

A response to the question: “Can computers think? Can they feel?”

by Josh Bacigalupi

This paper is a response to the above question as posed by *ask the expert*, a Stanford Engineering web column, April 2009 (Stanford 2009), and the ensuing thread that began among peninsula AI Meetup group members.

Firstly, I'd like to thank Ben Goertzel for his well considered essay on emotion (Goertzel 2004) and to also thank Dorian Pyle for posting the New Atlantis article by Ari N. Schulman (Schulman 2009). Both articles were informative and better researched than anything posted on this thread so far. Our discussion would be served by articulately agreeing or disagreeing with their points.

“Can computers think? Can they feel?”:

For myself it is not a question of whether a computer can or “can't [do] X...”. I do not claim that discrete binary state manipulators (i.e. computers) will never think or feel. But I do assert that the current implementation of the Church-Turing thesis do not think or feel as humans do. There are certain tasks humans do better than computers and vice versa. And if our goal is in any way to make a machine that approximates our skills or for our brains to approximate a computer's skills, we have work to do. Personally, I believe that a hybrid of the two is the ultimate endgame that will elicit the greatest benefit to the larger system in efficiency, adaptability and stability. But to even consider our thesis question above (thanks Kennita!), let alone the endgame I advocate, we must strive to understand our only proof of concept for thinking and feeling, namely, the mind; which is implemented, to the best of our knowledge, via an embodied brain.

The most efficient way to move forward is to employ the following design principle: don't reinvent the wheel. So, to the best of my ability, I'll summarize my personal take-away points from Goertzel and Schulman's writings and add a few references from some other respected references, namely: John Searle, Anthony Bell (Research Scientist at the Redwood Center for Theoretical Neuroscience, UCB) and Dharmendra S. Modha (Manager, Cognitive Computing, IBM Almaden Research Center).

My main thesis, since I am a Monist at heart, is that mind – and everything we roughly associate with it – can eventually be executed by some machine fabricated by our collective human minds. Importantly, however, the nature of this machine may depart markedly from what we now un-rigorously refer to as a computer and what I'll specify as a discrete binary state manipulator based on the Church-Turing thesis and currently implemented via multiple transistors, which are functionally grouped into logic circuits so as to perform logical operations on discrete inputs to elicit a priori mapped outputs (e.g. logic tables). My point in being this verbose is to illustrate the fact that computers, as defined and implemented, were designed to perform specific kinds of tasks that our minds were routinely inefficient at, namely, recursion; or the reduction of a larger problem into smaller similar problems that can be solved repetitively via automation.

In essence, our minds created a machine to do what our minds did not do efficiently. This machine's form in theory and practice is highly ordered to perform a particular task; as such, it is improbable (via second law of thermodynamics $\text{entropy}(S) = k \ln W$) that some other specific

configuration exists to perform the exact same task by accident. Furthermore, our mind has evolved to be equally, if not vastly more, complex than the computer it created. Entropy is essentially a statistical model. So, it follows that as a form becomes more complex to complete some function it becomes increasingly improbable that some other form – evolved to complete a *different* task – can be made to efficiently perform the initial form’s task. In other words, human’s evolved to perform certain things and computers were designed to do some other things that the human’s original form was inefficient at. As such, it is unlikely that 1) computers are of the same underlying form as the mind, and 2) that these computers can, in turn, be made to efficiently do what humans have been evolved to do for eons.

Schulman’s article, though it takes a while to get there, makes some similar points that, if headed by the computational community at large, will increase the efficacy of the “mind as machine” project. And, again, the point isn’t to say that thinking and feeling can’t be implemented via a machine like a computer; it’s to say that there are probably more efficient ways. And Schulman’s initial point is that to advance this efficiency it behooves the research community, interested in thinking machines, to understand the true nature of their computers as they relate to our understanding of mind/brain; the only existent implementation, to date, of thinking and feeling. He further makes this point as follows on p. 63-64:

Properly understood, the first question underlying the AI debate is:
Can the properties of the mind be completely described on their own terms as an algorithm? Recall that an algorithm has a definite start and end state and consists of a set of well-defined rules for transitioning from start state to end state. As we have already seen, it was the explicit early claim of AI proponents that the answer to this question was yes: the properties of the mind, they believed, could be expressed algorithmically (or “procedurally,” to use a more general term). But the AI project has thus far failed to prove this answer, and AI researchers seem to have understood this failure without acknowledging it. The founding goal of AI has been all but rejected, a rejection that carries great significance for the central presumption of the project but that has gone largely unremarked. As an empirical hypothesis, the question of whether the mind can be completely described procedurally remains open (as all empirical hypotheses must), but it should be acknowledged that the failure thus far to achieve this goal suggests that the answer to the question is no—and the longer such a failure persists, the greater our confidence must be in that answer.

Once the unlikelihood of procedurally describing the mind at a high level is accepted, the issue becomes whether the mind can be replicated at some lower level in order to recreate the high level, raising the next important question: *Are the layers of physical systems, and thus the layers of the mind and brain, separable in the same way as the layers of the computer?*
 (Schulman 2009, p. 63-64)

So, Schulman is not claiming that minds cannot be implemented on machines in general. He is only questioning the specific goals and relative success of a community that has claimed for decades to be able to implement intelligence, let alone mind, on a specific kind of machine; the computer. As a designer by trade, I’m continually amazed by the resources and quality people devoted to achieving these goals via such means, when we have a perfectly good example of a thinking feeling entity awaiting more thorough focus; our own minds. Of course, the reason is not a mystery because any self-organizing system, such as human culture, has a strong predilection for the familiar over the efficient. Computers are knowable by design and have been fantastically successful in circumscribed domains; as such, we naturally try to project that

success into other domains (e.g. the mind) until we are explicitly convinced that our chosen tool is simply ill-designed for such expanded tasks.

Whether or not you believe that computers can manifest mind as they are now, a really useful question in service of the “machine as mind” project is the one Schulman posits at the end of the excerpt above. He goes on to describe why it is difficult to model any portion of the brain as we might model any portion of a computer at any scale. Any portion of the computer can be modeled completely “because the behavior of any single level can be explained without recourse to some higher or lower level.” In other words, I know what will happen if I input a high and low into an XOR circuit group, just as I know what will happen if I use its corresponding bit operator at the programming level. The output is a designed function of the inputs. The mind, however, has repeatedly defied such simple explanations. A neuron’s firing rate, for example, is not a mere function of “upstream” neuronal action potentials. A number of studies have shown that Local Field Potentials (for references Wiki “LFP”) – the collective voltage gradient in the interstitial regions around neurons – can alter spike timing. The actual wiring of the brain is a direct correlate to the relationship between neuronal spike timing among neurons. This is an example of *not* being able to distill the brain down to separable parts that can then be definitively modeled. In essence, the behavior of each functional “part” of the brain is a function of the collective behavior of many other parts. Anthony Bell makes this point with more rigor and knowledge in his paper *Levels and loops: the future of artificial intelligence and neuroscience* (Bell 1999).

In this way, I find Ben Goertzel’s contribution very interesting since I interpret his view of emotions as a consolidation state, whereby relevant internal and external information is present, but the more focus-specific and time-bound functionality of consciousness is not equipped to process it (Goertzel 2004). Subsequently, based on the arguments above, I would suggest (maybe Ben has and I missed it) that thinking in the focused conscious regime is a function of feelings and emotions born of broader time and space stimulus, which are processed more within the subconscious regime. The really tricky part of mind is that the converse is also true. Feelings and emotions are similarly functions of not just environmental and subconscious inputs, but of conscious thoughts themselves. As such, there is no “black box” or fundamental layer of abstraction, since any such proposition is simultaneously a function of the system as a whole, which is dynamic; ergo, *a priori* procedures become very inefficient in defining such open systems. Essentially, the entire system is distributed and as Schulman points out it is convenient for us to generalize these interwoven subsystems into distinct layers, but we continue to do so at the peril of our understanding of the mechanism as a whole. Reductionism is a tool of exploration not a property of nature. We need to understand its limits.

