Can Machines Have First-Person Properties?

Mark F. Sharlow

1. Introduction

One of the most important ongoing debates in the philosophy of mind is the debate over the reality of the first-person character of consciousness.[1] Philosophers on one side of this debate hold that some features of experience are accessible only from a first-person standpoint. Some members of this camp, notably Frank Jackson, have maintained that epiphenomenal properties play roles in consciousness [2]; others, notably John R. Searle, have rejected dualism and regarded mental phenomena as entirely biological.[3] In the opposite camp are philosophers who hold that all mental capacities are in some sense computational - or, more broadly, explainable in terms of features of information processing systems.[4] Consistent with this explanatory agenda, members of this camp normally deny that any aspect of mind is accessible solely from a first-person standpoint. This denial sometimes goes very far - even as far as Dennett's claim that the phenomenology of conscious experience does not really exist.[5]

The reason for the deep divide between these two camps is not hard to recognize. It seems intuitively obvious that a physical system cannot develop any first-person properties by virtue of its information processing capacities alone. An information processing system is, after all, a whole composed of physical parts; the properties of such a whole should be as objective, and as third-person accessible, as the physical parts which make up the system. This applies to the system's logical properties (including its computational powers) as well as to the system's physical characteristics. The intuition that information processing cannot give a system objectively inaccessible properties lies close to the root of the debate over first-person character. If this intuition were not so compelling, the advocates of first-person properties might not have to deny computational views of mind to make room for first-person properties, while members of the opposite camp would not have to debunk first-person properties to carry out their explanatory agenda.

In this paper I will undertake a seemingly unreasonable task: I will argue that the intuition mentioned in the last paragraph is wrong. More precisely, I will argue that a system which implements information processing of a certain specific kind has properties which are both inaccessible to external observers and accessible, in a certain sense, to the system itself. My argument will consist of three parts. First, I will point out a property which some information processing systems possess, and will use some results of mathematical logic to argue that this property is inaccessible to external observers of those systems. Second, I will point out some facts from neurophysiology which appear to imply that human brains have this same property. Third, I will argue that certain things humans say about their experiences can plausibly be read as narratives about inaccessible properties of this sort. This last conclusion suggests that the properties in question might indeed be first-person accessible, third-person inaccessible properties of neural systems.

2. Logical Properties: A Brief Tour

Before beginning, I will review some basic facts about the properties of information processing systems (hereafter called simply "systems"). This review will allow me to highlight certain facts and ideas which will become important later on. In addition to reviewing known facts, I will introduce a definition which will prove useful later. I want to look at the properties of systems from a slightly unusual viewpoint: I will emphasize the distinction between a system's *physical* properties and its *logical* properties, and will explore in detail some of the logical properties which systems may have.

Any system has a number of physical properties, including its size, mass, physical structure, and electrical characteristics. In addition to these physical properties, such a system also has a number of logical properties. One can find many examples of such properties in the realm of simple electronic devices. Examples of logical properties of such devices are:

(i) the property of being an *or-gate* (a device whose output is the logical disjunction of its inputs);

(ii) the property of being an *adder* (a device whose output is the arithmetical sum of its inputs);

(iii) the property of having three inputs and two outputs;

(iv) the property of containing one positive feedback loop.

A system has logical properties by virtue of the way in which it processes information or signals, and hence, indirectly, by virtue of the way that its parts are interconnected.

To exhibit the behaviors of which it is physically capable, a system must have certain logical properties. A desktop computer, for example, cannot do the tasks expected of it unless a certain set of logical relationships obtains among the microscopic parts of its CPU. A human brain cannot orchestrate human behavior unless certain logical relationships hold between different parts of that brain. As an instance of this latter example, a brain cannot exercise its typical capacities for motor coordination unless it implements a certain sort of feedback involving the voluntary muscles and the visual and proprioceptive sensors. Sometimes a logical property of a system is *necessary* for a behavior of the system, in the sense that the system could not exhibit the behavior if it ceased to have the property. For example, to orchestrate the catching of a baseball, a brain must exemplify the kind of neural feedback which I just described. If the feedback loop is severed (say, by blindness) and is not replaced by some substitute feedback loop, then the brain cannot orchestrate the catching of a baseball.

