual sub-system could identify and locate in 3-D space, while the intermediate 2-D structures would include intermediate qualities that are used to derive the objective descriptions, and which could also be attended to by the meta-management sub-system, and distinguished from features of the external surface. Compare (Austin, 1959). For example, highlights and reflected wavelengths, and sometimes shadows, would change as the robot moved around the table and that would help the visual system infer the intrinsic physical properties of the surface, like texture, curvature, transparency, reflectivity, and colour.

A result of having not only the minimal capabilities required for seeing and avoiding objects (like some current robots), but also the meta-management capabilities with access to enduring records of low- and intermediate-level stages in visual processes, would be that *the robot knows what it is like to be moving round a 3-D table!*

The products of inspection by the meta-management layer, including both information about the environment (e.g. the fixed rectangular shape of table top) and information about the changing 2-D features in the intermediate sensory information structures would have all the describable features of P-C required by Chalmers theory, and would *necessarily* be causally connected with physical processes going on the machine's brain or computing system, just as the contents of virtual machines such as spelling checkers or meeting planners, or chess players are causally connected with physical processes in the computer in which the virtual machines are implemented.

The italicised word "necessarily" is justified because that sort of connection is exactly what software engineers know how to produce in the virtual machines they create. When debugging faulty systems they know how to make changes that *guarantee* that certain mistakes will not

recur (e.g. violating a rule of chess), as long as the hardware functions normally and the running virtual machines are not tampered with. So we have a collection of entities and processes that are not physical but are implemented in physical machinery, and which are causally connected with processes in physical machinery in *intelligible and predictable ways*. This means that they cannot be the causally disconnected phenomenal qualities of some philosophers, nor the P-C's that define Chalmer's hard problem because his P-Cs are only *contingently* connected with physical processes and the causal relations can only be studied empirically, if at all: they cannot be worked out.

As far as I understand it, Chalmer's theory entails that anything produced in computers or robots in the manner I have described would not be an example of P-C (or phenomenal consciousness) and could not solve or contribute to the hard problem. But that leads to a problem for him, which, as far as I know, nobody has noticed.

Incidentally, I think new forms of representation and new forms of information processing will need to be invented/discovered before we can give robots virtual machines with these meta-management capabilities. Just giving them the current video-technology and even things like current recognition and tracking software systems will be grossly inadequate.

3.22. *The potential duplication implied by Chalmers P-C*

The claim that P-Cs exist but their existence cannot be demonstrated to be a consequence of any physical organisation has a bizarre consequence.

Suppose we were able to demonstrate that a certain physical machine (PM1) did implement a kind of virtual machine (VM1) meeting a complex specification, relating its states and behaviours to very many other states and processes, in the manner required to produce a full explanation of the workings of a human mind, including all the cognitive functions, after centuries more research on the variety of types of thing that occur in minds. Then, if Chalmers' idea is coherent, there would be a residual problem that we could not answer, namely whether or not there is *in addition to VM1* some *other* entity or set of entities (P-C1) with the ontological status required by Chalmers.

So in addition to all the intelligible, and controllable, virtual machine processes and causal interactions within VM1 and between components of VM1 and PM1 or the external environment, there could be another collection of ongoing processes P-C1. Its changes, by hypothesis would be caused by changes in PM1, but its effects would duplicate those of VM1, and it could never be shown to be a necessary consequence of anything else going on in the machine. E.g. software engineers would not be able to explain why they occur. This is a very strange duplication: the virtual machine states and processes that play a crucial role in all the functions of the robot, and which are as intelligible as typical products of human engineering, would be paralleled by a host of P-Cs with the same structures and changing patterns, etc. which produce the same effects (e.g. memories, decisions, emotions, verbal reports) but which have a quite different metaphysical status because they are not necessarily connected with the design of the machine, even if correlations happen to exist. What's more, some philosophers demand that such parallel undetectable, functionless, P-Cs have to exist if the robot, is to avoid being a zombie. They are also needed to prevent humans being zombies.