But back to the proposed distributed open system. Evolutionarily it makes sense, since each “level” or degree of consolidation has its benefits and detriments. Rational thought is definitive, but at the expense of broader focus. Such “higher level” rational thoughts are likely predicated on a system of broader sensitivity with lower threshold for token-ability (i.e. know-ability in the conscious sense). This more sensitive and broadly focused system could have been selected for since the information gathered had fitness relevance to a given agent in some environment even though it lacked specificity in time and space. In other words, nature has evolved a mechanism to take advantage of both regimes (narrow and wide bandwidth) in a very efficient form; whereby the narrow system is intrinsically integrated with the wider system, thus, not only conserving energy in implementing both systems, but also conserving computation in

translating between regimes. In this way, pattern representation can be both distinct (e.g. sights and sounds), while simultaneously integrated (e.g. red lights and sirens...emergency). My conscious state can then near-instantly switch among these “levels” of representation because they are spatially and temporally coincident upon one another; a concept called meta-stability.

I hope the previous paragraphs put just one chink in the armor of our pride for the computer, not because I don't see computation as a critical part of the solution, but because research dollars need to also support more diverse engineering projects of mind that are soundly grounded in neuroscience and not just computer science. In any case, please indulge the following survey of current efforts.

For those of you that are staunch computationalists, watch Modha's lecture at the Decade of the Mind Symposium (Modha 2007). I also suggest you watch the videos for IBM Almaden Institute's Cognitive Computing symposiums '06 and '07 (Cog. Comp. 2006, 2007). They are replete with current efforts to mimic the mind and its various abilities by very fastidious and well funded people. Among these research technicians and scientists you will find a common thread, which probably many of you share; specifically, if we can just make a computer fast and powerful enough, intelligence and mind-like qualities will emerge. It's the quantity vs. quality argument. Modha, and I suspect many other intelligent people like him, advocate what he calls a “synaptic network” as opposed to a neural network. He argues that if we can simulate the brain's synaptic activity (analogous to square waves) we can simulate mind. To date, they have simulated a mouse brain for a few seconds on IBM's super computer. I am singularly impressed with the level of integration between massive amounts of hardware and software, but I don't believe they will find what they are looking for because mind is not the sum of its synapses. I can't help but think of a musical score with only the time signature, but with no notes. And I'm sure they aren't building LFP into their models let alone any number of other chemical and genetic superimposed systems as described in Bell's paper.

In support of a more diverse research agenda, John Searle was graciously invited as the only dissenting voice at IBM's Almaden Institute 2006 Cognitive Computing symposium. After watching a few lectures in favor of computationalism, such as Henry Markram (some very cool ideas) and Robert Hecht-Nielsen (some ok ideas), I strongly suggest you watch Searle's lecture (Searle 2006). He's probably not going to talk you out of computationalism if that's your bread and butter, but he will help you frame any project in search of a thinking and feeling computer in rigorous terms.

To sum up, I believe that a machine will eventually manifest some acceptable form of mind, but that it's core architecture will not be purely discrete binary computation. Such computers as we have today will certainly be part of the solution, but there is vast potential beyond the computational dogma. I respect and follow the work of computationalists even though I think their fundamental preconceptions regarding the model-ability of mind are wrong. Why? Because these thoughtful, intelligent and driven people are the system. I am not against them, but I hope to offer, along with a modicum of others, a slightly modified view that can increase the probability of near-term success in achieving a goal we collectively share.

References (links as of April 2009) - Right-click links to open in browser

Stanford 2009 – <http://soe.stanford.edu/research/ate/asktheexpert.html>

Goertzel 2004 – <http://www.goertzel.org/dynapsyc/2004/Emotions.htm>

Schulman 2009 – http://www.thenewatlantis.com/docLib/20090220_TNA23Schulman.pdf

Bell 1999 – <http://www.cnl.salk.edu/~tony/ptrsl.pdf>

Modha 2007 – <http://www.almaden.ibm.com:80/cs/people/dmodha/>

Cog. Comp. 2006 – <http://www.almaden.ibm.com/institute/2006/agenda.shtml>

Cog. Comp. 2007 – <http://www.citris-uc.org/CognitiveComputing07>

Searle 2006– <http://www.almaden.ibm.com/institute/2006/agenda.shtml>