Given a system and a set of behaviors which we take to be expected or typical for that system, we may distinguish properties which are necessary for the production of the expected behaviors from those which are not necessary. If a logical property P of the system is necessary for the production of one or more of the system's expected behaviors, then we may call P a *behaviorally necessary logical property*, or BNLP, for the system. (The decision as to what constitutes a BNLP depends on our choice of a set of expected behaviors for the system. Usually this set will be obvious from the context - for a pocket calculator, the summing of numerical inputs is expected, and so is the clearing of its display; for a thermostat, the regulation of a temperature is expected.) If a property P is a BNLP for a system, then the exemplification of P by that system is a necessary condition (though not generally a sufficient condition) for the production of the expected behaviors of that system.

It is clear that logical properties play roles in the production of behavior. For example, the brain's property of implementing certain feedback loops is relevant to the brain's production of ball-catching behaviors. Whether logical properties can be said to *cause* behaviors is a question which I will not take up here. It is sufficient to recognize that logical properties are involved in the causation of behavior and are relevant to the production of behavior. (Every computer programmer knows that a system's instantiating a program gives the system powers to behave in new ways.)

3. Structurally Complex Logical Properties

Some logical properties are fairly simple in their internal structure. For example, some systems have the property of possessing a single negative feedback loop; this property does not appear terribly complex and can be defined in a relatively small number of words. Other logical properties have more complex logical structures; they are built up from simpler logical properties, often in intricate ways. These complex logical properties include the following:

> (1) the property of containing a positive feedback loop whose reference signal is the output of a negative feedback loop;
>
> (2) the property of containing an and-gate whose inputs are the outputs of or-gates;
>
> (3) the property of being a system whose output is the sum of five of its inputs minus the sum of two of its inputs;
>
> (4) the property of being a system with a prime number of inputs and an odd number of outputs;
>
> (5) the property of being an integrator (that is, a system whose output signal is approximately equal to the integral of its input signal over some period of time);
>
> (6) the property of being a system composed of n subsystems, S_1 through S_n, which are connected so that some of the output signals of S_i are input signals of S_{i+1} for each i = 1, ..., n-1.

All of these properties are perfectly good logical properties; they can be, and in some cases are, instantiated in real electronic or neural systems. Property (3) is the sort of property which might belong to a neuron; some neurons effectively sum their excitatory inputs and subtract their inhibitory inputs from the total. Property (5) is a real property of some electronic devices; there really are circuits which are integrators. Property (6) could belong to a string of processors connected in series. If each processor in this series must be active for the whole system to produce its customary output, then property (6) is a BNLP for the system (the string of processors) as a whole. The only important general difference between properties like (1)-(6) above and simpler logical properties like (i)-(iv) is a difference in structural complexity.

We note in passing that definition (6) actually defines a family of logical properties, one for each natural number n, rather than a single property. Note also that in framing definitions (4) and (5), we used some notions from number theory and from calculus. This illustrates the fact that one can use mathematical notions (and not only those of formal logic) to define logical properties of systems.

Examples (1)-(6) also can be used to illustrate the fact that a logical property can be defined in more than one equivalent way. For instance, one could replace definition (3) with:

> (3a) the property of being a system whose output is the sum of a set of inputs whose number of elements is the third prime number, minus the sum of a set of inputs whose number of elements is the first even positive number.

Of course, we achieve no useful purpose by complicating definition (3) in this way. I offer this alternative definition only to make a point: that there usually is more than one mathematically equivalent way to describe a logical property of a system. Sometimes these equivalent descriptions may look strikingly different; one description may involve mathematical concepts that another, equivalent description does not involve. Definition (3a) also underscores the fact that logical properties with mathematically intricate definitions are legitimate logical properties - as legitimate as properties defined by the simplest definitions.

Another lesson to be learned from (3) and (3a) is this: if logical properties are defined by logically equivalent definitions, then those properties play exactly the same roles in the production of behavior. Suppose that a particular system's behavior depends crucially on the presence of a part which has property (3). Then that system's behavior also depends crucially on the presence of a part which has property (3a) - for any part that has (3) has (3a), and vice versa. It does not make sense to claim that the system's behavior depends on the property defined by (3), but not on the property defined by (3a). If we say this, then we are saying, in effect, that the behavior depends on the system's summing up *five* inputs, but not on its summing up *the third prime number* of inputs - or else that the behavior depends upon subtracting *two* inputs, but not upon subtracting *the first even positive number* of inputs. Out of respect for the truths of arithmetic, we should not say such things. In general, if two definitions of logical properties are logically equivalent, then the properties defined by those definitions play exactly the same roles in the production of behavior. (I will not take up the metaphysical and linguistic question of whether such equivalent definitions describe identical properties or only necessarily coextensive ones.)

