Macalester Journal of Philosophy

Volume 20 | Issue 1 Article 7

6-21-2012

Consciousness and AI: Reformulating the Issue

Patrick Holzman
Macalester College

Consciousness and AI: Reformulating the Issue

Patrick Holzman

Abstract In this paper, I explore the “issue” of consciousness in artificial intelligences, the problem of whether they can be conscious, specifically going for simply asking what consciousness involves, instead of more technical aspects of the field. I use Robert Kirk's concepts of the "Basic Package" as well as "Direct Activity" to outline what being conscious involves, and attempt to apply it to artificially designed and constructed beings. I assume that artificial conscious intelligences will be constructed, eventually; my goal is to suggest a specific and more useful way of thinking about consciousness, which will hopefully accelerate the inevitable.

The science of artificial intelligence deals with attempts to make programs or machines that can function in an intelligent way. What "intelligent" means is dependent on our own judgment and defined for the most part in terms of our own actions. Humans (and animals) act "intelligently," and so when we want to create an artificial intelligence, what we want is something that acts like us, that at least appears to make complex judgments and choices about its environment. Note that I have deliberately phrased this description of AI with phrases like "acts intelligently," or "appears to make judgments," or "functions in a certain way." That is, I've put these goals in terms of what the intelligences do, what their

behavior is, without any mention of their internal structure, and in doing so l leave open the question of consciousness and the "mind."

From my fairly limited understanding of the perspective of those working with AI, this is entirely reasonable. The goal of engineers working in AI is to create something that acts intelligently. The challenge is the execution, the structure of the program, but the goal itself is purely based on behavior. I might even be so bold as to say that many researchers in AI assume that "consciousness," and rational judgment, whatever these involve, will come out in the wash, around when we get things that can truly act like a person. However, I feel that consciousness should be a goal in itself,[1] and the path to consciousness will involve the amplification of some already existing "bare awareness" or "internal life" found in all systems. I will first talk a bit about the field of artificial intelligence, then consciousness in general (that is, attempt to define what I'm talking about in the first place), then introduce Robert Kirk's idea of the "Basic Package" or the "decider," and then finally his concept of "direct activity" and the "Basic Package Plus" to work out how one could judge whether a thing has consciousness or not. Using these, Kirk constructs a model of consciousness, or at least the salient aspects of consciousness in terms of testing for it. I agree with his views, and ultimately I will conclude that the phenomenal aspect of consciousness is not as important as the ability to make judgments and to actually understand the world.

---

[1] More precisely, I feel that creating an AI that would qualify as one of Kirk's "deciders" is worthwhile as a goal; I think it will be clear why after I describe Kirk's concepts.

The claim "AI researchers don't care about consciousness" is something of a straw man, but I want to dispel even the smallest hint of behaviorism. To be blunt, logical behaviorism, the opinion that all there is to having a mind is acting in a certain way, seems absolutely incomprehensible. In this sort of behaviorism, the statement "he believes that it will rain" is identical with the statement "he carries an umbrella and otherwise acts in certain ways." But this is quite false. Consider a hypothetical table-based system, wherein all conceivable inputs are associated with various outputs: input A at state J causes output X and state K, input B at state M causes output Y and state N, all down a table. Given a long enough table, and a fast enough method to access it, you could have an AI that perfectly replicated a human.[2] However, it seems rather obvious that this would not possess a rational mind, would not analyze the world or make judgments, but instead function purely through reaction.

A behaviorist would say that this is an unfair criticism; such a thing would be impossible to execute. If we were to create a being that acted like a human, and fully like a human, able to react to an indefinite variety of situations, and its hardware was limited to something the size of a human head, then it seems reasonable to assume that such a thing would likely be acting in a complex, rich way, actually having an internal functioning of similar convolution to ours, if likely with a different sort of structure. Phrased this way, behaviorism is much more about practical judgments about the nature of things we could encounter or build. This still misses the

[2] I do mean, though, an *extremely* long table, with a great many states and inputs. Essentially the false human's entire life story would count as a single state.

point behind asking whether something really possesses a mind or is conscious—unless there is something beyond the material, possession of a mind must line up with some physical state of things. Furthermore, advances in technology may allow us to make astonishingly complex AIs that, nonetheless, will have no true mind.