I am reminded of invisible, intangible, and in all other ways undetectable fairies dancing at the bottom of the garden, celebrating all the floral growth, and perhaps helping it along somehow, which might have been a coherent hypothesis before we acquired our modern understanding of biological phenomena and the underlying physics and chemistry – which admittedly is still incomplete, but has done enough to make the fairies unnecessary.

The P-C contents are like the fairies, for they are not needed to explain anything, since every detail of what they are supposed to explain is accounted for by the dense network of causal relations in the information processing systems, physical and virtual.

The fairy story may make a sort of sense as an internally consistent story. But as part of any serious theory about what's going on in the garden there is no reason to take it seriously because, by hypothesis, we have an alternative account that explains all the details.

Likewise the description of the P-Cs of the hard problem makes a sort of sense as an internally consistent story, but is completely redundant as part of an account of how minds work. I think claiming that both the scientific explanation and the P-C theory might be true simultaneously is incoherent, or at least seriously semantically flawed, in a manner similar to the theory of the universe as expanding continuously but undetectably because all measuring devices and laws of physics change so as to hide the effects of the expansion.

3.23. *I should have been clearer*

In recent months I have discovered from conversations with respected AI researchers that all that philosophical history about P-Cs, phenomenal consciousness, and the hard problem (only crudely summarised above) is not as widely known as I had thought. I should therefore not have taken so much for granted when I originally wrote the AWMC paper for the 2007 workshop.

I did not intend to write anything that denied the existence of experiences (are there both subjective and non-subjective experiences??) nor have I ever recommended a move back to behaviourism or physicalism. In fact I gave several lists of *non-physical internal phenomena* that behaviourist and physicalist theories cannot accommodate; and I have elsewhere begun to show how virtual machine architectures whose behaviours are not externally visible can account for some of those phenomena. Moreover, virtual machines with certain kinds of meta-semantic competence in a meta-management system would be expected to produce if not exactly the sorts of states and processes that humans label "subjective experience", including awareness of having experiences, then something very similar in general character, and perhaps at a later stage of this research something exactly similar, or as similar as the experiences of different humans are to one another. That similarity would include dynamic features such as the way visual contents change while the perceiver moves around a room and the 3-D percept remains fixed. (If possible at all, this replication could be centuries away, not decades away).

3.24. *The complexity and multiplicity of the phenomena*

Quick and simple answers are not to be expected when we are trying to understand products of millions of years of biological evolution, many thousands of years of cultural development, and many days months and years of individual development and learning.

Some simple kinds of artificial self-awareness are familiar in engineered control systems and conventional computer programs. More sophisticated forms are newer. A digital thermostat could try to predict the effects of its signals to a boiler and then compare its predicted temperature changes with measured temperature changes. It could discover that its predictions are systematically wrong, e.g. because the temperature swings are higher than predicted. Depending on the physical and virtual machinery used for its design, it might be able to alter itself to make more accurate predictions on the basis of what it has observed, and also alter its control behaviour perhaps by making smaller more frequent changes which produce smaller swings. That form of learning would require a sophisticated virtual machine architecture not commonly found in thermostats (none that I have ever encountered) though the main ideas are not new and I think they will be increasingly found in adaptive controllers.

There are many different possible designs, including some using self-modifying neural net mechanisms. All of those are familiar ideas, and very many variants are possible in self monitoring scheduling systems in self monitoring planners, and even in self-monitoring self monitors (Kennedy, 1999), provided that they have the right architecture and the right sorts of meta-semantic competence to perform the required functions, which will differ from case to case. But whether that will satisfy researchers seeking P-C is another matter. I suggest

that replicating all the many A-C phenomena will suffice to explain how human minds work, including all the varieties of consciousness and unconsciousness.

3.25. *Self-seduction*

Notice that all those "self-" capabilities can be produced without having a special internal mechanism called a "self". For X to do self-monitoring or self-control requires X to monitor X or control X, not for something else, a self, to monitor or control X.

