Crude, Cheesy, Second-Rate Consciousness

Joanna J. Bryson

September 10, 2008

Abstract

If we aren't sure what consciousness is, how can we be sure we haven't already built it? In this statement I will speak from the perspective of someone who routinely builds small-scale machine intelligence. I begin by discussing the difficulty in finding the functional utility for a convincing analog of consciousness when considering the capabilities of modern computational systems. I then move to considering several animal models for consciousness, or at least for behaviors humans report as conscious. I use these to propose a clean and simple definition of consciousness, and use this to suggest which existing artificial intelligent systems we might call conscious.

"If the best the roboticists can hope for is the creation of some crude, cheesy, second-rate, artificial consciousness, they still win." — Daniel Dennett (1994), *The Practical Requirements for Making a Conscious Robot*

While leading a group building a humanoid robot in the 1990s, Rodney Brooks complained about the term *robot brain* (Brooks and Stein, 1994). You can have a robot hand or arm or eye or even face. But as soon as you say you have a robot brain people say "That's not a brain." The aim of this article is to make you look at some artificially-intelligent systems and say "You know, maybe that *is* robot consciousness," about something that already exists.

From experience, I know this is hard to do. I remember sitting in a Cambridge, Massachusetts diner with other postdocs after Dennett had just given a seminar. The other postdocs asserted science would solve consciousness, but not in their lifetimes, not in a hundred years. While trying to understand why they could be so certain of this, I challenged them about how a computer could prove it is conscious. Almost anyone who owns a computer can make it say "I am conscious." As for generating human empathy, teddy bears and pet rocks do this with no intelligence at all. The postdocs said that consciousness was a special sort of self-knowledge, being aware of what you are thinking. But computer programs have perfect access to all their internal states. If you set up a program correctly, you can ask it exactly what line of code — what instruction — it is executing at any time, and precisely what values are in its memory. This is in fact the job of a debugging tool, such as an Interactive Development Environment (IDE) — a common type of program which is not even rountinely considered AI.

So if consciousness is just perfect memory and recall, then video recorders have it. If it also requires access to process as well as memory, then computers have that access. Maybe some people are committed enough to these definitions that they are already convinced computers are conscious. But most of us find this idea unsatisfactory.

Multiple Drafts and Concurrency

Dennett talks in his multiple drafts theory about the fact that brains have many things going on in them at one time (Dennett, 1991; Dennett and Kinsbourne, 1992). In his model, consciousness is a spotlight that shines on no more than one of these things at a time, at least it only shines brightly on one. But why is the brain doing so many things at once? The reason is because if many processors run at the same time, more can get done quickly. In computer science, this is called *concurrency*.

Concurrency is a great strategy for problems that can be taken apart into pieces. But the "hard problem" in concurrency comes if you need to combine all or even some of the answers you find back together again. This can be called the problem of *coordination*. For an example, think of bees. A colony of bees can explore a large space around their hive to find flowers by having each bee fly in a random direction. They will explore even more space by using simple distributed rules each bee can know, like "don't fly near another bee". But how does it help the colony if only one bee finds some really good flowers? When the bees communicate by the waggle dance, a lot of bees have to stop what they are doing to be involved, and one bee has to spend a *lot* of time and energy dancing (von Frisch, 1967). When you think about the complexity of this behavior and the time it took to evolve you realize it must be a huge advantage for the bees overall. They sacrifice this time as individuals, and on average each individual then has a better chance of finding food and bringing it home.

To return to the main topic, the suggestion I am making is that self awareness isn't enough for consciousness unless there is a significant portion of the self of which one might *not* be aware. Or put another way, there must be some process that acts as a "bottleneck" or constraint, a limit that makes some sub-part of the whole special. In the bee case, that limiting process is the communication to others when a really good source of food has been found by one bee — the recruitment of others to a particular location.

