Interior Grounding, Reflection, and Self-Consciousness

Marvin Minsky

MIT

In *Brain, Mind and Society*, Proceedings of an International Conference on Brain, Mind and Society, Graduate School of Information Sciences, Brain, Mind and Society, Tohoku University, Japan, September 2005. See http://www.ic.is.tohoku.ac.jp/~GSIS/

This PDF version was created by Aaron Sloman on 7 Jun 2016, derived from: http://web.media.mit.edu/~minsky/papers/Internal%20Grounding.html

==========

Some computer programs are expert at some games. Other programs can recognize some words. Yet other programs are highly competent at solving certain technical problems. However, each of those programs is specialized, and no existing program today shows the common sense or resourcefulness of a typical two-year-old child --- and certainly, no program can yet understand a typical sentence from a child's first-grade storybook. Nor can any program today can look around a room and then identify the things that meet its eyes.

This lecture will suggest some ideas about why computer programs are still so limited. Some thinkers might say that this is because computers have no consciousness, and that nothing can be done about this, because it is in the nature of machines to only what they are programmed to do --- and therefore they cannot be programmed to 'think'.

> *Citizen: I am convinced that machines will never have thoughts or feelings like ours, because machines lack vital ingredients that can only exist in living things. So they cannot have any feelings at all, no hopes or joys or fears or pains --- or motives, ambitions, or purposes. They cannot have the faintest sense of pride or shame, or of failure, achievement, or discontent, because they simply can't care about what they do, or even know they exist.*

It seems to me that we use such statements to excuse ourselves for our failures to understand ourselves. To do this, we collect the phenomena that we can't yet explain, and then pack them into such 'suitcase-like' words as *sentience*, *spirit*, or *consciousness* --- and then describe these "vital ingredients" as entities with mysterious traits that can't be explained in physical ways.

However, here I will take an opposite view. Whenever some seemingly 'basic' aspect of mind seems hard to explain, I will try to depict it as the product of some more complex network of processes --- whose activities may sometimes cooperate, but may also have ways to conflict and compete. Then in each of the examples below, a mystery that seemed inexplicable will then be replaced by a set of several different questions and problems, each of which may still be difficult, but at least won't seem so more intractable. We'll start by unpacking the set of phenomena for which we have come to use the word "consciousness." (This following section is condensed from chapter 4 of my forthcoming book, "The Emotion Machine.")

==========

What is Consciousness?

> *Aaron Sloman: "It is not worth asking how to define consciousness, how to explain it, how it evolved, what its function is, etc., because there's no one thing for which all the answers would be the same. Instead, we have many sub-capabilities, for which the answers are different: e.g. different kinds of perception, learning, knowledge, ... self-control, etc."* --- From a message in comp.ai.philosophy, 14 Dec. 1994

To see how many things human minds do, consider this fragment of everyday thinking.

> *Joan is part way across the street on the way to deliver her finished report. While thinking about what to say at the meeting, she hears a sound and turns her head --- and sees a quickly oncoming car. Uncertain whether to cross or retreat, but uneasy about arriving late, Joan decides to sprint across the road. She later remembers her injured knee and reflects upon her impulsive decision. "If my knee had failed, I could have been killed. Then what would my friends have thought of me?"*

It might seem natural to ask, *"How conscious was Joan of what she did?"*But rather than dwell on that *'consciousness'* word, let's look at a few of the things that Joan actually *"did."*

> *Reaction: Joan reacted quickly to that sound.*
> *Identification: She recognized it as being a sound.*
> *Characterization: She classified it as the sound of a car.*
> *Attention: She noticed certain things rather than others.*
> *Indecision: She wondered whether to cross or retreat.*
> *Imagining: She envisioned some possible future conditions.*
> *Selection: She selected a way to choose among options.*
> *Decision: She chose one of several alternative actions.*
> *Planning: She constructed a multi-step action-plan.*
> *Reconsideration: Later she reconsidered this choice.*

In the course of doing those things, other 'parts' of Joan's mind did other things.

