Loosemore, R.P.W. (2012a). Human and Machine Consciousness as a Boundary Effect in the Concept Analysis Mechanism. In: P. Wang & B. Goertzel (Eds), Theoretical Foundations of AGI. Atlantis Press.

Chapter 15

Human and Machine Consciousness as a Boundary Effect in the Concept Analysis Mechanism

Richard Loosemore

Mathematical and Physical Sciences, Wells College, Aurora, NY 13026, U.S.A.

rloosemore@wells.edu

To solve the hard problem of consciousness we observe that any cognitive system of sufficient power must get into difficulty when it tries to analyze consciousness concepts, because the mechanism that does the analysis will "bottom out" in such a way as to make the system declare these concepts to be both real and ineffable. Rather than use this observation to dismiss consciousness as an artifact, we propose a unifying interpretation that allows consciousness to be explicable at a meta level, while at the same time being mysterious and inexplicable on its own terms. This implies that science must concede that there are some aspects of the world that deserve to be called "real", but which are beyond explanation. We conclude that some future thinking machines will, inevitably, have the same subjective consciousness that we do. Some testable predictions are derived from this theory.

15.1 Introduction

The scope of this chapter is defined by the following questions:

- When we use the term "consciousness" what exactly are we trying to talk about?
- How does consciousness relate to the functioning of the human brain?
- If an artificial general intelligence (*AGI*) behaved as if it had consciousness, would we be justified in saying that it was conscious?
- Are any of the above questions answerable in a scientifically objective manner?

The ultimate goal is to answer the third question, about machine consciousness, but in order to make meaningful statements about the consciousness of artificial thinking systems, we need first to settle the question of what consciousness is in a human being. And before

we can answer that question, we need to be clear about whatever it is we are trying to refer to when we use the term "consciousness". Finally, behind all of these questions there is the problem of whether we can explain any of the features of consciousness in an objective way, without stepping outside the domain of consensus-based scientific enquiry and becoming lost in a wilderness of subjective opinion.

To anyone familiar with the enormous literature on the subject of consciousness, this might seem a tall order. But, with due deference to the many intellectual giants who have applied themselves to this issue without delivering a widely accepted solution, I would like to suggest that the problem of consciousness is actually much simpler than it appears on the surface. What makes it seem difficult is the fact that the true answer can only be found by asking a slightly different question than the one usually asked. Instead of asking directly for an explanation of the thing, we need to ask why we have such peculiar difficulty stating what exactly the thing is. Understanding the nature of the *difficulty* reveals so much about the problem that the path to a solution then becomes clear.

15.1.1 *The Hard Problem of Consciousness*

One of the most troublesome aspects of the literature on the problem of consciousness is the widespread confusion about what exactly the word "consciousness" denotes. In his influential book on the subject, Chalmers [2] resolved some of this confusion when he drew attention to the fact that the word is often used for concepts that do not contain any deep philosophical mystery. These straightforward senses include:

- The ability to introspect or report mental states. A fly and a human can both jump out of the way of a looming object, but a human can consciously think and talk about many aspects of the episode, whereas the fly simply does not have enough neural machinery to build internal models of its action. By itself, though, this ability to build internal models is not philosophically interesting.
- Someone who is asleep can be described as not being conscious, but in this case the word is only used for a temporary condition, not a structural incapacity.
- We occasionally say that a person *consciously* did something, when what we really mean is that the person did it *deliberately*.
- If a person *knows* a fact we sometimes say that they are *conscious* of the fact.

In contrast to these senses (and others in a similar vein), there is one meaning for the word "consciousness" that is so enigmatic that it is almost impossible to express. This

is the subjective quality of our experience of the world. For example, the core thing that makes our sensation of redness different from our sensation of greenness, but which we cannot talk about with other people in any kind of objective way. These so-called *qualia*—the quality of our tastes, pains, aches, visual and auditory imagery, feelings of pleasure and sense of self—are all experiences that we can talk about with other people who say they experience them, but which we cannot describe to a creature that does not claim to experience them. When a person who is red-green color blind asks what difference they would see between red and green if they had a full complement of color receptors, the only answer we can give is "It is like the difference between your color red/green and the color blue, only different." To the extent that this answer leaves out something important, that omitted thing is part of the problem of consciousness.

The terms "phenomenology" or "phenomenal consciousness" are also used to describe these core facts about being a conscious creature. This is in contrast to the *psychology* of being a thinking creature: we can analyze the mechanisms of thought, memory, attention, problem solving, object recognition, and so on, but in doing so we still (apparently) say nothing about what it is like to be a thing that engages in cognitive activity.

One way to drive this point home is to notice that it is logically possible to conceive of a creature that is identical to one of us, right down to the last atom, but which does not actually experience this inner life of the mind. Such a creature—a philosophical zombie—would behave as if it did have its own phenomenology (indeed its behavior, *ex hypothesi*, would be absolutely identical to its normal twin) but it would not experience any of the subjective sensations that we experience when we use our minds. It can be argued that if such a thing is logically possible, then we have a duty to explain what it means to say that there is a thing that we possess, or a thing that is an attribute of what we are, that marks the difference between one of us and our zombie twin [1, 5]. If it is conceivable that a thing could be absent, then there must be a "thing" there that can be the subject of questions. That thing—absent in the zombie but present in ourselves—is consciousness.

In order to make a clear distinction between the puzzle of this kind of consciousness, versus the relatively mundane senses of the word listed earlier, Chalmers [2] labeled this the "hard problem" of consciousness. The other questions—for example, about the neural facts that distinguish waking from sleeping—may be interesting in their own right, but they do not involve deep philosophical issues, and should not be confused with the hard problem.

