The Physical Symbol System Hypothesis: Status and Prospects

Nils J. Nilsson

Stanford Artificial Intelligence Laboratory
Stanford University
nilsson@cs.stanford.edu
http://ai.stanford.edu/∼nilsson

Abstract. I analyze some of the attacks against the Physical Symbol System Hypothesis—attacks based on the presumed need for symbol-grounding and non-symbolic processing for intelligent behavior and on the supposed non-computational and "mindless" aspects of brains.

The physical symbol system hypothesis (PSSH), first formulated by Newell and Simon in their Turing Award paper,[1] states that "a physical symbol system [such as a digital computer, for example] has the necessary and sufficient means for intelligent action." The hypothesis implies that computers, when we provide them with the appropriate symbol-processing programs, will be capable of intelligent action. It also implies, as Newell and Simon wrote, that "the symbolic behavior of man arises because he has the characteristics of a physical symbol system."

Newell and Simon admitted that

> The hypothesis could indeed be false. Intelligent behavior is not so easy to produce that any system will exhibit it willy-nilly. Indeed, there are people whose analyses lead them to conclude either on philosophical or on scientific grounds that the hypothesis is false. Scientifically, one can attack or defend it only by bringing forth empirical evidence about the natural world.

Indeed, many people have attacked the PSSH. Their arguments cluster around four main themes. One theme focuses on the presumption that computers can only manipulate meaningless symbols. Intelligence, some people claim, requires more than formal symbol manipulation; it requires some kind of connection to the environment through perception and action in order to "ground" the symbols and thereby give them meaning. Such connectedness is to be achieved through what some of its proponents call "embodiment." Intelligence requires a physical body that senses and acts and has experiences.

[1] Allen Newell and Herbert A. Simon, "Computer Science as Empirical Inquiry: Symbols and Search," *Communications of the ACM*. vol. 19, No. 3, pp. 113-126, March, 1976. Available online at:
http://www.rci.rutgers.edu/∼cfs/472_html/AI_SEARCH/PSS/PSSH1.html

M. Lungarella et al. (Eds.): 50 Years of AI, Festschrift, LNAI 4850, pp. 9–17, 2007.

Some even claim that to have "*human*-level" intelligence, a machine must have a human-like body. For example, Hubert Dreyfus argues that:[2]

> ... to get a device (or devices) with human-like intelligence would require them to have a human-like being in the world, which would require them to have bodies more or less like ours, and social acculturation (i.e. a society) more or less like ours.

In order to avoid arguments about what kind of body (if any) might be required, I think discussions about this theme would be less confusing if, instead of being about bodies, they were about the need for "grounding" symbols in whatever environment the intelligence is to function. Such an environment might be either the actual physical world or simulated, artificial worlds containing other agents.

Another theme focuses on the presumption that much that underlies intelligent action, especially perception, involves non-symbolic (that is, analog signal) processing. Of course, any physical process can be simulated to any desired degree of accuracy on a symbol-manipulating computer, but an account of such a simulation in terms of symbols, instead of signals, can be unmanageably cumbersome.

The third theme, related to the second, comes from those who claim that "computation," as it is ordinarily understood, does not provide an appropriate model for intelligence. Some have even said that it is time "to do away with the computational metaphor that has been haunting AI for 50 years: the brain is not a computer!"[3] Intelligent behavior requires "brain-style" (not computational) mechanisms.

A fourth theme is based on the observation that much that appears to be intelligent behavior is really "mindless." Insects (especially colonies of insects) and even plants get along quite well in complex environments. Their adaptability and efficacious responses to challenging situations display a kind of intelligence even though they manipulate no symbols. Jordan Pollack extends this claim even to human intelligence. He has written "Most of what our minds are doing involves mindless chemical activity ..."[4]

In light of these attacks, where does the PSSH stand today? Manifestly, we have not yet mechanized human-level intelligence. Is this shortcoming the fault of relying on the PSSH and the approaches to AI that it encourages? Might we need to include, along with symbol manipulation, non-symbolic processing modules in order to produce intelligent behavior? Of course, it could just be that

[2] Quote taken from http://en.wikipedia.org/wiki/Hubert_Dreyfus. Dreyfus's point of view about all this is explained in: Hubert L. Dreyfus, "Why Heideggerian AI Failed And How Fixing It Would Require Making It More Heideggerian," a paper written in connection with being awarded the APA's Barwise Prize, 2006.

[3] From a description of the "50th Anniversary Summit of Artificial Intelligence" at http://www.ai50.org/

[4] Jordan B. Pollack, "Mindless Intelligence," *IEEE Intelligent Systems*, p. 55, May/June 2006.

mechanizing intelligence is so much more difficult than we ever imagined it to be that it's not surprising that we haven't done it yet regardless of the approaches we have tried.