Some approaches to artificial intelligence similarly have concurrent processes which normally operate more or less independently. In AI as in some other disciplines such as Psychology or EvoDevo, this decomposition of the whole into some specialized subparts is called *modularity* (Bryson, 2000, 2005). Just as in Psychology and EvoDevo, the utility of modularity in AI is that more complicated systems can be developed more simply and operate more quickly. The problem of coordination in AI is called *action selection*. This problem emerges whenever multiple modules are contending for a single resource (Blumberg, 1996). An example of a "resource" in this sense can be as simple as physical location. I cannot stand and give a talk at a meeting at the same time as I enjoy myself in a café, so if I want to do both I have to find some sequential ordering for my actions. Another such resource is speech — we can only say one word at a time, so words must be sequenced. And, critically for Dennett's description here of consciousness, memory. Apparently, episodic memory is a constrained resource, and only some of the things we are thinking about or perceiving will wind up in it.

A Functionalist Hypothesis of Consciousness

Dennett in his statement here said the only common characteristic of conscious contents is "the historical property of having won a temporally local competition with sufficient

decisiveness to linger long enough to enable recollection at some later time". But the question of course is, competition for what? As Dennett points out later and I will return to in my conclusion, one element for humans is *public expression*. If your current thoughts made it so far as to become verbalized, they are now a part of the public awareness. In this case the "local competition" is not only internal but also external — with other speakers. The memory is not only your own but also that of any other hearers. But for most of my statement I will focus on individual consciousness. Here it may be a little harder to see why we are only conscious of one thing at a time. Some researchers have suggested that the reason for this individualization is that consciousness is related to another sequencing problem — action selection, which I mentioned before. Norman and Shallice (1986) suggest consciousness is a set of extra or special resources which are brought to the problem of sequencing behavior when the brain is either uncertain about the correct sequence (as in a new context or when working on a new task) or when such sequencing is particularly important (as in when performing a delicate operation.)

Norman and Shallice (and others) have always been a little fuzzy about what the special resources consciousness brings to such difficult situations might be. I am going to make a specific proposal here, but I won't entirely justify it until later in my essay. But my proposal is simple — I think consciousness and episodic memory are the parts of a process for adaptable action selection. This process consists of:

1. fixing an aspect of a behavior context in the brain and
2. allowing the brain to search for potential actions that might be best suited to this context.

This sort of action selection is exceptional — most aspects of behavior are predicted directly by their context and do not need such a process of search. However, because human behavior is unusually plastic, we spend quite a lot of our time doing this sort of thing, even when the next action is not particularly difficult or pressing. Either as a consequence of this or as a consequence of our language and culture, we can even use it to reason at various levels of abstraction. So we might think about an essay we are writing when driving home when the road itself does not demand much attention.

I got this model of interacting attention and action from vision researchers, Wolfe et al. (2000). The main point of that paper is that when performing a new task, one doesn't learn from that performance when one can use vision rather than memory to guide the behavior. But my hypothesis depends more on another model they describe. This model accounts for the difference in the time it takes to find some visual stimuli compared to others. Studies like these that measure the time for processing are called *reaction time* (RT) studies. In vision, if you have a field of dots where some are red and one is blue, you will find the blue one very quickly, and your RT will not depend on how many red dots there are. Similarly, if there are a number of **T**s on a screen and one **L**, you will not have trouble finding the one L, and you will find it quickly no matter how many Ts there are. *However*, if the screen has many Ts and many Ls, and Ts are both red and blue, but only one L is blue, it will take you longer to find the one blue L. And, the more other distracting objects (red Ls, blue Ts) there are, the longer it will take you to find the blue L.

Why is this? Vision researchers have long agreed part of the answer is because finding a blue object or a particular shape are both things problems that your eyes' concurrency

system can handle more or less by itself. The different cells in your early visual processing can identify whether they have a blue section or a T shape easily, and alert whatever motor system needs to signal this quickly. But apparently identifying that something is both blue *and* a T cannot be done this way. Wolfe and his colleagues proposed a relatively simple explanation for what happens in this case. One just randomly looks at items with one trait and checks if they also have the other trait, until one happens to look at the right one. So for example, you might just look at anything blue in the field (perhaps returning multiple times to some objects) and eventually you will either see that one is also a T or give up. Thus the process of recognizing and visually targeting blueness or T-ness is not very conscious, but the process of saying "is that *both* blue *and* a T" apparently must be.