Many philosophers would say that these subjective aspects of consciousness are so far removed from normal science that if anyone proposed an objective, scientific explanation for the hard problem of consciousness they would be missing the point in a quite fundamental way. Such an explanation would have to start with a bridge between the ideas of *objective* and *subjective*, and since the entire scientific enterprise is, almost by definition, about explaining objectively verifiable phenomena, it seems almost incoherent to propose a scientific (i.e. non-subjective) explanation for consciousness (which exists only in virtue of its pure subjectivity).

The story so far is that there is confusion in the literature about the exact definition of consciousness because it is ambiguous between several senses, with only one of the senses presenting a deep philosophical challenge. This ambiguity is only part of the confusion, however, because there are many cases where a piece of research begins by declaring that it will address the hard problem (for example, there is explicit language that refers to the mystery of subjective experience), but then shifts into one of the other senses, without touching the central question at all. This is especially true of neuroscience studies that purport to be about the "neural correlate of consciousness": more often than not the actual content of the study turns out to devolve on the question of which neural signals are present when the subject is awake, or engaging in intentional acts, and so on.

The eventual goal of the present chapter is to answer questions about whether machines can be said to be conscious, so it should be clear that the hard problem, and only the hard problem, is at issue here. Knowing that an artificial intelligence has certain circuits active when it is attending to the world, but inactive when it is not, is of no relevance. Similarly, if we know that wires from a red color-detection module are active, this tells us the cognitive level fact that the machine is detecting red, but it does not tell us if the machine is experiencing a sensation of redness, in anything like the way that we experience redness.

It is this subjective experience of redness—as well as all the other aspects of phenomenology—that we need to resolve. What does it mean to say that a human experiences a subjective phenomenal consciousness, and is it possible to be sure that an artificial intelligence of sufficient completeness would (or would not) have the same phenomenal experience?

15.1.2 *A Problem within the Hard Problem*

We now focus on the fact that even after we separate the hard problem of consciousness from all the non-hard, or easy problems, there is still some embarrassing vagueness in

the definition of the hard problem itself. The trouble is that when we try to say what we mean by the hard problem, we inevitably end up by saying that *something is missing* from other explanations. We do not say "Here is a thing to be explained," we say "We have the feeling that there is something that is not being addressed, in any psychological or physical account of what happens when humans (or machines) are sentient." It seems impossible to articulate what we actually want to see explained—we can only say that we consider all current accounts (as well as every conceivable future account) of the mechanisms of cognition to be not relevant to phenomenology.

The situation can perhaps be summarized in the form of a dialectic:

Skeptic: *If you give us an objective definition for terms such as "consciousness" and "phenomenology," then and only then can we start to build an explanation of those things; but unless someone can say exactly what they mean by these terms, they are not really saying anything positive at all, only complaining about some indefinable thing that ought to be there.*

Phenomenologist: *We understand your need for an objective definition for the thing that we want explained, but unfortunately that thing seems to be intrinsically beyond the reach of objective definition, while at the same time being just as deserving of explanation as anything else in the universe. The difficulty we have in supplying an objective definition should not be taken as grounds for dismissing the problem—rather, this lack of objective definition IS the problem!*

If we step back for a moment and observe this conflict from a distance, we might be tempted to ask a kind of meta-question. Why should the problem of consciousness have this peculiar indefiniteness to it? This new question is not the same as the problem of consciousness itself, because someone could conceivably write down a solution to the problem of consciousness tomorrow, and have it accepted by popular acclamation as *the* solution, and yet we could still turn around and ask: "Yes, but now please explain why the problem was so hard to even *articulate*!" That question—regarding the fact that this problem is different from all other problems because we cannot seem to define it in positive terms—might still be askable, even after the problem itself had been solved.

15.1.3 *An Outline of the Solution*

In fact, this meta-question needs to be addressed first, because it is the key to the mystery. I would like to propose that we can trace this slipperiness back to a specific cause: all intelligent systems must contain certain mechanisms in order to be fully intelligent, and a

side effect of these mechanisms is that some questions (to wit, the exact class of questions that correspond to consciousness) can neither be defined nor properly answered.

When we pose questions to ourselves we engage certain cognitive mechanisms whose job is to analyze the cognitive structures corresponding to concepts. If we take a careful look at what those mechanisms do, we notice that there are some situations in which they drive the philosopher's brain into a paradoxical mixed state in which she declares a certain aspect of the world to be both *real* and *intrinsically inexplicable*. In effect, there are certain concepts that, when analyzed, throw a monkey wrench into the analysis mechanism.

That is a precis of the first phase of the argument. But then there is a second—and in many ways more important—phase of the argument, in which we look at the "reality" of the particular concepts that break the cognitive mechanism responsible for explaining the world. Although phase one of the argument seemed to explain consciousness as a malfunction or short-circuit in the cognitive mechanism that builds explanations, in this second part we make an unusual turn into a new compromise, neither dualist nor physicalist, that resolves the problem of consciousness in a somewhat unorthodox way.

15.2 The Nature of Explanation

All facets of consciousness have one thing in common: they involve some particular types of introspection, because we "look inside" at our subjective experience of the world (qualia, sense of self, and so on) and ask what these experiences amount to. In order to analyze the nature of these introspections we need to take one step back and ask what happens when we think about any concept, not just those that involve subjective experience.