By an analogous argument, if a system's having a property P entails the system's having a property Q, then if P is a BNLP, then so is Q. If P is a BNLP for S, then the instantiation of P by S is necessary for the production of S's expected behaviors; since Q is instantiated whenever P is, the instantiation of Q by S also is necessary for this, so Q is a BNLP also. To deny this conclusion is to claim that a system may produce its expected behaviors while possessing P but not Q - a logical impossibility, given that "S has P" entails "S has Q".

4. Some Complex Logical Properties of Brains

In this section, we will look more closely at the family of properties specified by definition (6) above. We will define other logical properties in terms of these properties; some of the resulting definitions will be much more intricate than any definitions given so far. At first, these definitions may seem pointless; their importance to the project of the paper should become clear in the next section.

First, let us give names to the properties defined in (6). For each natural number n, we will say that a system S has *monitor depth* n if and only if S has n subsystems, S_1 through S_n, which are connected so that some of the output signals of S_i are input signals of S_{i+1} for each i = 1, ..., n-1. In applying this definition, we will count the system S itself as a subsystem (this could happen in the case n=1). Also, we will not require that the S_i form a partition of the system in any sense; we will not require that all the S_i in the definition be distinct (this is to allow for the possibility of feedback loops).

Despite the abstractness of this definition, the property of having monitor depth n has a simple intuitive meaning. If the input of a subsystem B in a given system is the output of another subsystem A, then B can be thought of as "monitoring" (in a *very* broad sense of that word) the output of A. A system's monitor depth is simply a measure of the number of these "monitoring" steps through which the input of the system may pass. Note that it is possible (indeed common) for a system to have more than one monitor depth.

For any natural number n, we will call the property of having monitor depth n a "monitor depth property," or simply a "monitor depth." Thus, a system which has monitor depths 2 and 3 can be said to possess two monitor depths.

A monitor depth can be a BNLP of a system. For example, monitor depth 1 is a BNLP for all systems; without this property, a system would not be able to process its input in one or more consecutive stages, and no processing would get done. For a system such as a human brain, where inputs often are processed through many subsystems, the property of having monitor depth n can be a BNLP for rather high values of n. As an example, consider the visual system.[6] This system receives its inputs from the optic nerves; most of these inputs first reach the lateral geniculate nucleus (LGN), whose outputs go to the visual cortex. The outputs of the visual

cortex go to other areas of the cortex. Because it contains the visual pathway, a human brain has a monitor depth of at least 4. If any of the connections in this pathway were to disappear (without being replaced by some other, substitute connections), then the brain could not experience vision, and hence could not produce the full range of behavior typical of most human brains. Thus, monitor depth 4 is a BNLP of human brains. So, for that matter, are monitor depths 1, 2 and 3. (Actually the visual system is far more complicated than my brief description suggests, and its monitor depths go far above 4. I will raise this point again later.)

It is important to note that all of the monitor depths which a human brain possesses are *equally necessary* for the production of the expected behaviors of the brain. If any of them is not exemplified, then the brain cannot control behavior in the same ways as it customarily does, and to the same extent that it usually does. Intuitively, one might be tempted to regard monitor depth 1 as more fundamental or "more necessary" than the other monitor depths. This intuition is correct only in the sense that the loss of monitor depth 1 might have more profound consequences for brain function than the loss of a higher monitor depth alone. (To make a brain lose the monitor depth 1, one would have to dismantle it rather thoroughly.) But all of the monitor depths are equally necessary for the brain's production of the full spectrum of typical behaviors.

Now let us extend the definition of monitor depth slightly, by allowing monitor depths to be indexed by *sets* of natural numbers (positive whole numbers) instead of only by single natural numbers. We do this in the following manner. If A is a set of natural numbers, we say that a system S has *monitor depth over* A if and only if for each number n in A, S has monitor depth n. (For example, if a particular brain has monitor depths 1, 2, 3 and 4, then we will say that that brain has monitor depth over the set {1, 2, 3, 4}.) In a similar vein, if P is a property of natural numbers, we will say that S has *monitor depth over* P if and only if for each natural number n having P, S has monitor depth n. (For example, if a particular brain has monitor depths 1, 2, 3 and 4, then we will say that that brain has monitor depth over the property of being a natural number less than 5.)