To try to convince you of my definition of consciousness, I will now work through two more experimental psychology examples. Then I will return to the question of conscious machines. Both of my examples concern something Dennett describes here as "imponderable" — consciousness in non-human species.

Animal Models of Consciousness

I first really studied animal consciousness when a colleague made passing reference to declarative memory in a rat. Whether or not rats are aware, I'm quite certain they don't declare anything, but it turns out there is reasonably good evidence rats have explicit episodic memory. We know this because of their behavior, and because of its analogies in humans, whom we can ask about their conscious experience. In this case, the person who was being asked is patient HM. HM had both of his hippocampuses removed to treat his severe epilepsy, and as a result lost the ability to form new episodic memories. When I was a psychology undergraduate in the 1980s, we were taught that he had lost the ability to *consolidate* short-term memories into long-term memories, but it turns out that this was wrong. Also, we were told that when rats had their hippocampuses removed they could still consolidate their memory, but they had certain problems with navigation, so apparently hippocampuses were for navigation in rats but memory consolidation in humans. It turned out this was wrong too — the real answer is both more parsimonious and more interesting.

What HM can't consolidate is that he can't remember an episode after that episode finishes. So if you distract him by teaching him a new task, he can't remember when he met you afterwards. But although he had his surgery in the 1950s, he started acquiring semantic knowledge about John F. Kennedy and rock music. One day, someone thought of giving HM the sort of task the lesioned rats were actually learning. So they brought in an apparatus and said "when that light goes on, push that button", and when he did they gave him a penny. After he'd done this for some time, they distracted him by asking him to count his pennies. After this he said he didn't know what the apparatus was for. But when the light went on, he pushed the button. When they asked him why he did that, he said "I don't know."

So what about the rats? One of the "navigational" tasks the rats had problems with was the radial arm maze — a maze with eight arms coming out from a center. The trick with this maze is to remember which three arms the scientists put food in, and to go to each of them and not the others because you only have a little time in the maze. Also, you

can't go to the three arms in a particular order, because little doors slide up and down to prevent that. You have to remember which of the three arms you've already been down today to make sure you go to each of them once. When the rats had no hippocampuses, they could still learn which three arms had the food, just like HM could tell you about the Beatles. But on any particular day, they didn't efficiently go down those three arms once each, like a normal rat would. Rather, they acted like they couldn't remember what they'd just been doing, just like HM. This is what my colleague had referred to as "declarative memory". The ordinary rats (the ones that still had their hippocampuses) were showing they had it by going down the three arms efficiently. For details and full referencing of the above experiments, see Carlson (2000). But the main point here for my argument, is that rats seem to have a special memory like humans, and like humans they lose that memory if they lose their hippocampuses.

So from this I hope we can accept that animals as much like us as rats have at least part of what we normally think of as *consciousness*, and that they use it for remembering things and choosing their actions. I will now move on to the third experimental psychology story. This one doesn't involve surgery, just getting older. One of the standard tasks studied in animal learning and reasoning is called *transitive inference*. You may remember this from math — if $A > B$ and $B > C$, then $A > C$. Now it turns out that for animals the $A > C$ part is easy — *if* they can learn the two premises. It turns out to be very, very hard to learn two different things about B, and it takes a lot of training to get them to learn the original pairs. Now another important part of this story is the reaction time. Say an animal (including a human) has learned a bunch of pairs making a big sequence: $A > B; B > C; C > D; D > E; E > F$. One characteristic of transitive reasoning in animals is that the further apart two stimuli are from each other in that chain, the *faster* the animal is at making their choice. This is called the Symbolic Distance Effect (SDE). So due to the SDE, the reaction time for answering $B?E$ is shorter than for answering $B?D$.