Of course there will be sub-systems that perform different sub-tasks, but labelling any of them as "the self" would just be an arbitrary linguistic change, and requiring one special "self" to be responsible for all self-monitoring and all self-modification would be a bad design. In humans the self-monitoring subsystem that manages your balance as you walk is distinct from the self-monitoring subsystem that corrects verbal errors as you speak or write (though it misses some of them), and both differ from the self-monitoring mechanisms constituting the immune system. Researchers should resist the seduction of the idea of a special unitary subsystem, the self. The only unitary self is the whole person, which may have different aspects, which we call mind and body (as noted in (Strawson, 1959)). But if Mary can hit John with a fish, then Mary can hit Mary with a fish, and we can express that unambiguously by saying "Mary hit herself with a fish" without implying that she hit some special entity inside Mary.

I have argued in (Sloman, 2008b) (following Hume) that talk of "the self", or "a self", is based on a failure to understand how ordinary languages work when they use indexicals or reflexive pronouns. The occurrence of words like "I", "me", "my" and "self" refer to humans, not special mysterious parts of humans. Reifying an "I" or a self is a bit like assuming that in addition to all the places we all know about, there are other special mysterious places referred to as "here", and in addition to ordinary times there are special entities referred to as "now". No doubt some philosopher somewhere has been seduced into talking about "the now" and "the here" as if they were peculiar entities distinct from ordinary times and places.

So when I say that the "Queen herself gave me my medal" I am not saying two things gave me my medal, one human being (with body and mind) and another special entity. I am simply denying that she got someone else to give it to me. And "Fred was pleased with himself" does not mean there were two things, Fred and the special thing he was pleased with inside himself. It just says that Fred was pleased with Fred – the whole Fred, not some mysterious special portion. (Freud's trinity: the id, the ego and the superego were different, for he was trying to construct an architectural theory while lacking suitable technical concepts. As AI develops we should be able to produce a much improved version of what he was trying to do. Compare (Minsky, 2006).)

3.26. *Conscious and unconscious contents*

An aspect of the theory that some may find confusing is the implication that what makes various entities, states and processes within the running virtual machine contents of introspective consciousness depends on whether those contents are accessed by sub-systems within the whole architecture concerned with summarising, reporting, and attempting to modulate, control, or improve performance. (The variety of such "meta-management" functions is still an open research topic.)

A consequence is that on this model there is no *intrinsic* difference between the contents of consciousness and some of the other information contents of running virtual machines. The very same information structure may at one time be introspected and at another time ignored. (This seems to be part of what happens during development of expertise as sub-processes in complex tasks become automated.) But even something that is not being attended to may nevertheless have the potential to be attended to (as in the demonstration in Footnote r): the information is there and being used, and on a similar occasion later on it might be monitored

and modulated – e.g. when a learner musician develops awareness of the need to modify the mode of control used when playing a particular passage previously performed fluently. This may require sensing and modifying muscular tension, about which information was previously available, and used (unconsciously) for other purposes, but not introspected. We could call the information structures that are sometimes contents of consciousness and sometimes not "Potential consciousness contents" (PCC). When the potential is realised they become examples of "access consciousness" (A-C).

Another consequence is that there may be subsystems accessing and using information in a PCC or modulating such information to perform a control function, which themselves are not accessed by the more global or central management mechanisms. In that case, we can talk of a *part* of a whole animal or robot (not necessarily a physically distinct part, but a virtual-machine that is a part of a larger virtual machine, being conscious of something that causes it to alter its (internal) behaviour (just as a fly's being conscious of something approaching it causes it to fly away) while the whole animal is not introspectively conscious of it.

So, while I do not deny the existence of any of the normal phenomena involved in what is normally called human consciousness in non-scientific contexts, the theory being developed implies that parts of a person (though typically not physical parts) can be conscious of (make use of information about) something that the person is not conscious of.