The goal at the moment is to make programs that solve problems humans are still not very good at, such as traffic control or chess. Generally this is done by formalizing the situation mathematically, then writing a program to manipulate this formalization and find what best fits a certain criteria. It is not that "is efficient" is the criterion, but rather that there is some variable in the formalization that the program attempts to minimize or maximize. Once the formalization is "translated" back into our own understanding of the problem, this variable is identified with efficiency, but a significant part of the work when making the artificial intelligence is this formalization, in determining how best to abstractly represent the problem. Even when researchers attempt to create AIs that learn through something like a "neural net," they must first create a domain within which the AI will function; the problem has changed from solving a certain problem in a certain language, to working out the language and what problem is involved while still using a certain other language.

Here I want to begin to use words like "syntax" and "semantics," but I think doing so would be dangerous—such words have been used many times before and have vague definitions.[3] It

[3] I also suspect the way I use these words, or at least what I consider important about them, is different from many others'.

might be best to carefully lay out what I'm talking about. When I refer to consciousness, I am to a certain extent going by Nagel's idea of having a "what it's like."[4] My eventual conclusion is, however, that the phenomenal aspect of consciousness, this feel, is not what is valuable. Rather, the value comes in the ability to make judgments about the world; whether our perception is immediate or has no phenomenological aspect is irrelevant. The distinction I'm trying to focus on is one between "consciousness" and "awareness," between "perception" and "sensation." Awareness, sensation, is simply having a first-person perspective from which certain things are experienced, while consciousness, perception, is to have some context, some interpretation of that sensation. Perception is sensation with internal context; consciousness is awareness with actual *meaningful* content. This is a very fuzzy distinction, one that will be better distinguished when I get to Kirk's deciders and the basic package. One possible way to think of it is by the concept of "raw feels," which here would just be sensation. The "raw feels," the sensations, are the raw bits of context-less information that comes into a system, which for some things is then interpreted and becomes perception, becomes conscious. For those things which are not conscious, sensation cause some reflexes to fire, and in this manner they are yet aware.

Terms and Assumptions

In earlier versions of this paper I freely used "consciousness" when what I meant was "a mind," in this sense of

[4] Nagel, Thomas. 1974. "What Is It Like to Be a Bat?" *The Philosophical Review.* 83 (4): 435-450.

"rational and analyzing" that I've been stressing. To jump ahead, the distinction between an unconscious and a conscious mind is, in Kirk's terms, the addition of "direct activity," the fact that for a conscious mind perception is irresistible and happens automatically. Perception of the world directly affects the mind, changing how that being will achieve its goals or what its goals are, with no need to reflect on its body of knowledge. This “direct activity” is what is entailed by consciousness, the “what it’s like.” The actual rationality should not be properly referred to as “consciousness,” except in that explicit sense of rationality. That is, we are not discussing “conscious vs. unconscious,” but rather “conscious vs. reflexive” or something along those lines.

By "system" I really do mean any sort of system. For the most part I'm talking about complex life-forms, but computer programs, robots, even things like toasters or thermostats count as a "system." The "basic package" and "deciders" will be detailed in more depth later, but essentially a "decider" is something that analyzes information about its environment, forms goals, and then executes those goals. I will say that systems that are deciders have "minds," and "mind" here means "rational, complex mind." "Consciousness" refers to minds or deciders that have direct activity, Kirk's "basic package plus." Unfortunately, I do not have a simple term for systems that are not deciders that still have direct activity. I think they could be called "non-rational sensing systems."

I will go ahead and assume there is nothing beyond our physical bodies at work when we speak of the mind. Our brains do things, and this activity, from a different perspective, is called “mind.” Brain activity does not “produce” minds over and above

the brain, they is not “caused” by that activity. Rather, the activity in the brain somehow is the activity in the mind. You could not have a human brain functioning the way it does without also having a mind. This is not a contingent fact, but rather a fact about the sort of activity that occurs in the brain, that it is also conscious mental activity, when viewed from the inside. I take this as a matter of faith, and feel no need to defend this. It seems for the most part obvious, and, honestly, not that interesting.