15.2.1 *The Analysis Mechanism*

In any intelligent system—either a biological mind or a sufficiently complete artificial general intelligence (AGI) system—there has to be a powerful mechanism that enables the system to analyze its own concepts. The system has to be able to explicitly think about what it knows, and to deconstruct that knowledge in many ways. If the degree of intelligence is high enough, the scope of this ***analysis mechanism*** (as it will henceforth be called) must be extremely broad. It must be able to ask questions about basic-level concepts, and then ask further questions about the constituent concepts that define basic-level concepts, and then continue asking questions all the way down to the deepest levels of its knowledge.

AGI systems will surely have this analysis mechanism at some point in the future, because it is a crucial part of the "general" in "artificial general intelligence," but since there is currently no consensus about how to do this, we need to come up with a language that allows us to talk about the kind of things that such a mechanism might get up to. For the purposes of this chapter I am going to use a language derived from my own approach to AGI—what I have called elsewhere a "molecular framework" for cognition [6, 7].

It is important to emphasize that there are no critical features of the argument that hinge on the exact details of this molecular framework. In fact, the framework is so general that any other AGI formalism could, in principle, be translated into the MF style. However, the molecular framework is arguably more explicit about what the analysis mechanism does, so by using the language of the framework we get the benefit of a more concrete picture of its workings.

Some AGI formalisms will undoubtedly take a different approach, so to avoid confusion about the role played by the MF in this chapter, I will make the following claim, which has the status of a postulate about the future development of theories of intelligence:

- Postulate (*Analysis Mechanism Equivalence*): Any intelligent system with the ability to ask questions about the meaning of concepts, with the same scope and degree of detail as the average human mind, will necessarily have an equivalent to the analysis mechanism described here.

Different forms of the analysis mechanism will be proposed by different people, but the intended force of the above postulate is that in spite of all those differences, all (or most) of those analysis mechanisms will have the crucial features on which this explanation of consciousness depends. So the use of the molecular framework in this chapter does nothing to compromise the core of the argument.

15.2.2 *The Molecular Framework*

The Molecular Framework (MF) is a generic model of the core processes inside any system that engages in intelligent thought. It is designed to be both a description of human cognition and a way to characterize a broad range of AGI architectures.

The basic units of knowledge, in this framework, are what cognitive psychologists and AGI programmers loosely refer to as "concepts," and these can stand for *things* [chair], *processes* [sitting], *relationships* [on], *actions* [describe], and so on.

The computational entities that encode concepts are found in two places in the system: the ***background*** (long-term memory, where there is effectively one entity per concept) and the ***foreground***, which is roughly equivalent to working memory, or the contents of consciousness, since it contains the particular subset of concepts that the system is using in its current thoughts and all aspects of its current model of the world.

The concept-entities in the foreground are referred to here as ***atoms***, while those in the background are called ***elements***. This choice of terminology is designed to make it clear that, in the simplest form of the molecular framework, each concept is represented by just one element in the background, whereas there can be many instances of that concept in the foreground. If the system happens to be thinking about several instances of the [chair] concept there would be several [chair] *atoms* in the foreground, but there would only be one [chair] *element* in the background.

For the purposes of this chapter we will almost exclusively be concerned with atoms, and (therefore) with events happening in the foreground.

The contents of the foreground could be visualized as a space in which atoms link together to form clusters that represent models of the state of the world. One cluster might represent what the system is seeing right now, while another might represent sounds that are currently being heard, and yet another might represent some abstract thoughts that the system is entertaining (which may not have any connection to what is happening in its environment at that moment). The function of the foreground, then, is to hold models of the world.

Theorists differ in their preference for atoms that are either active or passive. A passive approach would have all the important mechanisms on the outside, so that the atoms were mere tokens manipulated by those mechanisms. An active approach, on the other hand, would have few, if any, external mechanisms that manipulate atoms, but instead would have all the interesting machinery in and between the atoms. In the present case we will adopt the active, self-organized point of view: the atoms themselves do (virtually) all the work of interacting with, and operating on, one another. This choice makes no difference to the argument, but it gives a clearer picture of some claims about semantics that come later.

Two other ingredients that need to be mentioned in this cognitive framework are external sensory input and the system's model of itself. Sensory information originates at the sensory receptors (retina, proprioceptive detectors, ears, etc.), is then pre-processed in some way, and finally arrives at the "edge" of the foreground, where it causes atoms

representing primitive sensory features to become active. Because of this inward flow of information (from the sensory organs to the edge of the foreground and then on into the "interior" region of the foreground), those atoms that are near the edge of the foreground will tend to represent more concrete, low-level concepts, while atoms nearer the center will be concerned with more high-level, abstract ideas.

The ***self-model*** is a structure (a large cluster of atoms), somewhere near the center of the foreground, that represents the system itself. It could be argued that this self-model is present in the foreground almost all of the time because when the mind is representing some aspect of the world, it usually keeps a representation of its own ongoing existence as part of that world. There are fluctuations in the size of the self model, and there may be occasions when it is almost absent, but most of the time we seem to maintain a model of at least the minimal aspects of our self, such as our being located in a particular place. Although the self-model proper is a representation of the system, somewhere near to it there would also be a part of the system that has the authority to initiate and control actions taken by the system: this could be described as the *Make It So* place.

Finally, note that there are a variety of operators at work in the foreground, whose role is to make changes to clusters of atoms. The atoms themselves do some of this work, by trying to activate other atoms with which they are consistent. So, for example, a [cat] atom that is linked to a [crouching-posture] atom will tend to activate an atom representing [pounce]. But there will also be operators that do such things as *concept creation* (making a new atom to encode a new conjunction of known atoms), *elaboration* (where some existing atoms are encouraged to bring in others that can represent more detailed aspects of what they are already representing), various forms of *analogy building*, and so on.