5. Empirically Inaccessible Logical Properties?

In this section, we will focus on two particular logical properties of systems. These properties are defined in terms of monitor depth. At least one of these properties is widespread among machines; its relevance to the topic of consciousness will not become evident until later in the paper. This section makes use of some concepts from mathematical logic; readers unfamiliar with these concepts may want to refer to the relevant literature.[7]

An advance warning to the reader: I will use Gödel's second incompleteness theorem in this section. However, my argument should *not* be confused with Penrose's well-known arguments which deploy that theorem in the philosophy of mind.[8] What I am doing in this paper has very little connection with Penrose's views and should be judged separately from them. (My conclusions do not require one to accept or to reject Penrose's views.)

We define two logical properties as follows. We begin by choosing a formalized language L, and a formal theory T in the language L. Our choices of L and T must be such that T includes first-order arithmetic (that is, some first-order formulation of the Peano axioms, along with first-order logic). T is a well-defined mathematical object; in principle, one can define logical properties of systems in terms of T, just as we have defined such properties in terms of simpler mathematical objects and operations like prime numbers, integration, and addition. (Recall my earlier remarks about complex logical properties and the legitimacy of using mathematical concepts in defining such properties.)

Next, select a procedure for Gödel numbering of formulas of L. Let G be a function mapping formulas of L into their Gödel numbers according to this procedure. Then the pair (T, G), like T alone, is a legitimate mathematical object. In principle, one can define logical properties of systems in terms of this object. Next, use (T, G) to define two properties of natural numbers as follows. We call these properties P and Q:

Definition of P. Let n be a natural number. P(n) if and only if: n is the lowest Gödel number assigned by G to any formula of L undecidable in T, and n is the Gödel number under G of a true formula.

Definition of Q. Let n be a natural number. Q(n) if and only if: n is the lowest Gödel number assigned by G to any formula of L undecidable in T, and n is the Gödel number under G of a false formula.

These definitions are far more complex than any of the property definitions in the last section. However, P and Q still are legitimate mathematical properties, defined in terms of well-understood mathematical objects.

Now consider the following logical properties, A and B, of systems:

Definition of A. Let S be a system. A(S) if and only if S has monitor depth over P.

Definition of B. Let S be a system. B(S) if and only if S has monitor depth over Q.

Because P and Q are well-defined mathematical properties of natural numbers, it follows that A and B are well-defined logical properties of systems.

Do any actual systems have A and/or B? To answer this question we must ask another question: Which natural numbers have the properties P and Q? This question is easily answerable. For any natural number n, if n is not the lowest Gödel number under G for a formula undecidable in T, then both P(n) and Q(n) are false. Hence each of P and Q is true for at most one natural number. According to Gödel's second incompleteness theorem, there is a natural number n such that n codes a formula undecidable in T. Hence there is a lowest such natural number, say m. The formula F coded by m is either true or false. Hence either P(m) is true or Q(m) is true, but not both. Thus, one member of the pair of P and Q is true of exactly one natural number (namely m), while the other member of the pair is false of all natural numbers. This follows by formal logic from the definitions of P and Q.

Now we can provide a partial answer to the original question: which systems have A and/or B? If a system has monitor depth m, then it will have either A or B, depending on whether m has P or has Q. If P(m), then the system has A. If Q(m), then the system has B. A system without monitor depth m has neither A nor B.

So far, we have shown absolutely nothing important about systems. The statements in the last paragraph are simple consequences of definitions and well-known mathematical theorems; they tell us nothing new about the physical world. However, these statements do have one potentially interesting consequence. According to Gödel's second incompleteness theorem, we can never tell, by means of proofs formulated in T, whether F is true or false. This directly implies that we cannot prove, via arithmetical methods formalizable within T, whether it is P(m) or Q(m) that is true. Now let S be a system which we know to have monitor depth m. We know that A(S) if and only if P(m), and that B(S) if and only if Q(m). Therefore, we cannot prove, using methods formalizable in T, that A(S). Likewise, we cannot prove via such methods that B(S). However, we can prove that the disjunction, A(S) or B(S), is true.