Let's look first at the claim that the PSSH is based on manipulating formal (and thus meaningless) symbols and is false for that reason. John Searle, for example, has written:[5]

> What [a computer] does is manipulate formal symbols. The fact that the programmer and the interpreter of the computer output use the symbols to stand for objects in the world is totally beyond the scope of the computer. The computer, to repeat, has a syntax but not semantics.

Searle makes this claim as part of his argument that computers (viewed as symbol-processing systems) cannot be said "to understand" because the objects (in the world) that the symbols stand for are beyond their scope.

Rodney Brooks has also criticized the PSSH, and proposes (in supposed contrast) what he calls "nouvelle AI . . . based on the physical grounding hypothesis. This hypothesis states that to build a system that is intelligent it is necessary to have its representations grounded in the physical world."[6]

Searle and Brooks both seem to have ignored an important part of the PSSH. According to Newell and Simon:[7]

> A physical symbol system is a machine that produces through time an evolving collection of symbol structures. Such a system exists in a world of objects wider than just these symbolic expressions themselves.

Regarding this "world of objects," a physical symbol system includes (in addition to its means for formal symbol manipulation) the ability to "designate."

Here is Newell and Simon's definition (my italics):

"An expression [composed of symbols] designates an object if, given the expression, the system can either *affect the object itself* or *behave in ways dependent on the object*."

We see that the designation aspect of the PSSH explicitly assumes that, whenever necessary, symbols will be grounded in objects in the environment through the perceptual and effector capabilities of a physical symbol system. Attacks on the PSSH based on its alleged disregard for symbol grounding miss this important point.

In any case, in many applications, it isn't clear that symbol grounding is needed. For example, the "knowledge" possessed by expert systems—expressed

[5] John R. Searle, "Minds, Brains, and Programs," *Behavioral and Brain Sciences*, 3(3), pp. 417-457, 1980. Available online at: http://www.bbsonline.org/documents/a/00/00/04/84/bbs00000484-00/bbs.searle2.html

[6] Rodney A. Brooks, "Elephants Don't Play Chess," *Robotics and Autonomous Systems*, 6, pp. 3-15, 1990. Available online at: people.csail.mit.edu/brooks/papers/elephants.pdf

[7] Allen Newell and Herbert A. Simon, *op. cit.*

in symbolic form either as belief networks or as rules—has no direct connection to objects in the world, yet "formal symbol manipulation" of this knowledge delivers intelligent and useful conclusions. Admittedly, robots that perceive and act in real environments (as well as other systems that function in artificial, simulated environments) do need direct connections between some of their symbols and objects in their environments. Shakey most certainly had a body with sensors and effectors, but most of its processing was done by a physical symbol system.

Let's turn now to the second theme, namely that intelligent action requires non-symbolic processing. It is often claimed that much (if not most) of human intelligence is based on our ability to make rapid perceptual judgments using pattern recognition. We are not able to introspect about what underlies our abilities to recognize speech sounds, familiar faces and "body language," situations on a chess board, and other aspects of our environment that we "size-up" and act upon seemingly automatically. Because we cannot introspect about them, it is difficult to devise symbol-based rules for programming these tasks. Instead, we often use a variety of dynamical, statistical, and neural-network methods that are best explained as processing analog rather than discrete symbolic data.

Statistical and neural-network methods are quite familiar to AI researchers. The subject of dynamical systems, however, might not be. In an article in *The MIT Encyclopedia of Cognitive Science*, Tim van Gelder writes:[8]

> A dynamical system for current purposes is a set of quantitative variables changing continually, concurrently, and interdependently over quantitative time in accordance with dynamical laws described by some set of equations. Hand in hand with this first commitment goes the belief that dynamics provides the right tools for understanding cognitive processes.
> ...
> A central insight of dynamical systems theory is that behavior can be understood geometrically, that is, as a matter of position and change of position in a space of possible overall states of the system. The behavior can then be described in terms of attractors, transients, stability, coupling, bifurcations, chaos, and so forth—features largely invisible from a classical perspective.

I grant the need for non-symbolic processes in some intelligent systems, but I think they supplement rather than replace symbol systems. I know of no examples of reasoning, understanding language, or generating complex plans that are best understood as being performed by systems using exclusively non-symbolic processes.[9] Mostly this supplementation occurs for those perceptual and motor

[8] T. J. van Gelder, "Dynamic Approaches to Cognition" in R. Wilson, and F. Keil (eds.), *The MIT Encyclopedia of Cognitive Sciences*, pp. 244-246, Cambridge MA: MIT Press, 1999. Available online at: http://www.arts.unimelb.edu.au/~tgelder/papers/MITDyn.pdf

[9] In his article on dynamical systems, van Gelder writes "Currently, many aspects of cognition—e.g., story comprehension—are well beyond the reach of dynamical treatment."

activities that are in closest contact with the environment. This point has long been acknowledged by AI researchers as evidenced by the inclusion of "signal-to-symbol transformation" processes in several AI systems.[10]

Pandemonium, an early AI architecture proposed by Oliver Selfridge,[11] was non-commital about the symbolic versus non-symbolic distinction. Its hierarchically organized components, which Selfridge called "demons," could be instantiated either as performing non-symbolic or symbolic processes. In combination, his model would be a provocative proposal for a synthesis of those two processing methods.