All this implies that there can be many things going on inside an individual that are intrinsically very like things whose occurrence that individual is aware of, but simply happen not to be noticed by the individual either because the individual lacks the self-monitoring mechanisms required, lacks the meta-semantic concepts required for formulating them or because the self-monitoring mechanisms happened to be directed elsewhere at the time. (Much of this was said in chapter 10 of (Sloman, 1978).) The minimal similarity between introspectively accessed contents of consciousness and the PPCs that happen not to be accessed is that both are usable, and potentially useful, information structures generated in virtual machines. In addition an aspect of the visual field that is introspected at a certain time may exist and be used, but not introspected at another time. The unattended PCCs are no more physical events and processes than the ones that are attended and become instances of A-C, since both are information structures in virtual machines (Sloman, 2002, 2009b,c,d).

Explaining (as outlined later) why certain sorts of physical machinery can support such things, and in certain configurations necessarily will produce them, solves the alleged "hard problem" of consciousness and also helps to explain one of the reasons why the attempt to identify a concept of P-C is misguided, since instances of P-C will unnecessarily duplicate instances of A-C (or PCC) that occur in virtual machines of the right kind. The hypothesis that P-C instances exist explains nothing that is not better explained in a different way, including explaining how some PCCs can become instances of A-C. So P-C suffers from conceptual disengagement defined above in section 2.3. Moreover the mode of identification of instances of the concept of P-C has the characteristics of a kind of self-delusion often found in "ostensive" definitions, as illustrated in "the hard problem of spatial identity".

4. Conclusion

I have tried to summarise and justify critical comments I have been making over several decades about research on consciousness by philosophers, scientists and others. Explaining the background to the criticisms required explaining a variety of types of semantic flaw that can occur in various kinds of discussion including serious flaws associated with the concept of "phenomenal consciousness". Other features that are not inherently flaws in concepts, such as polymorphism and context sensitivity can, if not noticed or if misunderstood lead to flawed research goals and flawed descriptions of research achievements, some of them reminiscent of the faults lambasted in (McDermott, 1981).

I have tried to show (a) that concepts like "phenomenal consciousness" (P-C) as defined by Block, are semantically flawed and unsuitable as a target for scientific research or machine modelling, whereas something like the concept of "access consciousness" (A-C) with which it is contrasted refers (polymorphically) to phenomena that can be described and explained within a future scientific theory, and (b) that the "hard problem" is a bogus problem, because of its dependence on the P-C concept. It was compared with a more obviously bogus problem, the "the 'hard' problem of spatial identity" introduced as part of a tutorial on "semantic flaws". Different types of semantic flaw and conceptual confusion not normally studied outside analytical philosophy are distinguished. The semantic flaws of the "zombie" argument, closely allied with the flaws of the P-C concept are also explained. These topics are related both to the evolution of human and animal minds and brains and to requirements for human-like robots. The diversity of the phenomena related to the concept "consciousness" make it a *second-order polymorphic* concept, partly analogous to concepts like "efficiency" and others. As a result there cannot be one explanation of consciousness, one set of neural associates of consciousness, one explanation for the evolution of consciousness, one point in the development of a foetus at which consciousness is produced, nor one machine model of consciousness.

(Jack & Shallice, 2001) write: *"The closest that science can come to accounting for subjectivity is through elucidating the mechanisms that allow us to understand ourselves from our own point of view. Thus, our second step is to argue that a theory of consciousness must account for the processes underlying introspection."* That is part of what I am trying to do. But I also claim that when we have filled in the details we shall see what it is about those details that can produce introspective states of the very kind that produce philosophical theories about qualia, P-C, the "explanatory gap" and the "hard problem". This is a mixture of good and bad: the good is the (admittedly superficial) description of what is going on in a human-like virtual machine, e.g. during certain kinds of perception. The bad is failing to see that contents of virtual machines playing a vital role in the mechanisms of mind are amenable to scientific investigation and replication on machines of the future.