It is not hard to determine whether a system S has monitor depth m. We can determine that from observations of the structure and behavior of S. However, even if we determine that S has monitor depth m, there is a serious obstacle to our determining whether S has A or has B. This obstacle is of a logical nature. No matter how much observational data we have about S, we cannot determine whether S has A or S has B by using only those mathematical methods available within the theory T. Suppose, by way of *reductio*, that we could make observations which would allow us to infer whether it is A(S) or B(S) that is true. Once we know which of A(S) and B(S) is true, then we can determine whether F is true or false. Thus, any observational data that would let us decide whether S has A or has B also would let us decide whether F is true or false. But F is an undecidable formula of T; we cannot find out whether F is true or false via mathematical methods available in T - and no amount of information about how a particular physical system happens to be hooked together will help us to find this out! One cannot override Gödel's theorem by deploying information about a particular physical system; the

undecidability which that theorem imposes on F cannot be changed by the truth or falsehood of a bunch of contingent facts about the physical world.

Empirical studies of S can tell us exactly which monitor depths S has. But once we have this information, we have done all we can toward determining whether S has A or has B. Empirical facts about S cannot tell us whether F is true or false; hence, if S has monitor depth m, such facts cannot tell us whether S has A or S has B.

At this point, we have not done anything very interesting philosophically. We have simply shown that a system S with monitor depth m has a logical property which is not empirically detectable; this undetectability is a consequence of an incompleteness theorem. (This finding is no stranger than the fact that some numbers have mathematical properties which are not mathematically demonstrable, again because of an incompleteness theorem.) We can make our result more philosophically interesting by carefully choosing the language L and the theory T, as described in the following paragraph.

First, note that versions of the second incompleteness theorem hold, not only for arithmetic, but also for a wide range of mathematical theories which extend arithmetic, and also for ZFC set theory [9], in which all of classical mathematics (including arithmetic) is known to be representable. Let us try choosing T to be the theory generated by all mathematical theories used today in science - including arithmetic, basic algebra, linear algebra, group theory, calculus, differential equations, etc. For L, we must choose a formal language sufficiently powerful to express this theory. (Such a language might include some version of predicate logic with identity, together with notations for real number algebra, group theory, calculus, etc.; there is no reason in principle why the alphabet, syntactical rules, and semantical rules for such a language could not be written down.) We know that the question of whether S has A or has B cannot be decided by any means available within T. But for this particular choice of T, the means available within T include all mathematical methods presently used in science. Hence, if we choose L and T in this way, we find that the question of whether S has A or has B cannot be decided *by any mathematical means now used in science*. As I pointed out earlier, more empirical data will not help us to make this decision. Furthermore, it is not only the mathematical methods used in science today that cannot decide some question like this. Even if science is going to use other mathematical methods in the future, we can simply adjust L and T to encompass those methods, and then obtain a similar result for those methods.

The upshot is that if a system has monitor depth m, then *that system has a logical property which cannot be detected by any sort of empirical observations or scientific study.*

Some scientifically minded readers may be tempted to stop reading at this point. I urge those who feel this way to continue reading, as I am not going to advocate anything antiscientific in this paper. The result I just obtained may sound provocative and odd on a casual reading, but on closer inspection it turns out to be rather trivial. It does not tell us that there is anything special about the system S; the "undetectability" of the properties A and B is not a result of anything strange about the system, but is simply a corollary of a well-known mathematical result - Gödel's second incompleteness theorem. If anything is remarkable here, it is the incompleteness theorem, not my results or the system S. There is nothing metaphysically deep in my results so far.

6. Heterophenomenology and First-Person Access

Next, let us see what happens when we apply the conclusion of the last section to the most interesting information processing system of all - the human brain.

Human brains contain a wide variety of feedback loops. The visual system, which I discussed earlier, contains a number of these loops [10], and there are many other feedback loops in brains besides these visual loops. Earlier I alluded to the possibility of defining monitor depths for systems containing feedback loops. When one does this, one finds that such systems have very large sets of monitor depths. The following argument shows why.