> *Recollection: She retrieved descriptions of prior events.*
> *Representation: She interconnected a set of descriptions.*
> *Embodiment: She tried to describe her body's condition.*
> *Emotion: She changed major parts of her mental state.*
> *Expression: She constructed several verbal descriptions.*
> *Narration: She heard them as dialogs in her mind.*
> *Intention: She changed some of her goals' priorities.*
> *Apprehension: She was uneasy about arriving late.*
> *Reasoning: She made various kinds of inferences.*

Many of these activities involved mental processes that used descriptions of some of her other mental processes.

> *Reflection: She thought about what she's recently done.*
> *Self-Reflection: She reflected on her recent thoughts.*
> *Empathy: She imagined other persons' thoughts.*
> *Moral Reflection: She evaluated what she has done.*

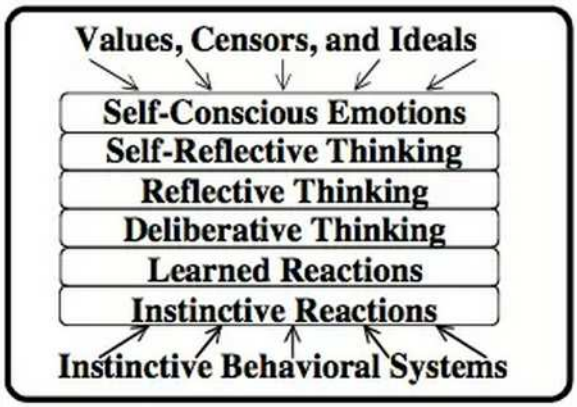
Self-Awareness: She characterized her mental condition.
Self-Imaging: She made and used models of herself.
Sense of Identity: She regarded herself as an entity.

That's only the start of a much longer list of aspects of how people feel and think --- and if we want to understand how our minds work, we'll need explanations for all of them. To do this, we'll have to take each one apart, to account for the details of how it works --- and then decide which of them to regard as aspect of what we call 'consciousness.'

==========

The mind seen as an organization of multi-level processes.

This section outlines a model of mind that shows how a system could reflect (to at least some extent) on what it was recently thinking about. There is not enough room to describe the whole idea here, but the reader can find more details at http://web.media.mit.edu/~minsky/E5/eb5.html.



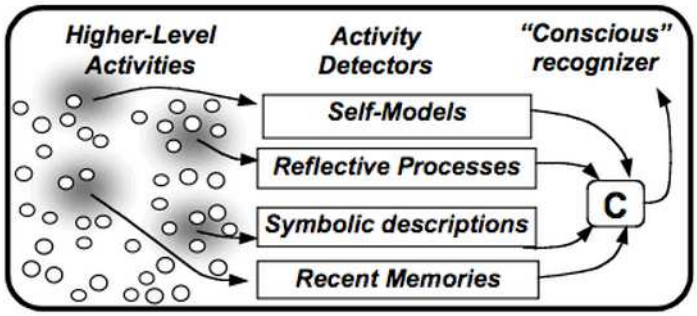
My associate Push Singh and I are at present developing a prototype of a system like this. We describe more details about this in http://web.media.mit.edu/~minsky/E4/eb4.html and in http://web.media.mit.edu/~push/CognitiveDiversity.html

I should note that this model is consistent with some of the early views of Sigmund Freud, who saw the mind as a system for resolving (or for ignoring) conflicts between our instinctive and acquired ideas.

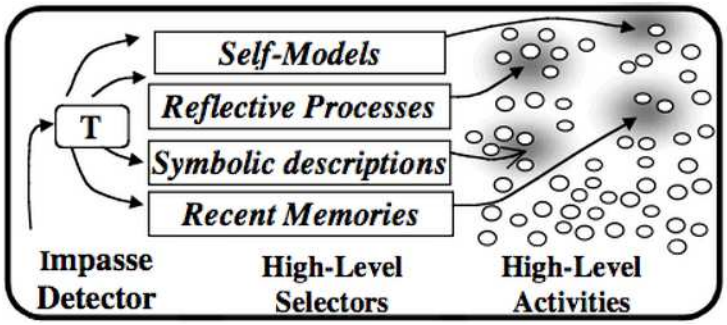
==========

How do we Recognize Consciousness?

Once we have a model in which the mind has several such layers of processes, we can start to construct hypotheses about what might be happening when a person claims to be thinking 'consciously'. For example, this might happen when a certain process in that person's brain detects some combination, such as this, of higher-level activities.