As I said earlier, reaction times are normally associated with cognition. So historically, people have been trying to discover what computation animals might be performing that does transitive inference and goes faster as a chain gets longer (Bryant and Trabasso, 1971; Shultz and Vogel, 2004). But what I think is really going on is that animals are conscious and thinking during the SDE delay. The more uncertain they are about the next action, the longer they hesitate, so their brain can search for a better, more certain solution, using a process like I described above for vision. I think this for two reasons. One is that I have spent some time researching mistakes children and monkeys make in performing transitive inference, and wound up supporting a model of the underlying process that explains everything *except* the SDE, so I (and some other people) think the SDE is not dependent on the transitive reasoning (Bryson and Leong, 2007; McGonigle and Chalmers, 1992). But the second reason is simpler — the SDE can go away and the animals still perform transitive inference. (Rapp et al., 1996) have shown that elderly rhesus macaques perform transitive inference more quickly than their juniors and just as accurately. However, they have no SDE. Also, they don't notice if the rewards change on one of the pairs. Because of an error in their experimental design, Rapp and his colleagues started rewarding all their monkeys on the pair $B?D$ at chance, so most of the monkeys stopped performing $B > D$ and rather went to chance on choosing B or D. But the old ones, who hadn't been hesitating, also didn't notice the change in reward and kept choosing B.

This is just one experiment and there's clearly a lot more work to be done. But I put forward as a hypothesis that the older lab monkeys are more likely to go into "auto-pilot" mode on a simple lab task. This could be adaptive for them, since if they'd lived that long in the wild they'd probably already know how to perform most tasks, and they might be losing scramble competitions (the way rhesus macaques forage) to younger, more agile monkeys in their troop. So learning is probably less important than speed for them. Of course, we can't be sure that they are performing their transitive inference decisions on auto pilot, because we can't ask them directly about their memory. But hopefully we will find a way to extend this research into human subjects.

Do We Have Conscious Machines Yet?

For now though I will return to the question of whether we have already achieved machine consciousness. Maybe not the full rich human pageantry of narrative with qualia, meta-reasoning and everything, but perhaps what Dennett has called "crude, cheesy, second-rate artificial consciousness" (Dennett, 1994, p. 137). What I have proposed here is that calling something "conscious" requires several things:

1. There must be multiple, concurrent candidate processes for conscious attention.
2. There must be some special process applied to a selected one of these processes.
3. This special process must achieve some function, probably concerning sequencing actions. And,
4. as a side effect, the object of this attention will normally be recorded in episodic memory, at least for a while.

Do any machines meet these criteria? I think probably yes. As pathetic as they are compared to humans or our science fiction, I think many of the humanoid robot systems which engage in dialog with human users and attempt to select objects from table tops can probably be thought of as meeting all these criteria in a crude, cheesy sort of way. Such robots are at MIT, Georgia Tech and the University of Birmingham, to name just a few (Roy and Pentland, 2002; Breazeal et al., 2006; Hawes et al., 2007).

If you think on a larger, Chinese-room sort of scale for a cognitive system, we might also see AI playing a part in other kinds of consciousness. For example, the Internet employs massive concurrency to create a world-wide database of useful information. If someone wants to act on a piece of that information, they employ a search engine to limit their view of all that data to say ten URLs with context on a single web-page. Under the definition of consciousness above, a page enters consciousness of the system as a whole at the same time it enters the consciousness of the human being who is doing the final selection of the page to be viewed. Reference to the selected Internet item and some summary details about it go into the episodic memory of the human, their browser and generally also their chosen search-engine provider (e.g. Google.) The browser will use this memory to suggest that page to the person again; the search company will use this memory to make it more likely this page is shown to other people who search, and the human will use the information for whatever they originally intended (or possibly something else). Thus in

a way a single consciousness is used concurrently by three different cognitive systems. And I think the two forms of consciousness that have AI elements are not too unlike what Dennett referred to here as "the publication competence", the making public of conscious information which he describes the final arbiter of what for a human is conscious.

Conclusion

As Dennett has said elsewhere, part of the reason we have trouble understanding consciousness is because the term has origins in folk-psychology and probably covers a large range of phenomena. What I have done here is concentrate on two criteria for consciousness I think Dennett has made very clear: that it is something that happens to one candidate process among many, and that it creates a lasting impression in something like episodic memory. From this I have proposed that consciousness is part of a particular sort of action selection — a sort that is triggered by uncertainty and allows for exploration and learning of new actions in a particular context. This is in contrast to the majority of action selection, which is more-or-less reducible to stimulus-response, possibly also with some automated arbitration (Prescott, 2007). Finally, I argued from my definition that we can find evidence of consciousness not only in animals but also in *existing* AI systems.

