**-EDITORIAL-**

# Banbury Bound, or Can a machine be conscious?

Eric Dietrich
Philosophy Dept.
Binghamton University
Binghamton, NY 13902-6000

In mid-May of 2001, I attended a fascinating workshop at Cold Spring Harbor Labs. The conference was held at the lab's Banbury Center, an elegant mansion and its beautiful surrounding estate, located on Banbury Lane, in the outskirts of Lloyd Harbor, overlooking the north shore of Long Island in New York. The estate was formerly owned by Charles Sammis Robertson. In 1976, Robertson donated his estate, and an endowment for its upkeep, to the Lab. The donation included the Robertson's mansion, now called Robertson House, and a large, seven-car garage that would become the actual conference center. The Center was opened on Sunday, June 14, 1977, by Francis Crick who gave a talk entitled "How Scientists Work." For us, Banbury was an idyllic location with great food where we could talk about the most difficult problem in all of science: *what is the nature and cause of consciousness?*

It *is* the most difficult problem in all of science; O yes it is. No one even knows what a solution would look like. And the proposed solutions we have today are so diverse that no underlying similarity unites them – save for the fact that most (yet not all) have something to do with the brain. But at Banbury, we were a plucky lot – we few, we happy few, we merry band of cognitive scientists, brave in the face of certain defeat. There were four philosophers (rather a lot, really), about ten or eleven neuroscientists and psychologists (depending on how you counted), and six or seven AI scientists, computer scientists, and engineers (again, depending on how you counted). Every single paper discussed some aspect of consciousness and its possible physical realization in

either a brain of some mammal or other or a machine – except my paper, of course, (co-written with Valerie Hardcastle).

The papers were packed with terrifically interesting information, suggestions, and insights (the papers, by the way are not currently available, but there is some talk of putting them on the web). The questions and discussions were enlightening, clever, and challenging. But at the end of each day, we were all left with the following nagging problem: for any suggested physical realization of consciousness, P, it was always possible to imagine P occurring in some brain or computer without consciousness thereby occurring – a point noted by several attendees. This dissociation either means that we never found the right P, or that we did find it but can never see it *as* the right P (the latter being my considered view, as I have mentioned previously in this forum, see Dietrich, 1999). Indeed, computers were the spoilers here, for they were great vehicles for exercising our imaginations: For every P, it never seemed as if implementing P in some machine would thereby cause the machine to be conscious. We could always say: "But if a computer did P, it doesn't seem as if it would thereby be conscious."

This dissociation between our felt experiences and the way we believe the world has to be is ubiquitous: all humans feel (perhaps with a little coaching) that their conscious experiences could be just what they are regardless of how the world is, that somehow our consciousnesses need not cohere with how the physical world actually is. And no matter how much faith we put in materialism (the view that consciousness just is a material property of the world, like mass), the intuition persists; even the most ardent materialists have this intuition; for example, Dennett admits to having it (he regards it as an illusion, of course). This intuition, that our experiences need not cohere with the world we live in, is called our *Cartesian intuition* (after the famous seventeenth century French philosopher, Descartes, who first articulated the modern view of dualism).

The Cartesian intuition is easy to conjure up. At one time or another, we have all dreamed that we are somewhere strange or have vividly imagined that we are doing something exciting or novel, yet we are not where we dreamed nor doing what we imagined. If we can dream we are hang gliding in Tibet when we are home in our beds,

then perhaps we are just dreaming we are at home in our beds. Perhaps everything is a dream. Perhaps boarding a jet, flying to Nepal, trekking in to Tibet, and hang gliding is a dream. We all have at one time or another thought something like, "What if none of my experiences are real? What if my entire experience of the world is one big dream? What if I am a brain in a vat somewhere, bathed in wet, warm nutrients merely dreaming that I am reading this essay? Or what if I am a body on life support somewhere, having philosophy-essay-in-an-AI-journal experiences? What if nothing is the way it appears to me; what if nothing is the way I experience it?" It's a scary thought. But it could be true, it seems. And the fact that it could be true is what scotches all attempts to reductively explain consciousness in material terms. In short, for some reason, humans are endowed with an intuition, our Cartesian intuition, that renders all physical, material explanations of consciousness completely unconvincing (actually, I know what the reason is: we have the Cartesian intuition precisely because we are conscious! – again, see my 1999 editorial).