Similarly, we also could ask about what might cause a person to initiate such a set of activities. This might happen, for example, when a certain kind of 'critic-like' process detects that your thinking has got into trouble. The effect of such a critic might then help you to the sorts of things that we sometimes describe as trying to 'focus' or 'concentrate.' The diagram below suggests one kind of process that one's brain might use to try to switch itself into some pattern of thinking engages more high-level processes --- for example, by activating resources like these:



It is important to emphasize that each of those sets of activities can be extremely complex, and also are likely to differ significantly between different individuals. This is yet one more reason why no one should expect to be able to find a simple description of what we call consciousness.

==========

How do our mental levels develop and grow?

Over the past few centuries, our scientists have discovered far more about atoms and oceans and planets and stars than about the mechanics of feelings and thoughts. Those sciences progressed because those scientists were successful at discovering very small sets of "simple" and "basic" laws that explained many different phenomena in the realms of physics and chemistry.

Why did that strategy work less well for the science of psychology? It seems to me that one reason for this was an almost universal belief that such functions as emotions and feelings must be essentially non-mechanical --- and that, therefore, they could not be explained in terms of physical processes. However, I suspect that the principal cause of this delay was the idea that psychologists, like physicists, should also seek simple "laws" of thought. In other words, it seems to me, that our psychologists and philosophers should not have tried so hard to use the methods that worked so well for those physical sciences. In fact, today we know that every human brain contains several hundred different, specialized kinds of machinery --- each of which must have evolved different processes that helped our ancestors to solve the various problems that they faced in thousands of different ancient environments. So tens of thousands of different genes must be involved with how people think.

This suggests that modern psychologists should consider taking an opposite view, and reject the urge to base their ideas on discovering small sets of simple laws. Whenever some aspect of mind seems hard to explain (such as affection, fear, or pain), we could attempt, instead, to replace it by a more complex set of interconnected processes. In other words, we'll take each mental phenomenon and try to depict it, not as so 'basic' and 'elementary' that it is inexplicable, but as resulting from the complex activities of big networks of different processes --- which sometimes cooperate and sometimes compete. Then each mystery will begin to disappear, because of having been replaced by a several new kinds of problems. Each of those problems may still be quite hard, but because they are far less mysterious, we'll be able to start to deal with them.

In other words, our main technique will be to demonstrate many seemingly separate 'features' of our minds are actually not single things but are aspects of what happens inside huge networks of different processes. To do this we'll need to accumulate ideas about how some of those processes work, and then we'll need to propose some ways that these might combine to produce the systems that we call our minds.

So now let us try to apply this idea to the question of how human learning works. It is easy enough to imagine machines with many levels of processes; indeed, many computer programs today are made of multiple layers of sub-programs. However, we still do not have good hypotheses about how our higher levels of brain-machinery come to do all the wonderful things that they do.

==========

The myth of "Grounding in Experience."

Most theories of human development assume that we begin by learning low-level reactions, and must wait for each stage to consolidate before we can learn to think more abstractly:

> *"Everything that we come to know --- from the simplest facts to our most abstract concepts --- is ultimately "grounded" on our experiences with the external world."*

More specifically, that 'standard theory' goes on to insist:

We begin by (somehow) learning to recognize particular sensory situations. Then we correlate our reactions with whether they lead to failure and success.

Then, in subsequent stages of development, we learn increasingly abstract ways to represent the objects and their relationships in the situations that we perceive.

However, this raises serious questions like these:

> *How do we recognize those "sensory situations"?*
> *How do we represent them?*
> *What determines how we react to them? ("Operants.")*
> *What constitute 'success' and 'failure'?*
> *How do we make those correlations?*

To answer such questions, it seems to me, we will need many new ideas about how to design such machinery. I doubt that it will ever suffice to assume (for example) that learning is basically a matter of statistical correlations, or that high-level concepts will spontaneously form in large neural networks with simple architectures --- or that we will come to understand much of human cognition by making small extensions to traditional concepts about "association of ideas" or "operant reinforcement." One great philosopher clearly recognized that those ideas had serious deficiencies:

> Immanuel Kant: *"That all our knowledge begins with experience there can be no doubt. For how is it possible that the faculty of cognition should be awakened into exercise otherwise than by means of objects which affect our senses, and partly of themselves produce representations, partly rouse our powers of understanding into activity, to compare, to connect, or to separate these --- and so to convert the raw material of our sensations into a knowledge of objects?"*
>
> *"But, though all our knowledge begins with experience, it by no means follows that all arises out of experience. For, on the contrary, it is quite possible that our empirical knowledge is a combination of that which we receive through impressions, and [additional knowledge] altogether independent of experience ... which the faculty of cognition supplies from itself, sensory impressions giving merely the occasion.*
> [Immanuel Kant, Introduction to *Critique Of Pure Reason*, Second edition, April 1787]

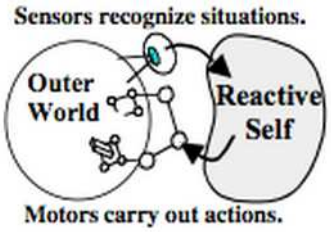
For although, as Kant remarked, sensations give us occasions to learn, this cannot be what makes us able to learn: in other words, it does not seem to explain how a person first could *learn to learn*. Instead, you need to begin with some 'additional knowledge' about how to *produce representations* and then *to connect* them. This is why, it seems to me, our human brains first had to evolve the kinds of complex architectures that our neuroscientists see.

For example, the traditional points of view do not begin to explain why the ’stages’ of children’s development so frequently seems highly abrupt; a child may spend an entire year expressing only "sentences" that contain no more than one or two words --- and then, more complex expressions may quickly appear. This has led to a belief that has been popular for many years: that such capabilities must simply be "innate," and are actually not "learned" at all. Accordingly, that viewpoint holds that the child needs only to "tune up" or, in some way, adapt that machinery to the language of its culture, so that it can automatically speak properly when the developmental "time is right." The following section suggests, instead, that different levels of learning could have been proceeding simultaneously throughout that period, but do not usually appear in overt behavior until the resulting processes have become sufficiently competent.

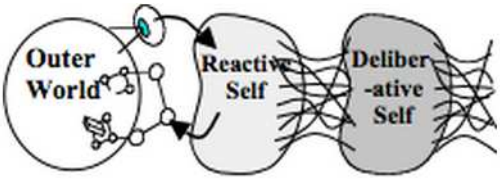
==========

A Theory of "Interior Grounding"

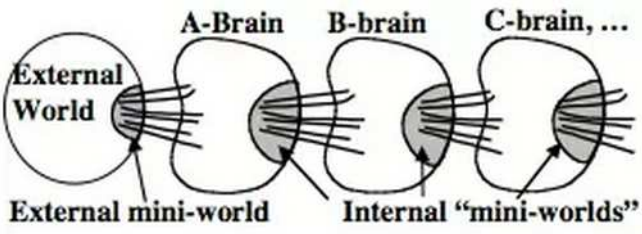
The old ’physical grounding hypothesis’ assumed that no ’higher cognitive level’ could start to learn before the levels below it have learned enough. In this view, mental development must begin with processes in which the child’s lowest level reactive systems acquire some knowledge about that child’s external environment:



Only then could the next level start to learn --- because (in that traditional view) the construction of each new structure must be based on the foundations of what the levels below it have learned.



However, we can imagine a different kind of process in which each of several levels of the brain can, at the same time, learn some ways to predict and control some of the activities in the parts of the brain to which it directly connects. In other words, each part of the brain exists inside its own ’local world’. Then we can make a new hypothesis: evolution could have provided each of those local worlds with what we might call "mini-worlds" that genetically have been already each equipped with potentially useful kinds of behaviors.

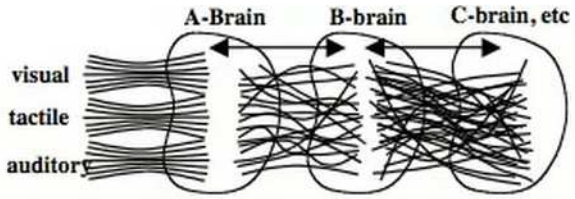


A typical *external* mini-world might consist of the system comprising some fingers and hand; then the reactive system can learn to predict how various combinations of finger-motions lead to different palm-sensations. Such a system could learn to predict that clenching the fingers will cause a sense of pressure on the palm. Similarly, an infant reactive level could learn to predict the effects of larger-scale motions of the limbs, or motions of the tongue in the mouth, or some visual effects of moving the eyes.