I will also assume that consciousness is an interesting and worthy topic of discussion. There is something that it is to be conscious; you and I can feel it just by thinking. This needs to be acknowledged, and explored. Finally, I will also assume that the mind could best be described as "activity within the brain when viewed from a different perspective." Searle gives this formulation: “Mental phenomena are caused by neurophysiological processes in the brain and are themselves features of the brain.”[5] I would agree, with a caveat about the exact use of language. I want to stress that the mind is not “caused” by the brain, but is the brain; I think I mean the same as Searle, but that I am insisting on a certain language. Searle talks about the mind being an emergent property of the brain, in the same way that wetness is an emergent property of H2O. Now, is wetness “caused” by the H2O? Not exactly, not in the same sense that a rock causes a window to shatter. H2O does not “produce” wetness, but rather it is wet, in sufficient quantities. “Produce” and “cause” evoke to me feelings of “extrude” and “impart,” not “possess.” However, if we are to say that the brain “causes” the mind in the same way that H2O “causes” wetness,

[5] John R. Searle, *The Rediscovery of the Mind* (Cambridge: MIT, 1992, 1.

then it's quite fine. Indeed, the other half of his formulation is that mental phenomena "are themselves features of the brain." Although I've focused on just this example, I think similar sorts of situations abound and are what actually make up many of the apparent differences among various theories.

What is tenuous, and interesting, is the use of "from a different perspective," when saying what the mind is. An objection can be raised that this physicalist explanation does not account for all aspects of the mental, that there is an "explanatory gap." Nagel's Bat[6] and Jackson's Mary[7] are paradigmatic thought experiments/arguments for this "internal perspective." The explanatory gap implies that knowledge of the physical world will not give you the knowledge of "what it's like." However, for now I am not focused so much on *what* it is like to be something, but rather *whether* there is a "what it's like" for any specific thing. I suspect there is a "what it's like" to be a cat, and that there is not a "what it's like" to be a rock, and furthermore that we can tell this empirically, and where the line is, just from physical facts. Even if physical facts can't tell *what* it is like, they can still tell *whether* there is a "what it's like." It is interesting to pursue whether we can tell "what it's like" to be something, and I will do so, somewhat, but the difficulties we have in doing so do not change our knowledge that there is a "what it's like" to be something.

---

[6] Thomas Nagel, "What is it like to be a bat?" (*Philosophical Review,* 1974).
[7] Frank Jackson, "What Mary didn't know" (*Journal of Philosophy,* 1986). I'm going to assume some familiarity with both of these.

Brain/Mind Identity

Before I get much farther, it may be valuable for me to clarify my position, especially towards mind/brain identity. The Stanford Encyclopedia of Philosophy states "The identity theory of mind holds that states and processes of the mind are identical to states and processes of the brain. [. . .] Consider an experience of pain, or of seeing something, or of having a mental image. The identity theory of mind is to the effect that these experiences just *are* brain processes, not merely *correlated with* brain processes."[8] On my view, there are two ways you can interpret this in terms of AIs possessing minds. One way is to say that, obviously, an AI cannot have a mind, as minds are identical to brains and thus non-brain possessing AIs will not possess minds. The other way, my way, is to say that AIs can quite easily possess minds, just minds that are very unlike our own, as their brains are unlike our own, *in structure*. What is the mind in the brain is the brain's structure—the mind is not a non-physical object whose parts can be identified with the parts of the physical object of the brain, but rather the organizational relations of the mind are identified with the relations in the brain. The mind is already nothing more than a set of relations; what the mind is identical with in the brain is those relations of the parts of the brain.

The sort of identity theory I agree with is an odd sort of token-token identity. A token of some activity in the brain is identical with a token of some activity of the mind. Types of tokens in the mind are defined in terms of behavior and similar

[8] J. J. C. Smart, "The Identity Theory of Mind" (*The Stanford Encyclopedia of Philosophy*, 2008).

phenomenological characteristics, and those tokens in the brain have similarities as well, but the link of types of mental tokens to types of physical tokens are on account of the linkage of the tokens themselves. “Pain is identical to c-fiber firing” does not mean that a being with no c-fibers cannot feel pain. Rather, individual tokens of pain in humans are found to link to tokens of c-fiber firing, but the link between pain and c-fiber firing *in general* is only on account of the commonality of the tokens of pain. In a being without c-fibers that still feels pain, such as a robot, we could say “pain is identical to a red wire firing,” or whatever the case is.