(McCarthy, 1995) writes *"Thinking about consciousness with a view to designing it provides a new approach to some of the problems of consciousness studied by philosophers. One advantage is that it focuses on the aspects of consciousness important for intelligent behavior. If the advocates of qualia are right, it looks like robots won't need them to exhibit any behavior exhibited by humans."* I argue that they do need the contents of consciousness that have been used as examples of qualia, as do dogs, chickens, fish and birds, but they need not know that they need them, that they are using them, how they are using them, and so on. And they certainly don't need the ghostly, causally disconnected, P-C surrogates.

5. Acknowledgements

I am grateful to Antonio Chella for inviting me to talk to the AAAI Fall Symposium 2007 workshop he chaired (Sloman, 2007) and for inviting me to submit a very slightly revised version of that workshop paper with the title "An Alternative to Working on Machine Consciousness" as a target for commentary in IJMC. I am also grateful to readers who took the trouble to write commentaries which convinced me of the need to write this background paper, namely Alexei V. Samsonovich, Cristiano Castelfranchi, Colin Hales, Catherine Legg, Drew McDermott, Elizabeth Irvine, Giuseppe Trautteur, Igor A Aleksander, John G. Taylor, Piotr Boltuc, Roberto Cordeschi and Ricardo Sanz. Catherine Legg in particular read and usefully commented on a draft of part of this paper. Luc Beaudoin made very useful comments on the penultimate version of this paper.

References

Austin, J. L. [1959] *Sense and Sensibilia* (OUP, Oxford).

Baars, B. J. [1988] *A cognitive Theory of Consciousness* (Cambridge University Press, Cambridge, UK).

Baars, B. J. [1997] *In the Theather of Consciousness: The Workspace of the Mind* (Oxford University Press, New York, Oxford).

Beaudoin, L. [1994] *Goal processing in autonomous agents*, PhD thesis, School of Computer Science, The University of Birmingham, Birmingham, UK, http://www.cs.bham.ac.uk/research/projects/cogaff/81-95.html#38.

Berthoz, A. [2000] *The Brain's sense of movement* (Harvard University Press, London, UK).

Block, N. [1995] On a confusion about the function of consciousness, *Behavioral and Brain Sciences* **18**, 227–47.

Block, N. [1999] Ridiculing social constructivism about phenomenal consciousness, *Behavioral and Brain Sciences* **22**(01), 199–201, doi:10.1017/S0140525X99221802.

Chappell, J. and Sloman, A. [2007] Natural and artificial meta-configured altricial information-processing systems, *International Journal of Unconventional Computing* **3**(3), 211–239, http://www.cs.bham.ac.uk/research/projects/cosy/papers/#tr0609.

Craik, K. [1943] *The Nature of Explanation* (Cambridge University Press, London, New York).

Degenaar, J. and Keijzer, F. [2009] Workspace and Sensorimotor Theories: Complementary Approaches to Experience, *Journal of Consciousness Studies* **16**(9), 77–102.

Gibson, J. [1966] *The Senses Considered as Perceptual Systems* (Houghton Mifflin, Boston).

Gibson, J. J. [1979] *The Ecological Approach to Visual Perception* (Houghton Mifflin, Boston, MA).

Goodale, M. and Milner, A. [1992] Separate visual pathways for perception and action, *Trends in Neurosciences* **15**(1), 20–25.

Jack, A. I. and Shallice, T. [2001] Introspective Physicalism as an Approach to the Science of Consciousness, *Cognition* **79**(1–2), 161–196, http://nivea.psycho.univ-paris5.fr/ philipona/biblio/Author/SHALLICE-T.html.

Johnson-Laird, P. [1988] *The Computer and the Mind: An Introduction to Cognitive Science* (Fontana Press, London), (Second edn. 1993).

Kennedy, C. M. [1999] "Distributed reflective architectures for adjustable autonomy," in *International Joint Conference on Artificial Intelligence (IJCAI99), Workshop on Adjustable Autonomy* (IJCAI, Stockholm, Sweden).

Legg, C. [2004] Review of *Machine Consciousness* by Owen Holland (Editor) Imprint Academic, 2003, *Metapsychology Online Reviews* **8**(36), http://metapsychology.mentalhelp.net/poc/view_doc.php?type=book&id=2305.