This cognitive framework depicts the process of thought as a collective effect of the interaction of all these atoms and operators. The foreground is a molecular soup in which atoms assemble themselves (with the help of operators) into semi-stable, dynamically changing structures. Hence the use of the term "molecular framework" to describe this approach to the modeling of cognition.

15.2.3 *Explanation in General*

Atoms can play two distinct roles in the foreground, mirroring the distinction between *use* and *mention* of words. When the word "cat" appears in a sentence like "The cat is on the chair," it is being used to refer to a cat, but when the same word appears in a sentence like "The word *cat* has three letters," the word itself, not the concept, is being mentioned.

In much the same way, if the foreground has atoms representing a chair in the outside world a [chair] atom will be part of the representation of that outside situation, and in this case the [chair] atom is simply being used to stand for something. But if the system asks itself "What is a chair?", there will be one [chair] atom that stands as the target of the cluster of atoms representing the question. There is a strong difference, for the system, between representing a particular chair, and trying to ask questions about the *concept* of a chair. In this case the [chair] atom is being "mentioned" or referenced in the cluster of atoms that encode the question. It helps to picture the target atom as being placed in a special zone, or bubble, attached to the cluster of atoms that represent the question—whatever is inside the bubble is playing the special role of being examined, or mentioned. This is in contrast to the ordinary role that most atoms play when they are in the foreground, which is merely to be used as part of a representation.

So, when an atom, [x], becomes the target of a "What is x?" question, the [x] atom will be placed inside the bubble, then it will be elaborated and unpacked in various ways. What exactly does it mean to elaborate or unpack the atom? In effect, the atom is provoked into activating the other atoms that it would normally expect to see around it, if it were part of an ordinary representation in the foreground. Thus, the [chair] atom will cause atoms like [legs], [back], [seat], [sitting], [furniture] to be activated. And note that all of these activated atoms will be within the bubble that holds the target of the question.

What the question-cluster is doing is building a model of the meaning of [chair], inside the bubble. The various features and connotations of the [chair] concept try to link with one another to form a coherent cluster, and this coherent cluster inside the bubble is a model of the meaning, or definition, of the target concept.

One important aspect of this [meaning-of-"chair"] cluster is that the unpacking process tends to encourage more basic atoms to be activated. So the concepts that make up the final answer to the question will tend to be those that are subordinate features of the target atom. This is clearly just a matter of looking in the opposite direction from the one that is normally followed when an atom is being recognized: usually the activation of a cluster of atoms like [legs], [back] and [seat] will tend to cause the activation of the [chair] atom (this being the essence of the recognition process), so in order to get the meaning of [chair], what needs to happen is for the [chair] atom to follow the links backwards and divulge which other atoms would normally cause it to be activated.

We can call this set of elaboration and unpacking operations the *analysis mechanism*. Although it is convenient to refer to it as a single thing, the analysis mechanism is not really

a single entity, it is an open-ended toolkit of flexible, context-dependent operators. More like a loosely-defined segment of an ecology than a single creature. However, at the core of all these operators there will still be one basic component that grabs the target atom and starts following links to extract the other atoms that constitute the evidence (the features) that normally allow this atom to be activated. All other aspects of the analysis mechanism come into play after this automatic unpacking event.

If this were about narrow AI, rather than AGI, we might stop here and say that the essence of "explanation" was contained in the above account of how a [chair] concept is broken down into more detailed components. In an AGI system, however, the analysis mechanisms will have extensive connections to a large constellation of other structures and operators, including representations of, among other things:

- The person who asked the question that is being considered;
- That person's intentions, when they asked the question;
- Knowledge about what kinds of explanation are appropriate in what contexts;
- The protocols for constructing sentences that deliver an answer;
- The status and reliability of the knowledge in question.

In other words, there is a world of difference between a dictionary lookup mechanism that regurgitates the definition of "chair" (something that might be adequate in a narrow AI system), and the massive burst of representational activity that is triggered when a human or an AGI is asked "What is a chair?". The mental representation of that one question can be vastly different between cases where (say) the questioner is a young infant, a non-native-speaker learning the English language, and a professor who sets an exam question for a class of carpentry or philosophy students.

15.2.4 *Explaining Subjective Concepts*

In the case of human cognition, what happens when we try to answer a question about our subjective experience of the color red? In this case the analysis mechanism gets into trouble, because any questions about the essence of the color red will eventually reach down to a [redness] concept that is directly attached to an incoming signal line, and which therefore has no precursors. When the analysis mechanism tries to follow downward links to more basic atoms, it finds that this particular atom does not have any! The [redness] concept cannot be unpacked like most other concepts, because it lies at the very edge of the foreground: this is the place at which atoms are no longer used to represent parts of the

world. Outside the foreground there are various peripheral processing mechanisms, such as the primitive visual analysis machinery, but these are not within the scope of the operators that can play with atoms in the foreground itself. As far as the foreground is concerned the [redness] atom is activated by outside signals, not by other atoms internal to the foreground.

Notice that because of the rich set of processes mentioned above, the situation here is much worse than simply not knowing the meaning of a particular word. If we are asked to define a word we have never heard of, we can still talk about the letters or phonemes in the word, or specify where in the dictionary we would be able to find it, and so on. In the case of color qualia, though, the amount of analysis that can be done is precisely zero, so the analysis mechanism returns nothing.