Obviously this leaves the question of whether the pain in the robot is the same as the pain in us. I feel it is not, unless we’ve made an effort to make a robot with the same physical and mental structure as us, but I still feel that it is reasonable to say it has pain, as long as it has connections to its physical body that induce unpleasant sensations in it and that serve a similar role as pain in us.; “unpleasant” will be dependent on whatever reward mechanisms we design it to have. If a robot is has a mind, has a way of forming goals, some of which include the preservation of itself, has ways of gathering information about damage to its body, and has some sort of unavoidable phenomenological sensation that carries this information to its mind which encourages it to avoid that damage, then it has sensations that can be usefully called “pain.” It may not be pain *like ours*, but it is no *less* pain that ours is.

Kirk's Basic Package[9]

In order to deal with things like consciousness, or the mind, or the idea of an "internal perspective," you need fairly strong definitions, or at least reasonably clear guidelines of what will constitute such things. Instead of trying to form yet another new framework, I've decided to use Robert Kirk's ideas of the "basic package," and "deciders," as well as his "direct activity," because the entire system seems to be the most reasonable and acceptable one I've read yet. A decider is something that makes judgments about the world, analyzes it, and forms goals and what it sees as the most appropriate paths to those goals. This is in contrast with systems that act purely on reflex, and Kirk uses this contrast extensively to lay out what he means by a "decider." An example of a general reflex system would be a clam, which shuts its shell when exposed to certain sensations. It is important to remember the distinction I tried to make between consciousness and a "mind"—a conscious mind is a mind with this direct activity, but a system does not need to have a full rational mind to have direct activity. A clam still has sensation, and an internal perspective, despite not having the full, rich consciousness it would possess with the basic package. What Kirk focuses on is "perception," which refers to sensation in a system that can learn, and which is an integrated part of a conscious mind. Fully conscious systems are partially defined by perceiving their environment and learning

[9] Kirk uses the concept of the "basic package" extensively. It is developed through chapter 6, and put forward on p. 89-96, and throughout the rest. The concepts of various sorts of reflex systems, and deciders, are developed first, through p. 77-89. The discussion of sensation and consciousness is from p. 58-61, as well as p. 92-94.

from their perceptions; similarly, perception is only perception in the sense that the system can do something consciously with the information, instead of having the sensation just cause a reflex.

By Kirk's view, there is a succession of increasingly rich reflex systems. Initially, there is the “pure reflex system,” such as a clam. These are systems with hardwired responses to stimuli, which are genetic (for biological systems) and cannot be altered by the system itself, nor are they designed to be altered by the external world. The “road to the decider” is not simply a matter of increasing complexity—a complex organism like an oyster is just as much a pure reflex system as a protozoon, although biological organisms with greater complexity are generally partially that way to allow for more complex responses. There are then “pure reflex systems with acquired stimuli,” where there is a slight amount of room for new responses to develop, and “built in triggered reflex systems,” wherein certain stimuli open up subsections of the list of responses, which themselves otherwise stay inactive. Finally, just before we cross the threshold into the deciders, are “triggered reflex systems with acquired conditions.” Kirk’s example is the dragonfly, which learns to have a specific nest, but for whom that learning process is automatically set up to happen. That is, the dragonfly does not decide “this is where I’ll set my perch,” but rather certain conditions cause the variable “perch” to get permanently filled in, which then gets plugged into the triggered reflex system.

The threshold between this and the decider is the capacity of “monitoring and controlling the responses,” and is the important part Kirk as emphasizes:

> We have reached a highly significant watershed. For a system to monitor and modify its own behaviour involves a major break with the reflex pattern. Monitoring and modifying must involve not only the organism's being able to perceive its own behaviour, or at least the effects of its behaviour on its environment, but also to adjust its behaviour in ways appropriate to its goals. That requires it to be able to control its own behaviour on the basis of its information, in a way that none of the types of systems so far considered is capable of. [. . .] It seems probable that what we can conveniently refer to as 'monitoring', modifying', and 'controlling' are highly complex processes, capable of being realized to a greater or lesser degree, at different levels of organization in the system as a whole, and in an indefinitely wide range of possible internal structural patterns.[10]

He further says that what is important is the *integration* of all these processes. There is no requirement of how these processes must be executed, just that there are capabilities. To be a decider, to have the "basic package," is for something to be able to—

(i) *Initiate and control* its own behavior on the basis of incoming and retained information: information that it can use;

(ii) *Acquire and retain information* about its environment;

(iii) *Interpret* information;

(iv) *Assess its situation*;

(v) *Choose* between alternative courses of action on the basis

[10] Robert Kirk, *Zombies and Consciousness* (New York: Oxford, 2005), 87.

of retained and incoming information (equivalently, it can *decide* on a particular course of action); and

(vi) *Have goals.*

Moreover, all of these must be unified and integrated.