Suppose that inputs entering the feedback loop pass through a string of t subsystems before reaching the loop. Suppose that t is at least 1 (the argument would be a little different for t=0). Call these subsystems S_i, ..., S_t. Suppose that these subsystems are connected in series, with some of the outputs of one forming inputs of the next. Then these subsystems form a chain of the sort called for by the definition of monitor depth. Hence we can conclude that the system has all of the monitor depths 1, 2, ..., t. The feedback loop itself consists of at least one subsystem, so we can trace the flow of information even further. Consider a feedback loop comprised of v subsystems (where v is at least 1), called U_1, ..., U_n, connected in series so that some of the outputs of each subsystem in the sequence are the inputs of the next. Since this is a feedback loop, some of the outputs of U_n are inputs for U_1. Let U_1 be the system into which the outputs of S_t feed. By going around the feedback loop once, we trace the chain of subsystems S_i, ..., S_t, U_1, ..., U_n. This chain is of the sort used in the definition of monitor depth; hence we conclude that the system has monitor depths 1, 2, ..., t+n. However, we can go around the loop *again*, and examine the chain of subsystems S_i, ..., S_t, U_1, ..., U_n, U_1, ..., U_n. Recall that the subsystems in the definition of monitor depth do not have to be distinct; the chain used in that definition can pass through the same subsystems over and over again. Hence we can show that the system has monitor depths 1, 2, ..., t+2n. We can continue this argument by going around the loop again and again. Finally, we discover that the system has *all* monitor depths 1, 2, 3, ... It has monitor depth n for each natural number n.

A system with a feedback loop has all natural numbers as its monitor depths. From this it follows that such a system has monitor depth m, where m is the number defined in the last section. Hence the system has either property A or property B, as defined in the last section.

Beginning with the facts noted in the last two paragraphs, we can show that A and B are BNLPs for human brains. The argument for this is as follows. The presence of feedback loops in a brain is necessary for that brain to exercise its customary functions. Two cases are possible: either (1) there is a natural number m such that P(m), or (2) there is a natural number m such that Q(m). Suppose that (1) is true. Then a normal brain, by virtue of having a feedback loop, has A. If that brain lacked A, it could not produce the normal behaviors we expect of it, since the lack of A would imply the absence of feedback loops, and without these loops the brain could not do what it normally does. Thus, A is a BNLP for the brain. Now suppose (2) is true; by an argument similar to the one for case (1), B is a BNLP for the brain. Hence either A or B is a BNLP for the brain, depending upon which of the two cases is true.

So far, we have learned nothing truly new about brains. These results are mere corollaries of known mathematical results; they are applicable to systems much simpler than brains - even television sets contain some feedback loops. Hence we have not yet learned anything new about consciousness. We have only learned how to apply Gödel's theorem to machines in amusing (or repulsive?) new ways.

There is one complication which may change this.

Daniel C. Dennett has championed the view that conscious processes can be investigated by means of the method of *heterophenomenology*.[11] In heterophenomenological studies, one examines an organism's reports of its own experiences, and tries to interpret these reports as narratives about objects. The heterophenomenologist may regard the objects discovered in this way as fictional objects. Alternatively, he may identify such an object with a real item or event inside the organism, if that object's properties, as described in the narrative, sufficiently resemble the actual properties of some internal item or event.[12]

Now for a thought experiment. We will apply the heterophenomenological method to the narratives of a human organism whose brain has the property A defined above. (If the subject's brain has B instead, the argument will be similar.)

Let us suppose that a person is reporting on his experience of color. At first, he sits in a red room with his eyes closed. At time t, he opens his eyes and begins to see the color red. Then he is asked to describe what he experienced immediately after opening his eyes. He might say something like:

(I) I started seeing red.

If he has never experienced red before (say, if he was blind from birth and his sight had just been restored), then he might well say something like this:

(II) I started to experience something - I can't really describe it in words.

This narrative reflects the familiar "ineffability" of color [13] - the difficulty which people encounter when trying to say exactly what a color experience feels or seems like. If the subject who uttered (II) is knowledgeable about neuroscience and about the cellular and molecular makeup of human brains, then he might also come up with a statement like this:

(III) The walls seem to have a special property. That property sure isn't like anything a scientist could find inside my brain.

For convenience, let us call the second sentence in (III) by a new name, (IIIa). The subject's utterance of the sentence (IIIa) seems to reflect the common feeling that conscious experiences are private or have first-person character. Certainly it can be read as saying that the property in question is not accessible to external observers. The question of the literal truth or falsehood of (IIIa) is the sort of question over which the parties to the first-person character debate might have a good fight.

The sentence (IIIa) is only a subject's description of how things seem to him. Yet if we consider the arguments of the last section, we are forced to conclude that this sentence is *literally true*. There *really is* a property in the subject's brain which is unlike anything an objective observer could find in the subject's brain! This property is the property A.

The narrative (III) leaves the heterophenomenologist with two choices. Either he reads the narrative as a fiction about a property which does not really exist, or else he reads it as being about something inside the organism. The heterophenomenologist can take the latter course if there is something in the organism (most likely a property within the organism) which the narrative describes fairly accurately.