Well, what were some of the suggested causes of or functions of consciousness -- some of the P's mentioned above? There were the usual suspects: the ability to generate language, our ability to integrate complex information into a coherent whole, intentional control of our body, awareness and attention, reportability of our internal states, having an identifiable self, performing nonstereotyped, nonlearned behavior, recursively thinking about our own thoughts, and so forth. These suspects fall into two categories: 1) they easily succumb to the Cartesian intuition, hence one can imagine a computer doing them yet not being conscious (e.g., language generation, information integration), or 2) they are as mysterious as consciousness itself (selfhood, awareness, and attention fall into this category). So, in truth, this list is merely "processes associated with consciousness."

The list of insights about consciousness from the conference was long. It was inspirational to see the power of the various cognitive and neural sciences brought to bear on the problem of consciousness – we humans know a lot interesting facts about the brain and its mind. One of the most interesting was a comment on Demetri Psaltis's paper. Psaltis outlined some results obtained by his graduate student, Greg Billock, and

him. Billock has a program that tests various strategies to play a "ground-acquisition" game called "desert survival" or "the camel game." Of course, it's a complicated game with lots of parameters and different strategies. The object is to use the camels in one's herd to capture oases in a desert. The player with the most oases at the end, wins.

A very interesting and successful strategy, called the savvy player strategy, relies on an "awareness window." The board is broken up into a set of neighborhoods around the oases; the size of the neighborhoods can be varied. The neighborhoods are then used to construct the awareness window. This window allows the program to concentrate its computational resources in a small area of the game board which is of crucial interest. Thus focused, the program computes values of parameters and plans camel deployment in ways that would be too expensive if applied to the whole game board. The trade off is that if the awareness window is too small, the algorithm will run quickly but won't have much to work with (most of the board as been excluded), hence will lose often. On the other hand, if the window is too large, then the expense of the computations will start to matter (there will be too much information), hence the program will run too slow, missing its turns and giving outdated information to its camels. It turns out, there is an optimum in the size of the awareness window – sort of a Goldilocks phenomenon: neither too large, nor too small, but just right.

Now, the "awareness window" is just a name, right? Computer scientists, and especially AI researchers, do that sort of thing all the time: it's called "hopeful naming." For example, they name their data structures "knowledge representations," when such structures neither represent, nor store knowledge (not knowledge represented to the machine itself). Or consider the famous General Problem Solver, which was not general at all (means-end analysis was supposed to be general, but it wasn't for interesting theoretical reasons having to do with concepts). The list is long. My point is that naming a computational-resource-focusing-window the "awareness" window doesn't give us conscious awareness (aka consciousness), does it? Well . . . mumble. *Something* has to give us conscious awareness. And, because of the Cartesian intuition, whatever that something is will *never* seem like the material basis of consciousness.

It was at this point that Andy Clark asked Psaltis: "Is your computer conscious because it uses the awareness window?" To which Psaltis replied: "No, not yet; integration is needed." Good answer. But what else could he say: "Yep, that baby is conscious all right! We're legally bound not to turn it off"? That would have flown like a lead balloon.

But then Christof Koch commented that perhaps the first conscious computer won't be dramatic or dramatically different from its predecessors, just like the first conscious animal wasn't dramatically different from its non-conscious predecessors. So perhaps Billock's and Psaltis's computer *is* conscious. There was a moment of silence.

The problem with this comment, though, is that while the intelligent behavior of a conscious computer may not be dramatically different from that of its predecessors, its being conscious *will be* dramatically different. Why? Because consciousness is a binary affair. It is all or nothing. You are either conscious or you are not. If you look at an apple and don't experience red, but gray, say, then you are still conscious of perceiving gray. Being not conscious means not experiencing anything at all, it doesn't mean only experiencing a few things or only experiencing things blandly or incoherently. For, few, bland, or incoherent, they are still experiences. So the first conscious computer will differ from its predecessors in something like the way humans differ from rocks. Of course we will almost certainly not notice the difference, and more's the pity.

So, can a computer be conscious? You bet. Are there conscious computers now? Who knows? Is the internet conscious? Maybe, maybe not? Should we care whether there are conscious computers now? Absolutely? How would we test for one? We can't, yet. A conscious computer, when we succeed in building one, will almost certainly help us develop a theory of consciousness, but it will be a thin theory: it will not help us deeply and satisfyingly understand consciousness at all. This is because we *are* conscious. Consciousness itself causes our Cartesian intuition, and this intuition blocks any explanation of this most mysterious phenomenon from being satisfying, compelling. So then what do we do next? Continue doing what we are doing – a boring

answer, I know, but given our ignorance, what choice do we have? And while we are continuing on, we can always hope.

Dietrich, E. (1999). Fodor's gloom, or What does it mean that dualism seems true?" *J. of Exper. and Theor. AI,* 1999, 11 (2), 145-152.