So far, this is the conventional view, in which all of our learning is finally based on what we learn from our experience with the external world. However, we could also imagine that some similar processes could also work at higher levels inside the same brain. For example, some higher levels could begin with connections to small systems that behave like simple finite-state machines. An example of such a system might have three state and two actions ’move left’ and ’move right.’

If that system behaves like three points on a line, then the B-brain could learn to predict (for example) that performing ’move left’ two or more times will always put the system into the leftmost state. There are many things that could be learned from this: that some actions are reversible, while other are not --- and that this can depend, in various ways, on the situation that the system is in. There are many other important things that could be learned by such machines: for example, about how different sequences of actions can be combined, or about the effects of various kinds of such modifications.

How could such a system evolve? The simplest hypothesis would be the each of the major cognitive parts of our brains is based on mutated copies of structures that already existed. Then each new such level might contain mutated versions of older learning machines, already equipped with primitive innate goals to predict the effects of imagined action-chains. Then several parts of an infant’s mind could each learn, simultaneously, some ways to predict and control its ’local environment.’



Eventually, these almost-separate systems would expand so that each of the levels inside that brain goes on to develop more powerful ways to exploit the abilities that its neighbors have learned.

==========

Representations of Knowledge

Any theory of learning must try to include some ideas about how the learning machine might represent the knowledge information that it acquires. Most traditional theories assume that learning is somehow based on making connections --- but only rarely go on to suggest the character of the things that are being connected. The situation is quite the opposite in the context of Computer Science, and practitioners frequently argue about what is the best way to represent knowledge. Sometimes such arguments go like this:

"It is always best to use rigorous Logic."

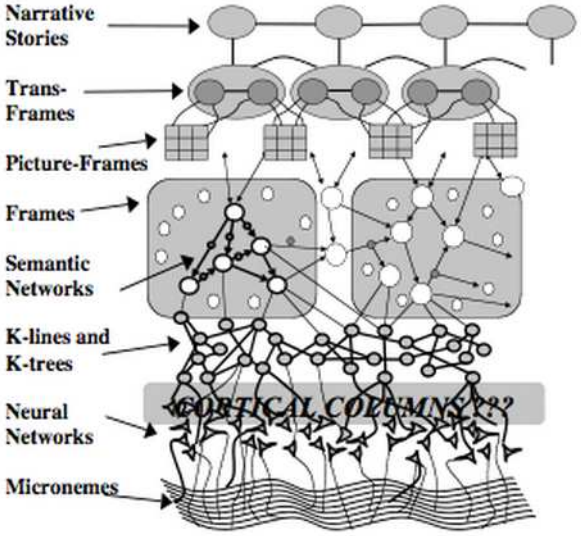
"No. Logic is too inflexible. Use Neural Networks."

"No, Neural Nets are even more rigid. They describe things with numbers, instead of abstractions. Why not simply use Natural Language."

"No, use Semantic Networks instead --- where different ideas are connected by concepts! Such networks are better than sentences are --- and have fewer ambiguities."

"No, Semantic Nets are too flexible --- and can lead to inconsistencies. Only Logic protects us from paradox."

Chapter 8 of *The Emotion Machine* discusses this in more detail, and concludes that so far as human brains are concerned, we must use many different ways to represent different kinds of knowledge. That discussion concludes by imagining that human common sense knowledge must use a variety of different methods and processes that results in arrangements that might look like this:



==========

The problem of Subjective Experience

Many philosophers have claimed that the hardest problem we need to face, both in psychology and philosophy, is to understand the nature of Subjective Experience. For example, here is one statement of this.

David Chalmers: "The hard problem, in contrast, is the question of how physical processes in the brain give rise to subjective experience. This puzzle involves the inner aspect of thought and perception: the way things feel for the subject. When we see, for example, we experience visual sensations, such as that of vivid blue. Or think of the ineffable sound of a distant oboe, the agony of an intense pain, the sparkle of happiness or the meditative quality of a moment lost in thought.â¦ It is these phenomena that pose the real mystery of the mind."