It's possible that a thing could have faculties similar to some of these, but to have these fully they must be all present and interrelated. Put another way, it makes no sense to talk of "goals" without something being able to acquire and interpret information, or to choose between various actions, nor does it make sense to talk about controlling behavior unless a thing has goals, or interpreting or assessing information unless it's going to be put to a use, to a choice. A thing can "sort of" interpret information, a thermometer for example, but it will not be doing so for itself. This again has a great deal to do with perception, which is just sensation that conveys information, that a decider can then act upon. Sheer sensation, experience, can be found in the simple reflex systems, without there being any understanding or perception, despite there often being some apparently intelligent reaction. This relates back to Kirk's definition of perception—sensation that conveys information, that a decider can then act upon. Sheer sensation, experience, can be found in the simple reflex systems, without there being any understanding or perception, despite there being reaction, often seemingly intelligent reaction.

Bringing this back to the subject of artificial intelligence, what we deal with when we have seemingly intelligent systems is instead this very bare pure reflex system. Kirk will freely admit that he does not know enough of the subject of animal neurology to give clear examples of each sort of reflex system. Similarly, I will say that I am not sufficiently familiar with the programming of AI

to say what sort of system any one example is. However, by my earlier outline of an AI, what we currently have is often still a very simple sort of reflex system. Even the "learning systems" are likely only so-called "triggered reflex systems with acquired conditions," where certain approaches to learning are acquired, but are still within the reflexive framework set up beforehand by the programmer. It is entirely possible, though, that I am wrong here, and that what is causing me to hesitate is something else.

Direct Activity[11]

In Kirk's view, the basic package is not sufficient for phenomenal consciousness. What is also needed is "direct activity," or the direct action of sensation on the creature's decision-making process. We all experience direct activity, when any sort of sensation comes our way, because we cannot help but sense it. Initially it's difficult to even understand what Kirk means by direct activity, because it's unclear what the alternative would be. The simplest example of information gained indirectly is subliminal information—when we do sense something, and file it away somehow, but do not notice it and actually perceive it at the time. The information has been acquired, and can be used to alter our goals or our methods, but in order to do so we must indirectly access them after the fact. Kirk stresses instantaneity and priority in direct activity. The perceptual information is instantly available to an organism, and it also holds priority, immediately changing our goals and choices about the world.

[11] Another important concept, direct activity is detailed in Chapter 9 of *Zombies*, pp. 140-163.

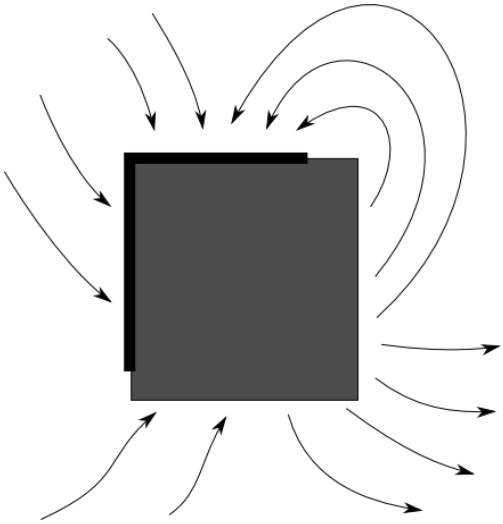
He uses what he calls a “rabbitoid” as an example,[12] stating that a “rabbitoid” is like a rabbit in all ways, except that sensory information does not act on it directly, but through some other method. It is difficult to imagine how this would work, but a possibility would be that the rabbitoid constantly queries its store of knowledge. When a fox comes up from behind a hill, the rabbitoid notices a second later during its regular “scan” of its knowledge base, and then bounds away, relying on its stored model of the environment to navigate. A conscious rabbit has an advantage over a non-conscious rabbitoid in that it will automatically notice changes in its environment, and will be able to alter its immediate goals accordingly, while the rabbitoid would always have some sort of delay in action. The very best that a rabbitoid could do, would be to constantly re-scan its knowledge multiple times per second. This distinction still holds if you assume that rabbits do not possess full rational minds; the reflex system possessed by a rabbitoid would still function better if information about the environment directly affected its system instead of it needing to constantly retrieve stored information about the world.