Or does it return nothing? What exactly would we expect the analysis mechanism to do in this situation? Bear in mind that the mechanism itself is not intelligent (the global result of all these operations might be intelligent, but the individual mechanisms are just automatic), so it cannot know that the [red] concept is a special case that needs to be handled differently. So we would expect the mechanism to go right ahead and *go through the motions* of producing an answer. Something will come out of the end of the process, even if that something is an empty container where a cluster of atoms (representing the answer to the question) should have been.

So if the analysis mechanism does as much as it can, we would expect it to return an atom representing the concept *[subjective-essence-of-the-color-red]*, but this atom is extremely unusual because it contains nothing that would allow it to be analyzed. And any further attempt to apply the analysis mechanism to *this* atom will yield just another atom of the same element. The system can only solve its problem by creating a unique type of atom whose only feature is itself.

This bottoming-out of the analysis mechanism causes the cognitive system to eventually report that "There is definitely something that it is like to be experiencing the subjective essence of red, but that 'something' is ineffable and inexplicable." What it is saying is that there is a perfectly valid concept inside the foreground—the one that encodes the raw fact of redness—but that the analysis of this concept leads beyond the edge of the foreground (out into the sensory apparatus that supplies the foreground with visual signals), where the analysis mechanism is not able to go. This is the only way it can summarize the peculiar circumstance of analyzing [red] and getting [red] back as an answer.

This same short-circuit in the analysis mechanism is common to all of the consciousness questions. For qualia, the mechanism hits a dead end when it tries to probe the sensory

atoms at the edge of the foreground. In the case of emotions there are patterns of activation coming from deeper centers in the brain, which are also (arguably) beyond the reach of the foreground. For the concept of self, there is a core representation of the self that cannot be analyzed further because its purpose is to represent, literally, itself. The analysis mechanism can only operate within the foreground, and it seems that all aspects of subjective phenomenology are associated with atoms that lie right on the boundary.

In every case where this happens it is not really a "failure" of the mechanism, in the sense that something is broken, it is just an unavoidable consequence of the fact that the cognitive system is powerful enough to recursively answer questions about its own knowledge. If this were really a failure due to a badly designed mechanism, then it might be possible to build a different type of intelligent system that did not have this problem. Perhaps it would be possible to design around this problem, but it seems just as likely that any attempt to build a system capable of analyzing its own knowledge without limitations will have a boundary that causes the same short-circuit. Attempts to get the system to cope gracefully with this problem may only move the boundary to some other place, because any fix that is powerful enough to make the system not sense a problem, for these special concepts, is likely to have the unwanted side effect of causing the system to be limited in the depth of its analytic thought.

If a system has the ability to powerfully analyze its own concepts, then, it will have to notice the fact that some concepts are different because the cannot be analyzed further. If we try to imagine a cognitive system that is, somehow, not capable of representing the difference between these two classes of concepts, we surely get into all kinds of trouble. The system can be asked the direct question "When you look at the color red, what is the difference between that and the color blue? Because my friend here, who has never been able to see the color blue, would like to know." In the face of that direct question, the system is not only supposed to find no difference between its internal ability to handle the analysis of the [redness] concept and its handling of others, like the [chair] concept, it is also supposed to somehow not notice that its verbal reply contains the peculiarly empty phrase "Uh, I cannot think of any way to describe the difference." At some level, it must surely be possible for us to draw the attention of this hypothetical cognitive system to the fact that it is drawing a blank for some kinds of concept and not for others—and as soon as we can draw its attention to that fact, it is on a slippery slope toward the admission that there is a drastic difference between subjective phenomenology and objective concepts. There is something approaching a logical incoherence in the idea that a cognitive system can have

a powerful (i.e. human-level) analysis mechanism but also be immune to the failure mode described above.

15.2.5 *The "That Misses The Point" Objection*

The principal philosophical objection to the above argument is that it misses the point. It explains only the *locutions* that philosophers produce when talking about consciousness, not the actual experiences they have. The proposed explanation looks like it has slipped from being about the phenomenology, at the beginning, to being about the psychology (the cognitive mechanisms that cause people to say the things they do about consciousness) at the end. That would make this entire proposal into a discussion of a non-hard problem, because the philosopher can listen to the above account and yet still say "Yes, but why would *that* short circuit in my psychological mechanism cause *this* particular feeling in my phenomenology?"

Here we come to the crux of the proposed explanation of consciousness. Everything said so far could, indeed, be taken as just another example of a non-hard sidetracking of the core question. What makes this a real attempt o address the hard problem of consciousness is the fact that there is a flaw in the above objection, because *it involves an implicit usage of the very mechanism that is supposed to be causing the trouble*.

So if someone says "There is something missing from this argument, because when I look at my subjective experience I see things (my qualia!) that are not referenced in any way by the argument", what they are doing is asking for an explanation of (say) color qualia that is just as satisfactory as explanations of ordinary concepts, and they are noticing that the proposed explanation is inferior because it leaves something out. But this within-the-system comparison of consciousness with ordinary concepts is precisely the kind of thought process that will invoke the analysis mechanism! The analysis mechanism inside the mind of the philosopher who raises this objection will then come back with the verdict that the proposed explanation fails to describe the nature of conscious experience, just as other attempts to explain consciousness have failed.

The proposed explanation, then, can only be internally consistent with itself if the philosopher finds the explanation wanting.