(See http://eksl-www.cs.umass.edu/~atkin/791T/chalmers.html and http://consc.net/papers/puzzle.html. For more details, see *Journal of Consciousness Studies* 2(3): 200-19, 1995 or http://consc.net/papers/facing.html.)

Chalmers went on to propose an answer to this, by advocating a form of dualism, in which that sense of experience is regarded as a fundamental feature or property of the world.

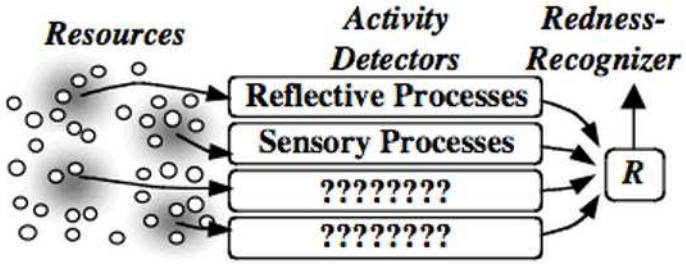
> *Chalmers: "This leads to a natural hypothesis: that information (or at least some information) has two basic aspects, a physical aspect and a phenomenal aspect. This has the status of a basic principal that might underlie and explain the emergence of experience from the physical. Experience arises by virtue of its status of one aspect of information, when the other aspect is found embodied in physical processing. ... Of course, the double-aspect principle is extremely speculative and is also underdetermined, leaving a number of key questions unanswered.*"

Similarly, many thinkers have maintained that our sensations have certain 'basic' or 'irreducible' qualities that stand by themselves and can't be 'reduced' to anything else. For instance, in such a view, each color like *Green* and each flavor like *Sweet* has its own indescribable character, which is unique and can't be explained. For if such qualities do not have any smaller parts or properties, then there's no possible way to describe them.

Those thinkers call this the problem of 'Qualia', and argue that qualities of sensations cannot be explained in physical terms, because they have no physical properties. To be sure, it is easy to measure the amounts of Red light that comes a splotch of paint, or how much sugar is in each piece of a peach, but such comparisons (those philosophers claim) tells us nothing about the nature of the *experience* of seeing *redness* or tasting *sweetness*.

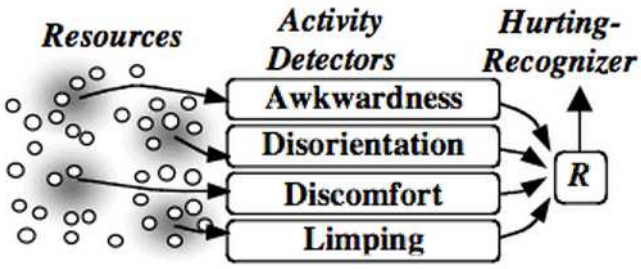
This subject might seem important because, if we cannot explain such 'subjective' things, that would undermine the whole idea that we can explain the human mind entirely in terms of such physical things as the machinery inside our brains. For, if the sensation of *sweetness* can never be measured or weighed, *or detected in any physical way*, then it must exist in a separate mental world, where it cannot possibly interact with any physical instruments.

Well, let's first observe that this claim must be wrong, because it is self-contradictory. For, if you can tell me that you have experienced *sweetness* then, somehow, *that sensation has caused your mouth to move!* So clearly, there must be some 'physical instrument' in your brain that recognized the mental activity that embodies your experience. In other words, we are simply facing, again, the same kind of problem that we solved in the previous section: we simply need another one of those internal "condition-detecting" diagrams, like the ones that we used above to account for why and how a person might talk about consciousness.



Similarly, Joan might first notice a change in her gait, or that she's been favoring her injured knee. Indeed, her friends may be more aware than she is, of how much that pain was affecting her. Thus, one's first awareness of pain may come after detecting signs of its effects --- by using the kind of

machinery.



Of course, we do not yet quite know how to construct and connect those condition detectors. However, so far as I can see, this is merely another instance of where our popular psychology assumes that some mental phenomenon is far simpler than it actually is. Perhaps in just a few years from now we shall be able to ask a brave philosopher to enter a suitable scanning device so that we can discover which brain-cells best distinguish the conditions that we wish to detect.