None of my arguments are meant to belittle consciousness in any way, although obviously as a functionalist I am happy if they help demystify it. I am not claiming consciousness is emergent or epiphenomenal. Rather, consciousness is a central process to the part of intelligent behavior I am most happy to call "cognitive". Explaining how something works is by no means the same as explaining it away. Even the crude, cheesy, second-rate artificial consciousnesses I describe are not I think belittled by the description — anything but. Hopefully with a more informed perspective on the situation, we will begin building more useful — and more conscious — cognitive systems.

References

Blumberg, B. M. (1996). *Old Tricks, New Dogs: Ethology and Interactive Creatures*. PhD thesis, MIT. Media Laboratory, Learning and Common Sense Section.

Breazeal, C., Berlin, M., Brooks, A., Gray, J., and Thomaz, A. L. (2006). Using perspective taking to learn from ambiguous demonstrations. *Robotics and Autonomous Systems*, 54(5):385–393.

Brooks, R. A. and Stein, L. A. (1994). Building brains for bodies. *Autonomous Robots*, 1(1):7–25.

Bryant, P. E. and Trabasso, T. (1971). Transitive inferences and memory in young children. *Nature*, 232:456–458.

Bryson, J. J. (2000). Cross-paradigm analysis of autonomous agent architecture. *Journal of Experimental and Theoretical Artificial Intelligence*, 12(2):165–190.

Bryson, J. J. (2005). Modular representations of cognitive phenomena in AI, psychology and neuroscience. In Davis, D. N., editor, *Visions of Mind: Architectures for Cognition and Affect*, pages 66–89. Idea Group.

Bryson, J. J. and Leong, J. C. S. (2007). Primate errors in transitive 'inference': A two-tier learning model. *Animal Cognition*, 10(1):1–15.

Carlson, N. R. (2000). *Physiology of Behavior*. Allyn and Bacon, Boston.

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

Dennett, D. C. (1994). The practical requirements for making a conscious robot. *Philosophical Transactions: Physical Sciences and Engineering*, 349(1689):133–146.

Dennett, D. C. and Kinsbourne, M. (1992). Time and the observer: The where and when of consciousness in the brain. *Brain and Behavioral Sciences*, 15:183–247.

Hawes, N., Sloman, A., Wyatt, J., Zillich, M., Jacobsson, H., Kruijff, G.-J., Brenner, M., Berginc, G., and Skočaj, D. (2007). Towards an integrated robot with multiple cognitive functions. In *Proceedings of the Twenty-Second Conference on Artificial Intelligence (AAAI-07)*, pages 1548–1553.

McGonigle, B. O. and Chalmers, M. (1992). Monkeys are rational! *The Quarterly Journal of Experimental Psychology*, 45B(3):189–228.

Norman, D. A. and Shallice, T. (1986). Attention to action: Willed and automatic control of behavior. In Davidson, R., Schwartz, G., and Shapiro, D., editors, *Consciousness and Self Regulation: Advances in Research and Theory*, volume 4, pages 1–18. Plenum, New York.

Prescott, T. J. (2007). Forced moves or good tricks in design space? landmarks in the evolution of neural mechanisms for action selection. *Adaptive Behavior*, 15(1):9–31.

Rapp, P. R., Kansky, M. T., and Eichenbaum, H. (1996). Learning and memory for hierarchical relationships in the monkey: Effects of aging. *Behavioral Neuroscience*, 110(5):887–897.

Roy, D. K. and Pentland, A. P. (2002). Learning words from sights and sounds: A computational model. *Cognitive Science*, 26(1):113–146.

Shultz, T. R. and Vogel, A. (2004). A connectionist model of the development of transitivity. In *The* 26^{th} *Annual Meeting of the Cognitive Science Society (CogSci 2004)*, pages 1243–1248, Chicago. Lawrence Erlbaum Associates.

von Frisch, K. (1967). *The Dance Language and Orientation of Bees*. Harvard University Press, Cambridge, MA.

Wolfe, J. M., Klempen, N., and Dahlen, K. (2000). Postattentive vision. *The Journal of Experimental Psychology: Human Perception and Performance*, 26(2):293–716.