If the heterophenomenologist takes the former path, he does not regard the subject as actually talking about property A or any other actual property. Instead, he constructs the subject's heterophenomenological world so as to include a fictitious property which is like A in some respects, including the very important respect of being inaccessible, in a sense, to external observers. Even if the heterophenomenologist does not identify this property with A, he is stuck with the fact that in uttering (IIIa), *the subject has spoken truly*. The subject has made a true statement about his inner goings-on. Further, it is not merely accidental that the subject made such a statement; the property A actually played a role in the production of the subject's statement. Normal color vision cannot occur without the action of feedback loops; hence it also cannot occur without a part of the brain instantiating property A. We have seen that for any properties P and Q, if "x has P" strictly entails "x has Q," then if P is a BNLP, then Q must be a BNLP also. The instantiation of A by the subject's brain is a necessary condition for the occurrence of normal vision. The presence of the property A is necessary for the production of the narrative in (III); it is just as necessary for this as the (seemingly) simpler property of having a feedback loop in the visual system. Thus, the subject has made a true statement to the effect that there is a property which a third-person observer couldn't find in his brain - and the fact that he does have such a property in his brain was part of the circumstances which triggered him to make this statement.

The heterophenomenologist's other option is to try to identify the property in the subject's heterophenomenological world with some actual property or item within the subject. The property A is an actual property within the subject; it is inaccessible in the way the subject claims the reported property to be, and it actually played some role in triggering the subject to produce the narrative. These considerations make A a prime candidate for identification with the reported property.

One might object that the property A really is nothing like the property which the subject reports. After all, A is defined in a very intricate way; its definition involved formalized languages, theories, and the like. The subject's actual narrative contains no trace of these mathematical objects; the subject's experience is not an experience of formalized languages and theories. To see why this objection fails, we need only consider the nature of neural

systems. Any informational state of a neural system has a terribly complex fine structure at the neuronal level. If we wanted to describe in microscopic detail the brain properties which give rise to conscious experience, we would, at very least, have to talk about the numbers of inputs to many different neurons, the summing and differencing of excitatory and inhibitory inputs, the connectivity of the neural system (including the structure of networks and feedback loops), and so forth. Yet the subject of experience neither talks about nor notices any of this fine structure. It is a characteristic of conscious experiences that the logical properties of the underlying neural systems are not apparent to the conscious subject. Thus, the fact that the subject reports a property apparently lacking the logical structure of A is no argument against the identification of the reported property with A. If we refused to identify reported properties with neural properties on the grounds that the reported properties lack the detailed structure of the neural properties, then we would not be able to identify any reported properties with realistic neural properties.

Another argument against the identification of the reported property with A is more cogent, but not fatal to the identification. This is the objection that A is not the only property to which the subject's narrative could reasonably be taken to refer. A brain with property A also has many other properties besides A which are both empirically inaccessible and behaviorally necessary to the production of the subject's narrative. One can construct such properties by using undecidable propositions of number theory; there are many such propositions. Consider an undecidable mathematical statement of the form "There exists a natural number n such that R." Then let C be a property of systems defined as follows: S has C if and only if S has a monitor depth over the property given by R. By an argument similar to the one given earlier for A, C is a property inaccessible to an external observer of a system, but a system which can produce narratives about its inner contents may well produce a narrative which can be read as a narrative about C. There may even be other properties of this sort which have nothing to do with monitor depth. Thus, it is reasonable to doubt the identification of the reported property with A; the reported property might equally well be identified with other logical properties of the system.

To meet this objection, the heterophenomenologist must change his tactics a bit. Instead of identifying the reported property specifically with A, he must identify it with some sort of composite of all the externally inaccessible properties involved in the production of the narrative. A simple way is to take the conjunction of all these properties, and identify the resulting complex property with the reported property. (In the sequel, we will call this conjunctive property W.) Once this is done, the heterophenomenologist has identified the reported property with the one property which both fits the narrative description and does not leave any strong suspicion that an entirely different property might be an equally good fit.

The upshot of this argument is that a human brain possesses properties which are inaccessible to third-person observers, but which are accessible, at least in a strained sense of accessibility, to the brain which has those properties. We have found a logical property of brains which appears to have at least a weak sort of first-person character.