The Red Herring of Thought

I want to interject a bit about conscious *thought*, and then about bats, before returning to consciousness. Thought is often considered a very important aspect of being human, and seems conflated with consciousness itself. But what happens when we think? One might say, we become aware of what’s going on in our

[12] Ibid. pp. 142.

mind, that we "look inside." I think there's a problem with this, that our everyday sensation of thought and introspection is too naïve, and problematic. Imagine this extremely simplified and kind of silly picture of the mind, gained (of course) from introspection: there is a sort of "black box," in which all mental activity occurs. This box takes certain inputs, many of which are "conscious," although some are not, and produces various outputs. These outputs include motion, activity, and speech, but (here is the point) also include "thought," which is nothing more than aborted speech and self-produced sensation, re-routed back into the box. On only part of the "edge of the box" is the "membrane of consciousness." In terms of this metaphor, things are conscious only as a result of passing through this membrane. Our knowledge of our mental activity is known only so far as we produce thoughts that are then reintroduced into consciousness. The activity within the black box is completely unknowable, and can only be inferred from the thoughts produced.



This is a very flawed picture. Consciousness is not a membrane, there is not a line when things "become conscious" in the brain. However, the salient point is that introspection is not directly accessing or monitoring our mental processes. Instead, thought is output that is reintroduced back into the system. This is

also flawed in that the reintroduction likely does not happen all at the same level—it would be better to imagine the circular arrows happening within the box, making loops of various sizes. But again, the point is that thought is activity re-routed, not the brain actively looking at itself.

This seems like it would be efficient, more than growing some specialized "introspection" capability. If the brain has taken the time to create systems dedicated to processing, say, language, it makes sense that when we think in terms of language we simply route the output of our thoughts, as if we were speaking, back into the language processing bits, using the same hardware we'd use if we were hearing, instead of developing a new system to "monitor" our thoughts. Similarly, at earlier stages, our ability to remember and imagine things significantly overlaps with our capacity to sense things, and so it seems reasonable that instead of developing a new "imagination" capacity, we rather develop the ability to stimulate those systems dedicated to dealing with perception. This also explains the sensation of thought, why it actually has a "sound," instead of just being abstract activity.

So, What is it Like to be a Bat?

I assume a bat has sensation, and also consciousness. What I mean is, there is something it is "like" to be a bat. Perhaps it doesn't have active thought, but it makes decisions, and its actions are complex and nuanced, reasoned. Nagel asked what it is like to be a bat; he, and others, concluded that we cannot know, that the life of a bat is fundamentally alien to us.

But at least attempt to imagine what being a bat is like. The problem, initially, and as Nagel stresses, seems to be echolocation,

something we have no real analogue for. But this doesn't seem entirely impossible, merely very difficult. Try this:

Close your eyes (perhaps read the instructions first, or, imagine closing your eyes). You can still hear, can you not? With your eyes closed, drop something like a book to your desk, and notice how you intuit its position from your impression of the sound. If you were to reach out, you could grasp it with some difficulty. If it were to make noise constantly, you could grasp it with near ease. A sharp sound to your left will give you an impression of "something" there. With your eyes closed, a man walking around a room, or a floor above with a thin ceiling, will give you an impression of motion, of presence. Focus on that impression of presence, separating it from the sensation of the sound itself.

Now, with your eyes closed, feel out your surroundings. You can tell that this is a box, or that is a sphere. You can feel the dimensions of your desk, and you have an almost visual experience of this the size and shape of things. These sensations can be deceptive (how large are your teeth, when sensed with your tongue, and then when felt with your fingers?), but that is not surprising.