There is something wickedly recursive about this situation. The proposed explanation does not address the question of why the phenomenology of the color red should be the way that it is—so in a certain respect the explanation could be said to have failed. But at the core of the explanation itself is the prediction that when the explanation is processed through

the head of a philosopher who tries to find objections to it, the explanation must necessarily cause the philosopher's own analysis mechanism to become short-circuited, resulting in a verdict that the explanation delivers no account of the phenomenology.

Do all of the philosophical objections to this argument fall into the same category (i.e. they depend for their force on a deployment of the analysis mechanism that is mentioned in the argument)? I claim that they do, for the following reason. The way that Chalmers [2] formulated it, there is a certain simplicity to the hard problem, because whenever an objection is lodged against any proposed resolution of the problem, the objection always works its way back to the same final point: the proposed explanation fails to make contact with the phenomenological mystery. In other words, the buck always stops with "Yes, but there is still something missing from this explanation." Now, the way that I interpret all of these different proposed explanations for consciousness—and the matching objections raised by philosophers who say that the explanation fails to account for the hard problem—is that these various proposals may differ in the way that they approach that final step, but that in the end it is only the final step that matters. In other words, I am not aware of any objection to the explanation proposed in this chapter that does not rely for its force on that final step, when the philosophical objection deploys the analysis mechanism, and thereby concludes that the proposal does not work *because* the analysis mechanism in the head of the philosopher returned a null result. And if (as I claim) all such objections eventually come back to that same place, they can all be dealt with in the same way.

But this still leaves something of an impasse. The argument does indeed say nothing about the nature of conscious experience, *qua* subjective experience, but it does say why it cannot supply an explanation. Is explaining why we cannot explain something the same as explaining it? This is the question to be considered next.

15.3 The Real Meaning of Meaning

This may seem a rather unsatisfactory solution to the problem of consciousness, because it appears to say that our most immediate, subjective experience of the world is an artifact of the operation of the brain. The proposed explanation of consciousness is that subjective phenomenology is a thing that intelligent systems *must* say they experience (because their analysis mechanism would not function correctly otherwise)—but this seems to put consciousness in the same category as visual artifacts, illusions, hallucinations and

the like. But something is surely wrong with this conclusion: it would be bizarre to treat something that dominates every aspect of our waking lives as if it were an artifact.

I believe that condemning consciousness as an artifact is the wrong conclusion to draw from the above explanation. I am now going to make a case that all of the various subjective phenomena associated with consciousness should be considered just as "real" as any other phenomena in the universe, but that science and philosophy must concede that consciousness has the special status of being unanalyzable. The appropriate conclusion is that consciousness can be predicted to occur under certain circumstances (namely, when an intelligent system has the kind of powerful analysis mechanism described earlier), but that there are strict limits to what we can say about its nature. We are obliged to say that these things are real, but even though they are real they are beyond the reach of science.

15.3.1 *Getting to the Bottom of Semantics*

The crucial question that we need to decide is what status we should give to the atoms in a cognitive system that have the peculiar property of making the analysis mechanism return a verdict of "this is real, but nothing can be said about it".

To answer this question in a convincing way, we need to understand the criteria we use when we decide:

- The "realness" or validity of different concepts (their epistemology);
- The meaning of concepts, or the relationships between concepts and things in the world (their semantics and ontology);
- The validity of concepts that are used in scientific explanations.

We cannot simply wave our hands and pick a set of criteria to apply to these things, we need to have some convincing reasons to make one choice or another.

Who adjudicates the question of which concepts are "real" and which are "artifacts"? On what *basis* can we conclude that some concepts (e.g. the phenomenological essence of redness) can be dismissed as "not real" or "artifactual"?

There seem to be two options here. One would involve taking an already well-developed theory of semantics or ontology—an off-the-shelf theory, so to speak—and then applying it to the present case. The second would be to take a detailed look at the various semantic/ontological frameworks that are available and find out which one is grounded most firmly; which one is secure enough in its foundations to be *the* true theory of meaning.

Unfortunately, both of these options lead us into a trap. The trap works roughly as follows. Suppose that we put forward a Theory of Meaning (let's call it Theory X), in the hope that Theory X will be so ontologically complete that it gives us the "correct" or "valid" method for deciding which concepts are real and which are artifacts; which concepts are scientifically valid and which are illusory/insufficient/incoherent.

Having made that choice, we can be sure of one thing: given how difficult it is to construct a Theory of Meaning, there will be some fairly abstract concepts involved in this theory. And as a result the theory itself will come under scrutiny for its conceptual coherence. Lying at the root of this theory there will be some assumptions that support the rest of the theory. Are those assumptions justified? Are they valid, sufficient or coherent? Are they necessary truths? You can see where this is leading: any Theory of Meaning that purports to be the way to decide whether or not concepts have true meaning (refer to actual things in the world) is bound to be a potential subject of its own mechanism. But in that case the theory would end up justifying its own validity by referring to criteria that it already assumes to be correct.

The conclusion to draw from these considerations is that any Theory X that claims to supply *absolute* standards for evaluating the realness or validity of concepts cannot be consistent. There is no such thing as an objective theory of meaning.

This circularity or question-begging problem applies equally to issues like the meaning of "meaning" and explanations of the concept of "explanation," and it afflicts anyone who proposes that the universe can be discovered to contain some absolute, objective standards for the "meanings" of things, or for the fundamental nature of explanatory force.