McCarthy, J. [1995] "Making robots conscious of their mental states," in *AAAI Spring Symposium on Representing Mental States and Mechanisms* (AAAI, Palo Alto, CA), revised version: http://www-formal.stanford.edu/jmc/consciousness.html.

McCarthy, J. [2008] The well-designed child, *Artificial Intelligence* **172**(18), 2003–2014, http://www-formal.stanford.edu/jmc/child.html.

McDermott, D. [1981] Artificial intelligence meets natural stupidity, in J. Haugeland (ed.), *Mind Design* (MIT Press, Cambridge, MA).

McDermott, D. [2007] Artificial Intelligence and Consciousness, in P. D. Zelazo, M. Moscovitch & E. Thompson (eds.), *The Cambridge Handbook of Consciousness* (Cambridge University Press, Cambridge), pp. 117–150, http://www.cs.yale.edu/homes/dvm/papers/conscioushb.pdf.

Minsky, M. L. [1987] *The Society of Mind* (William Heinemann Ltd., London).

Minsky, M. L. [2006] *The Emotion Machine* (Pantheon, New York).

M.Minsky [2005] "Interior Grounding, Reflection, and Self-Consciousness," in *Brain, Mind and Society, Proceedings of an International Conference on Brain, Mind and Society* (Graduate School of Information Sciences, Brain, Mind and Society, Tohoku University,

Japan), http://web.media.mit.edu/ minsky/papers/Internal Grounding.html.

O'Regan, J. and Noë, A. [2001] A sensorimotor account of vision and visual consciousness, *Behavioral and Brain Sciences* **24**, 939–1031.

O'Regan, J., Rensink, R. and Clark, J. [1999] Change-blindness as a result of 'mudsplashes', *Nature* **398(6722)**, 34.

Ryle, G. [1949] *The Concept of Mind* (Hutchinson, London).

Shanahan, M. [2006] A cognitive architecture that combines internal simulation with a global workspace, *Consciousness and Cognition* **15**, 157–176.

Sloman, A. [1969] How to derive "better" from "is", *American Phil. Quarterly* **6**, 43–52, http://www.cs.bham.ac.uk/research/cogaff/sloman.better.html.

Sloman, A. [1970] "Ought" and "better", *Mind* **LXXIX**(315), 385–394, http://www.jstor.org/view/00264423/di984453/98p0251x/0.

Sloman, A. [1978] *The Computer Revolution in Philosophy* (Harvester Press (and Humanities Press), Hassocks, Sussex), http://www.cs.bham.ac.uk/research/cogaff/crp.

Sloman, A. [1982] Image interpretation: The way ahead? in O. Braddick & A. Sleigh. (eds.), *Physical and Biological Processing of Images (Proceedings of an international symposium organised by The Rank Prize Funds, London, 1982.)* (Springer-Verlag, Berlin), pp. 380–401, http://www.cs.bham.ac.uk/research/projects/cogaff/06.html#0604.

Sloman, A. [1989] On designing a visual system (towards a gibsonian computational model of vision), *Journal of Experimental and Theoretical AI* **1**(4), 289–337, http://www.cs.bham.ac.uk/research/projects/cogaff/81-95.html#7.

Sloman, A. [1999] What sort of architecture is required for a human-like agent? in M. Wooldridge & A. Rao (eds.), *Foundations of Rational Agency* (Kluwer Academic, Dordrecht), pp. 35–52, http://www.cs.bham.ac.uk/research/projects/cogaff/96-99.html#21.

Sloman, A. [2002] The irrelevance of Turing machines to AI, in M. Scheutz (ed.), *Computationalism: New Directions* (MIT Press, Cambridge, MA), pp. 87–127, http://www.cs.bham.ac.uk/research/cogaff/00-02.html#77.