Another possible objection to this argument is that it is excessive to identify the reported property with the complicated property W when a simpler property will do. We have seen that the instantiation of A is implied by the existence of a feedback loop. Why suppose that the reported property involves A, when one could suppose that it involves the simpler property of having a feedback loop? All the other logical properties which make up the conjunction W are results of the physical makeup of the brain; hence for each property D which is a conjunct in W, there is another property D' which is a more "physical" property whose instantiation implies that of D. Why not just replace each D in W with its corresponding D', and thereby replace W with a conjunct (call it W') of seemingly simpler properties?

One can reply to this objection in three ways. First, the property W' simply is not a good fit to the subject's description of the reported property. W' is present when the subject sees red - but it does not answer to the part (IIIa) of the subject's description, and hence is a less plausible candidate than W for identification with the reported property. Second, the "simplicity" of W' relative to W does not automatically make W' a better candidate than W for the identification. We have seen that many neural properties relevant to consciousness have terribly complex internal or logical structures. The belief that a property is somehow more relevant to the production of behavior just because it is simpler can be a badly mistaken belief in many cases. Third, we have

seen that if the instantiation of a property is necessary for the production of a behavior, then all the properties whose instantiation this entails also are necessary for the production of that behavior. To regard the conjuncts of W' as necessary for the production of the narrative, and the conjuncts of W as not necessary for this, is unfounded.

7. Concluding Remarks

In this paper I have argued that human brains can have logical properties which are not directly accessible to third-person investigation but nevertheless are accessible (at least in a weak sense) to the brain itself. It is important to remember that these properties are not metaphysically mysterious in any way; they are simply logical properties of neural systems. They are natural properties, arising entirely from the processing of information by various subsystems of the brain. The existence of such properties can pose no threat to the scientific understanding of the mind.

The existence of these logical properties contradicts the widespread feeling that information processing in a machine cannot have features inaccessible to objective observers. But despite this offense against intuition, these findings support a view of first-person access which may be far more congenial to a scientific understanding of the mind than the alternative views that first-person character is either irreducible or unreal. Our conclusion suggests a way to bypass an important obstacle to a reductionistic account of consciousness. Indeed, it suggests that consciousness may be reducible to information processing even if experience does have genuine first-person features.

REFERENCES

Dennett, D.C. (1991). *Consciousness Explained* (Boston: Little Brown & Co.).

Searle, J.R. (1992). *The Rediscovery of the Mind* (Cambridge, MA: The MIT Press).

Jackson, F. (1982). "Epiphenomenal Qualia", *The Philosophical Quarterly* **32**, 127-136.

Penrose, R. (1989). The Emperor's New Mind (N.Y & Oxford: Oxford University Press).

Van Essen, D.C., Anderson, C.H., & Felleman, D.J., (1992). "Information Processing in the Primate Visual System: An Integrated Systems Perspective", *Science* **255**, 419-423.

Goldstern, M., and Judah, H. (1995). *The Incompleteness Phenomenon* (Wellesley, MA: A.K. Peters).

NOTES

1. The adjective "first-person," applied to mental phenomena, has been used perhaps most characteristically by Searle (Searle 1992; see especially pp. 20-21). I will follow Searle's terminology, perhaps with some un-Searlean deviations, throughout this paper.

2. Jackson 1982.

3. Searle 1992, especially pp. 1 and 26. Although Searle regards mental phenomena as biological, he does not accept either materialism or dualism (Searle 1992, Chs. 1 & 2, especially p. 26).

4. Throughout this paper I will use the term "information processing system," or simply "system," to characterize systems (including computers and brains) which engage in the processing of information. I will avoid the more popular terms "computational system" or "computer," since these terms may be narrower in meaning.

5. This claim is developed in detail in Dennett 1991, ch. 11 (see especially p. 365).

6. I draw primarily on Van Essen et al. 1992 for information on the visual system.

7. The information on Gödel theory used in this paper may be found in logic texts dealing with the incompleteness theorems. See, for example, Goldstern & Judah 1995.

8. See especially Penrose 1989, Ch. 4 and pp. 416-418.

9. Goldstern & Judah 1995, p. 233.

10. There is a considerable amount of scientific literature on visual feedback. See, for example, Van Essen et al. 1992.

11. Dennett 1991, Ch. 4.

12. Dennett 1991, p. 85.

13. This ineffability is discussed in Dennett 1991 (especially pp. 49-50 and 382-383), among other places in the literature.