15.3.2 *Extreme Cognitive Semantics*

There is only one attitude to ontology and semantics that seems capable of escaping from this trap, and that is an approach that could be labeled "Extreme Cognitive Semantics"—the idea that there is no absolute, objective standard for the mapping between symbols inside a cognitive system and things in the world, because this mapping is entirely determined by the purely contingent fact of the design of real cognitive systems [3, 8]. There is no such thing as the pure, objective meaning of the symbols that cognitive systems use, there is only the way that cognitive systems actually do, as a matter of fact, use them. Meanings are determined by the ugly, inelegant design of cognitive systems, and that is the end of it.

How does this impact our attempt to decide the status of those atoms that cause our analysis mechanisms to bottom out? The first conclusion should be that, since the meanings and status of all atoms are governed by the way that cognitive systems actually use them, we should give far less weight to an externally-imposed formalism—like the possible-worlds semantics popular in artificial intelligence [4]—which says that subjective concepts point to nothing in the real world (or in functions defined over possible worlds) and are therefore fictitious.

Second—and in much the same vein—we can note that the atoms in question are such an unusual and extreme case, that formalisms like traditional semantics should not even be expected to handle them. This puts the shoe on the other foot: it is not that these semantic formalisms are capable of dismissing the consciousness-concepts and *therefore* the latter are invalid, it is rather that the formalisms are too weak to be used for such extreme cases, and therefore they have no jurisdiction in the matter.

Finally, we can use the Extreme Cognitive Semantics viewpoint to ask if there is a way to make sense of the idea that various concepts possess different degrees of "realness."

In order to do this, we need to look at how concepts are judged to be or not be "real" in ordinary usage. Ordinary usage of this concept seems to have two main aspects. The first involves the precise content of a concept and how it connects to other concepts. So, unicorns are not real because they connect to our other concepts in ways that clearly involve them residing only in stories. The second criterion that we use to judge the realness of a concept is the directness and immediacy of its phenomenology. Tangible, smellable, seeable things that lie close at hand are always more real. Abstract concepts are less real.

Interestingly, the consciousness atoms that we have been considering in this argument ([redness], [self] and so on) score very differently on these two measures of realness. They connect poorly to other concepts on their downward side because we cannot unpack them. But on the other hand they are the most immediate, closest, most tangible concepts of all, because they *define* what it means to be "immediate" and "tangible." When we say that a concept is more real the more concrete and tangible it is, what we actually mean is that it gets more real the closer it gets to the most basic of all concepts. In a sense there is a hierarchy of realness among our concepts, with those concepts that are phenomenologically rich being the most immediate and real, and with a decrease in that richness and immediacy as we go toward more abstract concepts.

15.3.3 *Implications*

What can we conclude from this analysis? I believe that the second of these two criteria of "realness" is the one that should dominate. We normally consider the concepts that are closest to our phenomenology to be the ones that are the best-connected and most thoroughly consistent with the rest of our conceptual systems. But the concepts associated with consciousness are an exception to that rule: they have the most immediacy, but a complete lack of connections going to other concepts that "explain" what they are. If we are forced to choose which of the two criteria is more important, it seems most coherent to treat immediacy as the real arbiter of what counts as real. Perhaps the best way to summarize the reason why this should be so is to consider the fact that in ordinary usage "realness" of a concept is to some extent inherited: if a concept is defined in terms of others that are considered very real, then it will be all the more real. But then it would make little sense to say that all concepts obey the rule that they are more real, valid and tangible, the closer they are to the phenomenological concepts at the root of the tree ... but that the last layer of concepts down at the root are themselves *not* real.

Given these considerations, I maintain that the correct explanation for consciousness is that all of its various phenomenological facets deserve to be called as "real" as any other concept we have, because there are no meaningful *objective* standards that we can apply to judge them otherwise. But while they deserve to be called "real" they also have the unique status of being beyond the reach of scientific inquiry. We can talk about the circumstances under which they arise, but we can never analyze their intrinsic nature. Science should admit that these phenomena are, in a profound and specialized sense, mysteries that lie beyond our reach.

This seems to me a unique and unusual compromise between materialist and dualist conceptions of mind. Minds are a consequence of a certain kind of computation; but they also contain some mysteries that can never be explained in a conventional way. We cannot give scientific explanations for subjective phenomena, but we can say exactly *why* we cannot do so. In the end, we can both explain consciousness and not explain it.

15.4 Some Falsifiable Predictions

This theory of consciousness can be used to make some falsifiable predictions. We are not yet in a position to make empirical tests of these predictions, because the tests would seem to require the kind of nanotechnology that would let us rewire our brains on the fly,

but the tests can be lodged in the record, against the day that some experimentalist can take up the challenge of implementing them.

The uniqueness of these predictions lies in the fact that there is a boundary (the edge of the foreground) at which the analysis mechanism gets into trouble. In each case, the prediction is that these phenomena will occur *at exactly that boundary*, and nowhere else. Bear in mind, however, that we do not yet know where this boundary actually lies, in the implementation that is the human brain.

If we are able to construct AGI systems that function at the human level of intelligence, with a full complement of cognitive mechanisms that includes the analysis mechanism described earlier, then these predictions will be testable by asking the AGI what it experiences in each of the following cases.

15.4.1 *Prediction 1: Blindsight*

Some kinds of brain damage cause people to experience 'blindsight', a condition in which the person reports little or no conscious awareness of a certain visual stimulus, while at the same time they can sometimes act on the stimulus as if it were visible [9].

The prediction in this case is that some of the visual pathways in the human brain will be found to lie within the scope of the analysis mechanism, while others will be found to lie outside. The ones outside the scope of the analysis mechanism will be precisely those that, when spared after damage, allow visual awareness without consciousness.