Sloman, A. [2007] "Why Some Machines May Need Qualia and How They Can Have Them: Including a Demanding New Turing Test for Robot Philosophers," in A. Chella & R. Manzotti (eds.), *AI and Consciousness: Theoretical Foundations and Current Approaches AAAI Fall Symposium 2007, Technical Report FS-07-01* (AAAI Press, Menlo Park, CA), pp. 9–16, http://www.cs.bham.ac.uk/research/projects/cosy/papers/#tr0705.

Sloman, A. [2008a] "Architectural and representational requirements for seeing processes, proto-affordances and affordances," in A. G. Cohn, D. C. Hogg, R. Möller & B. Neumann (eds.), *Logic and Probability for Scene Interpretation, Dagstuhl Seminar Proceedings*, 08091 (Schloss Dagstuhl - Leibniz-Zentrum fuer Informatik, Germany, Dagstuhl, Germany), `http://drops.dagstuhl.de/opus/volltexte/2008/1656`.

Sloman, A. [2008b] "The Self" – A bogus concept, Http://www.cs.bham.ac.uk/research/projects/cogaff/misc/the-self.html.

Sloman, A. [2009a] Some Requirements for Human-like Robots: Why the recent over-emphasis on embodiment has held up progress, in B. Sendhoff, E. Koerner, O. Sporns, H. Ritter & K. Doya (eds.), *Creating Brain-like Intelligence* (Springer-Verlag, Berlin), pp. 248–277, http://www.cs.bham.ac.uk/research/projects/cosy/papers/#tr0804.

Sloman, A. [2009b] Virtual Machines and the Metaphysics of Science, Http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#mos09, PDF presentation for AHRC Metaphysics of Science Conference.

Sloman, A. [2009c] "What Cognitive Scientists Need to Know about Virtual Machines," in N. A. Taatgen & H. van Rijn (eds.), *Proceedings of the 31st Annual Conference of the Cognitive Science Society* (Cognitive Science Society, Austin, TX), pp. 1210–1215, http://www.cs.bham.ac.uk/research/projects/cogaff/09.html#901.

Sloman, A. [2009d] Why the "hard" problem of consciousness is easy and the "easy" problem hard. (And how to make progress), Online tutorial presentation: http://www.cs.bham.ac.uk/research/projects/cogaff/talks/#cons09.

Sloman, A. [2010] An Alternative to Working on Machine Consciousness, *Int. J. Of Machine Consciousness* http://www.cs.bham.ac.uk/research/projects/cogaff/09.html#910.

Sloman, A. [(to appear)] What's information, for an organism or intelligent machine? How can a machine or organism mean? in G. Dodig-Crnkovic & M. Burgin (eds.), *Information and Computation* (World Scientific, New Jersey), http://www.cs.bham.ac.uk/research/projects/cogaff/09.html#905.

Sloman, A. and Chrisley, R. [2003] Virtual machines and consciousness, *Journal of Consciousness Studies* **10**(4-5), 113–172, http://www.cs.bham.ac.uk/research/projects/cogaff/03.html#200302.

Sloman, A. and Logan, B. [1999] Building cognitively rich agents using the Sim_agent toolkit, *Communications of the Association for Computing Machinery* **42**(3), 71–77, http://www.cs.bham.ac.uk/research/projects/cogaff/96-99.html#49.

Strawson, P. F. [1959] *Individuals: An essay in descriptive metaphysics* (Methuen, London).

Trehub, A. [1991] *The Cognitive Brain* (MIT Press, Cambridge, MA), http://www.people.umass.edu/trehub/.

Turing, A. [1950] Computing machinery and intelligence, *Mind* **59**, 433–460, (reprinted in E.A. Feigenbaum and J. Feldman (eds) *Computers and Thought* McGraw-Hill, New York, 1963, 11–35).

Velmans, M. [2009] How to define consciousness - and how not to define consciousness, *Journal of Consciousness Studies* **16**(5), 139–156.

Wittgenstein, L. [1953] *Philosophical Investigations* (Blackwell, Oxford), (2nd edition 1958).