Imagine the sensation you experience when a noise is heard, the sense of location and position. Isolate the feeling of position, the feeling of "a presence," from the sensation of the noise itself, that it is a noise. Focus on the feeling of position and presence. Now, imagine the sensation of feeling the shape of an object, and isolate the impression of the form and size from the feeling of touch itself. Merge those feelings of position, as if you were

experiencing a thing's form through a constant torrent of sound, where the sensation of any individual sound was drowned out by the ubiquity of the torrent, leaving only the feeling of position, form, size, and distance. What is there is not the sound, but the almost eerie sense of "something there," the odd itching at your back, like the feeling of being watched without experiencing the watcher.

Pretend you were blind, and had to live off of touch and sound to navigate, but then were able to somehow merge the impression of presence you get from sound, having the sounds themselves fade into the background, and then were to combine this with the feeling of form and shape gained through touch, having the feeling of touch itself be replaced with that background noise, extended to the range of your hearing. You would reach out constantly, as if touching through sound.

Nagel would say that this is what it would be like for a human to be a bat (and even then, only barely), and would press the point, asking what is it like for a *bat* to be a bat. A human has its own beliefs, desires, goals, and so on, and to imagine what a bat's internal life is like is impossible since these will always interfere with our attempts. However, I feel that you can run into the same sorts of problems with asking a question as apparently simple as "what is it like to be yourself?"

First ask, what *was* it like to be yourself? Imagine yourself ten years ago, or even a day ago. How do you do this? Well, you extrapolate. I myself at this moment a day ago was bumbling about, taking a shower, not really interested in anything, assuming I'd wake up a little in an hour or so and figure out what to do then.

Right now I'm coming off of a long, caffeine-fueled writing session. Ten years ago I would be napping on an hour-long bus ride; if I was awake at this moment ten years ago, I'd have been just woken up for some reason. Certainly, both of these involve being fairly groggy and sleepy, and to that extent I can imagine what it would have been like. However, the phenomenal quality of each experience is very different. The sleepiness I feel now is very different from ten years ago, and it is difficult to evoke that feeling in myself because to do so I would need to overwrite my current feeling. I cannot remember what it feels like, I can only extrapolate, evoke the feeling.

But is this only because sleepiness is a muddled, vague feeling? Consider pains. When I was younger, I stubbed my toe. I've done so many times over the years, in fact. And yet can I accurately remember what it felt like? No, only that there was an accompanying feeling of suffering. If anything, what I remember is the suffering, not the pain itself, and even that suffering is extrapolated. How I related to pain then is much different than how I relate to it now. What I feel when I imagine that pain is not what it was like for past me to feel pain, but what it would be like for present me to feel past me's pain, *and only poorly*. How different is this from trying to think what it's like to be a bat? Not impossibly so—and it is not a matter of kind, but of degree. It is much easier to imagine what it was like to be me feeling pain than what it's like to be a bat; but neither is perfect.

What if I asked, what is it like to be you, a second ago? No no no, that's silly, surely. But pinch yourself. Ow. What was it like? Well… it hurt, yes, but can you evoke that sensation again? Not really. You can recall the suffering, and what the pain was sort

of like, and how it still hurts a little now, but none of that is what it was like to be you a second ago, feeling that pain. So what is it like to be you, right now? Anytime you try to focus on that, you can only evoke the feelings a second later. What it is like to be you is constantly slipping away. You can only experience "what it's like" to be anything, namely you, as it is experienced. To actually feel what it's like, you need to have the feeling at the moment. This also somewhat makes sense evolutionarily—why would we go through incredible effort and cost to repeat pleasurable actions if we could merely evoke the pleasure in our minds on command?

That we cannot imagine what it is like to be a bat is not so surprising when we cannot even imagine what it is like to be ourselves. And yet this does not tell us that there was *nothing* that it was like to be ourselves, and it does not tell us that there is *nothing* that it is like to be a bat, and this says nothing about telling whether there is a "what it's like" for any entity through physical observation.[13] Our memories are not just "not as vivid," but entirely false, constructed. We cannot know the past "what it's like," or others' "what it's like"-s, just as we cannot know the "what it's like" for a bat—but we would not deny consciousness to our past selves, or to other people.

What is it Like to be a Thermostat?