15.4.2 *Prediction 2: New Qualia*

If we built three sets of new color receptors in the eyes, with sensitivity to three bands in the ultraviolet range, and if we built enough brain wiring to supply the foreground with new concept-atoms triggered by these receptors, this should give rise to three new color qualia. After acclimatizing to the new qualia, we could then swap connections on the old color receptors and the new UV pathways, at a point that lies *just outside the scope of the analysis mechanism*. The prediction here is that the two sets of color qualia will be swapped in such a way that the new qualia will be associated with the old visible-light colors, and that this will only occur if the swap happens beyond the analysis mechanism.

If we then removed all trace of the new UV pathways and retinal receptors, outside the foreground (beyond the reach of the analysis mechanism), then the old color qualia would disappear, leaving only the new qualia. The subject will have a ghost of a memory of the old color qualia, because the old concept atoms will still be there, but those atoms will only

be seen in imagination. And if we later reintroduce a set of three color receptors and do the whole procedure again, we can bring back the old color qualia if we are careful to ensure that the new visual receptors trigger the foreground concept-atoms previously used for the visible-light colors. The subject would suddenly see the old qualia again

15.4.3 *Prediction 3: Synaesthetic Qualia*

Take the system described above (after the first installation of new qualia) and arrange for a cello timbre to excite the old concept-atoms that would have caused red qualia. Cello sounds will now cause the system to have a disembodied feeling of redness.

15.4.4 *Prediction 4: Mind Melds*

Join two minds so that B has access to the visual sensorium of A, using new concept-atoms in B's head to encode the incoming information from A. B would say that she knew what A's qualia were like, because she would be experiencing new qualia. If B were getting sounds from A's brain, but these were triggering entirely new atoms designed especially to encode the signals, B would say that A did not experience sound the way she did, but in an entirely new way. If, on the other hand, the incoming signals from A triggered the same sound atoms that B uses (with no new atom types being created), then B will report that she is hearing all of A's sonic input mixed in with her own. In much the same way, B could be given an extra region of her foreground periphery exclusively devoted to the visual stream coming from A. She would then say that she had two heads, but that she could only attend to one of them at a time. With new atoms for the colors, again, she would report that B's qualia differed from her own.

Note that any absolute comparison between the way that different people experience the world is not possible. The reported qualia in these mind-meld cases would be entirely dependent on choices of how to cross-wire the systems.

15.5 Conclusion

The simplest explanation for consciousness is that the various phenomena involved have an irreducible dual aspect to them. On the one hand, they are explicable because we can understand that they are the result of a powerful cognitive system using its analysis mechanism to probe concepts that happen to be beyond its reach. But on the other hand, these concepts deserve to be treated as the most immediate and real objects in the universe,

because they define the very foundation of what it means for something to be real. These consciousness concepts—such as the subjective phenomenological experience of the color red—cannot be explained by any further scientific analysis. Rather than try to resolve this situation by allowing one interpretation to trump the other, it seems more parsimonious to conclude that both are true at the same time, and that the subjective aspects of experience belong to a new category of their own: they are real but inexplicable, and no further scientific analysis of them will be able to penetrate their essential nature.

According to this analysis, an Artificial General Intelligence designed in such a way that it had the same problems with its analysis mechanism that we humans do (and I have argued that this would mean any fully sentient computer capable of a near-human degree of intelligence, because the analysis mechanism plays such a critical role in all types of general intelligence) would experience consciousness for the same reasons that we do. We could never prove this statement the way that we prove statements about objective concepts, but that is part of what it means to say that consciousness concepts have a special status (they are real, but beyond analysis). The only way to be consistent about our interpretation of these phenomena is to say that, insofar as we can say anything at all about consciousness, we can be sure that the right kind of artificial general intelligence would experience a subjective phenomenology comparable in scope to human subjective consciousness.

Bibliography

[1] Cambell, K. K. (1970). *Body and Mind* (Doubleday, New York).
[2] Chalmers, D. J. (1996). *The Conscious Mind: In Search of a Fundamental Theory* (Oxford University Press, Oxford).
[3] Croft, W. and Cruse, D. A. (2004). *Cognitive Linguistics* (Cambridge University Press, Cambridge).
[4] Dowty, D. R., Wall, R. E., and Peters, S. (1981). *Introduction to Montague Semantics* (D. Reidel, Dordrecht).
[5] Kirk, D. (1974). Zombies versus materialists, *Aristotelian Society* **48** (*suppl.*), pp. 135–52.
[6] Loosemore, R. P. W. (2007). Complex Systems, Artificial Intelligence and Theoretical Psychology, in B. Goertzel and P. Wang (eds.), *Advances in Artificial General Intelligence: Concepts, Architectures and Algorithms* (IOS Press, Amsterdam), pp. 159–173.
[7] Loosemore, R. P. W. and Harley, T. A. (2010). Brains and Minds: On the Usefulness of Localization Data to Cognitive Psychology, in M. Bunzl and S. J. Hanson (eds.), *Foundational Issues in Human Brain Mapping* (MIT Press, Cambridge, MA), pp. 217–240.
[8] Smith, L. B. and Smith, L. K. (1997). Perceiving and Remembering: Category Stability, Variability, and Development, in K. Lamberts and D. Shanks (eds.), *Knowledge, Concepts, and Categories* (Cambridge University Press, Cambridge).
[9] Weiskrantz, L. (1986). *Blindsight: A Case Study and Implications* (Oxford University Press, Oxford).