In other words, to understand how feelings work in more detail, we'll have to stop looking for simple answers, and start to explore more complex processes. The sensory systems in human brains include dozens of different processors. So, when you try to tell someone else about the 'sensations' you 'experience, those pathways are so complex and indirect that you will be telling a story based on sixth-hand reports that have gone through many kinds of transformations. So despite what those philosophers claim, there is no basis to insist that what we 'experience' is uniquely 'direct.'

When a ray of light strikes your retina, signals flow from that spot to your brain, where they affect other resources, which then transmit other kinds of reports that then influence yet other parts of your brain. [NOTE: In fact, a single spot of red may not be sensed as being red; in general the colors we see depend, to a large extent, on which other colors are in its neighborhood. Also, some readers might be surprised to hear that the visual system in a human brain includes dozens of different processing centers.]

Also, at the same time signals from the sensors in your ears, nose, and skin will travel along quite different paths, and all these streams of information may come to affect, in various ways, the descriptions the rest of your mind is using. So, because those pathways are so complex and indirect, when you try to tell someone about what sensation you feel, or what you are experiencing, you'll be telling a story based on sixth-hand reports that use information that has gone through many kinds of transformations. So despite what some philosophers claim, we have no basis to insist that what we call our sense of 'experience' is uniquely direct.

The old idea that sensations are 'basic' may have been useful in its day, the way the four kinds of 'atoms' of antiquity were supposed to be elementary. But now we need to recognize that our perceptions are affected by what our other resources may want or expect.

Now some philosophers might still complain that no theory like this can truly describe or explain the *experience* of seeing that color or feeling that touch. Listen again to the best of those philosophers:

> *David Chalmers: "When we visually perceive the world, we do not just process information; we have a subjective experience of color, shape, and depth. We have experiences associated with other senses (think of auditory experiences of music, or the ineffable nature of smell experiences), with bodily sensations (e.g., pains, tickles, and orgasms), with mental imagery (e.g., the colored*

shapes that appear when one rubs ones eyes), with emotion (the sparkle of happiness. the intensity of anger, the weight of despair), and with the stream of conscious thought.

*"[That we have a sense of experiencing] is the central fact about the mind, but it is also the most mysterious. Why should a physical system, no matter how complex and well-organized, give rise to experience at all? Why is it that all this processing does not go on "in the dark", without any subjective quality? Right now, nobody has good answers to these questions. This is the phenomenon that makes consciousness a *real* mystery."* See http://consc.net/papers/puzzle.html or http://consc.net/papers/facing.html.

Here is how I would deal with that 'mystery.' When you see your friend Jack react to things, you cannot see the machinery that makes him react in those ways --- and so you have few alternatives to simply saying that, *"he reacts to what he experiencing."* But then, you must be using the word *'experience'* as an abbreviation for what you would say if we knew what had happened inside Jack's --- such as, *"He must have detected some stimuli, and then made some representations of these, and then reacted to some of those by changing some of the plans he had made, etc."*

In other words, if your brain can begin to speak about some 'experience' it must already have access to some representations of that event; otherwise, you would not remember it --- or be able to say that you have experienced it! So your very act of discussing that 'experience' shows that 'it' cannot be a simple or basic thing, but must be a complex process that is involved with the high-level networks of representations that you call your Self.

When seen this way, the problem which Chalmers calls 'hard' is not really a single problem at all, because it condenses the complexity of all those many steps by squeezing them into the single word, 'experience' and then declares this to be a mystery. From this point of view, there should be nothing surprising about the fact that you find it so hard to talk about your sensations and feelings? You look at a color and see that it's Red. Something itches your ear and you know where to scratch. Then, so far as you can tell, that's all there seems to be to it; you recognize that experience --- and nothing like "thinking" seems to intervene. Perhaps this is what leads some people to think that the qualities of such sensations are so basic and irreducible that they will always remain inexplicable.

However, I prefer to take the opposite view --- that what we call *sensations* are complex reflective activities. They sometimes involve extensive cascades in which some parts of the brain are affected by signals whose origins we cannot detect --- and therefore, we find them hard to explain. So, I see no exceptional mystery here: we simply don't yet know enough about what is actually happening in our brains. But when you think enough about anything, then you see this is also the case with everything.