To ascribe emotions, desires, or beliefs to a thermostat is silly. When I say that a thermostat has experience or sensation, I do not mean anything approaching our own experience. As Kirk

[13] A significant amount of this is paraphrased from Kirk (2005), ch. 5, especially p. 61-68.

would say, we are deciders, we interpret information and make decisions based on that information according to goals. A thermostat whose setting aligns with the ambient temperature does not "feel content." A thermostat set to a higher temperature does not "desire to make things hotter." A thermostat does not "believe maintaining the temperature is good." A thermostat does not *perceive*, because it does not interpret its sensations, does not work with information. Emotions, desires, and beliefs are fantastically complex and important aspects of our experience. Some day we will make a machine that does experience emotion and desire, and have beliefs, but it will be no time soon[14]. This is a reasonable and worthy goal, but it is important to realize how difficult it will be. So if we cannot say that a thermostat "desires to make things hotter," in what sense does it have an internal experience?

When I ask "what is it like to be a thermostat," I'm speaking of something that it is very, very difficult to imagine. It is hard enough to imagine what being a dog is like; harder still to imagine the life of a slug, and of a bacterium; so when we get down to something as bare and simple as a thermostat, we are truly a long ways away from our own experience. It is not even enough to try to sense things thoughtlessly, as the sensation of a thermostat is nothing like ours in any way. A thermostat is simpler even than an individual neuron.

All I mean is that the thermostat "senses." It senses the temperature the same way a protozoon senses light levels and moves accordingly, or the same way a bacterium senses a certain

[14] No, I don't have support for this, but I consider it the same sort of statement as "someday we will colonize other planets." Barring something horrible happening, or the discovery of some extreme limiting factor, it seems so.

chemical in its environment and stops dividing[15]—really it must be far more simple than that, but it is the same sort of "basic reflex" as a clam closing its shell in response to certain stimuli, or a slug retracting a feeler when it touches something rough. In the same way that animals have moved from those basic reactions to our own, we should attempt a similar project to move from the thermostat (well maybe something else) to a conscious being like us. What this requires is a move from the "reflex system with acquired conditions" to the actual "decider."

Conclusion: Does "Consciousness" Matter When Thinking of Artificial Intelligence?

It depends on what the connotations of "consciousness" are, which brings us back to Kirk's direct activity. If the difference is between having direct activity or not, between being a rabbit or a rabbitoid, it seems in fact that consciousness is of no importance, and the focus on "what it's like" is missing the point. If, instead, "consciousness" is taken to deal with the difference between sensation and perception, between acting on reflex, or making judgments, having goals, and so on, then it is obviously of high value. A system that can actually analyze the world and make judgments will have an advantage over something that acts on predefined rules, assuming it is meant to deal with the sorts of

[15] Certain protozoa sense light, and then move their flagella to move toward it, but only when it is fairly mild; bright, constant light has no effect. Colonies of certain bacteria maintain a size by having each bacterium secrete a chemical, and then stop division when the chemical reaches a certain concentration, which lines up with a certain population.

complex and variable situations that humans and other animals can handle.

In other words, we should not be asking whether computers will be conscious; that is a matter of how they relate to information. What matters is how they process it, and the incidental aspects of consciousness (the instantaneity, the priority) should not be taken as essential to having a mind.

Bibliography

Churchland, Patricia S., Sejnowski, Terrence J. "Neural Representation and Neural Computation." *Mind and Cognition*, ed. W. G. Lycan (1990, 1994). Cambridge:Blackwell. pp. 224-252. (From *Neural Connections, Mental Computations*, ed L. Nadel et all, MIT Press (1989))

Jackson, Frank. "What Mary didn't know." *Journal of Philosophy* (83): 291–295. 1986.

Kirk, Robert. *Zombies and Consciousness.* New York: Oxford, 2005.

Lycan, William G. *Consciousness and Experience.* Cambridge: MIT, 1996.

Nagel, Thomas. 1974. "What Is It Like to Be a Bat?" *The Philosophical Review.* 83 (4): 435-450.

Searle, John R. "Can Computers Think." *Philosophy of Mind.* ed. D. Chalmers (2002) pp. 669-675. New York: Oxford, 1983.

Searle, John R. *The Rediscovery of the Mind.* Cambridge: MIT, 1992.

Smart, J. J. C. "The Identity Theory of Mind." *The Stanford Encyclopedia of Philosophy*, 2008.ed. Edward N. Zalta. http://plato.stanford.edu/archives/fall2008/entries/mind-identity/