Now, let's analyze the phrase "the brain is not a computer," which is the main point of the third theme of attacks against the PSSH. People who make this claim often stress distinctions like:

> Computers have perhaps hundreds of processing units whereas brains have trillions.
> Computers perform billions of operations per second whereas brains perform only thousands.
> Computers are subject to crashes whereas brains are fault tolerant.
> Computers use binary signals whereas brains work with analog ones.
> Computers perform serial operations whereas brains are massively parallel.
> Computers are programmed whereas brains learn.
> Etc.

Aside from the fact that many of these distinctions are no longer valid,[12] comparisons depend on what is meant by "the brain" and what is meant by "a computer." If our understanding of the brain is in terms of its component neurons, with their gazillions of axons, dendrites, and synaptic connections, and if our understanding of a computer is in terms of serial, "von Neumann-style" operation—reading, processing, and writing of bits—all accomplished by transistor circuitry, well then of course, the brain is not *that* kind of a computer. So what?

We don't understand "computation" (the metaphor we are being persuaded to abandon) by reference only to a low-level, von Neumann-style description. We can understand it at any one of a number of description levels. For example,

[10] P. Nii, E. Feigenbaum, J. Anton, and A. Rockmore, "Signal-to-Symbol Transformation: HASP/SIAP Case Study," *AI Magazine*, vol 3, Spring 1982.

[11] Oliver. G. Selfridge, "Pandemonium: A Paradigm for Learning," in D. V. Blake and A. M. Uttley, editors, *Proceedings of the Symposium on Mechanisation of Thought Processes*, pages 511-529, London: Her Majesty's Stationary Office, 1959.

[12] For example, a paper written in 2003 claimed that "Google's architecture features clusters of more than 15,000 commodity-class PCs with fault-tolerant software." Undoubtedly, Google uses many more networked computers today. See: Luiz André Barroso, Jeffrey Dean, and Urs Hölzle, "Web Search for a Planet: The Google Cluster Architecture," *IEEE Micro*, March-April, 2003. Available online at: http://labs.google.com/papers/googlecluster-ieee.pdf

computation might be understood as a collection of active recursive functions operating on symbolic list structures. Alternatively, it might be understood as parallel-operating "knowledge sources" reading from, transforming, and writing complex symbolic expressions on a "blackboard." Other possible computational models are a collection of symbol-processing Pandemonium demons, a "dynamic Bayes network" of symbolically-represented propositions,[13] or a loosely-coupled society of simple computational "agents."[14]

Perhaps our gradually increasing understanding of how the brain operates will lead to other useful computational models, such as the graphical models of the neo-cortex proposed by Hawkins; by Hinton, Osindero, and Teh; by Lee and Mumford; and by Dean.[15] Our ideas about what "computation" can be are ever expanding, so those who want to claim that the brain is not a computer will need to be more precise about just what kind of computer the brain is not.

Engineers have no difficulty using several levels of description and neither will brain scientists. Transistors and synapses are best understood and explained using the vocabularies of physics and chemistry. But database systems, for example, are best understood and programmed using higher-level computational concepts—which, by the way, had to be *invented* for those purposes. Similarly I predict, understanding how brains represent declarative knowledge, understand and generate language, and make and carry out plans will require levels of description higher than that of neural circuitry. And just as engineers already have a continuum of bridges connecting an explanation of how transistors work with an explanation of how computers perform database searches, brain scientists will eventually have bridges connecting their explanations of how neurons work with their yet-to-be perfected explanations of how brains carry out those processes we call intelligent.

There is already some exciting progress on developing symbol-based theories of brain operation and on connecting these theories with neural circuitry. For example, Randall C. O'Reilly, writes that the pre-frontal cortex "is critical for maintaining current context, goals, and other information in an active state that guides ongoing behavior in a coherent, task-relevant manner."[16] He even suggests that neural circuits protect against noise in the same way that computers do, namely by employing binary encoding, and that neural circuits are capable of "limited variable binding."

---

[13] Stuart Russell and Peter Norvig, *Artificial Intelligence: A Modern Approach*, Second Edition, Chapter 15, Upper Saddle River, New Jersey: Pearson Education, Inc., 2003.

[14] Marvin Minsky, *The Society of Mind*, New York: Simon and Schuster, 1985.

[15] Jeff Hawkins with Sandra Blakeslee, *On Intelligence*, New York: Times Books, 2004; G. Hinton, S. Osindero, and Y. Teh, "A Fast Learning Algorithm for Deep Belief Networks," *Neural Computation*, 2006, to appear; T. S. Lee and David Mumford, "Hierarchical Bayesian Inference in the Visual Cortex, *J. Opt. Soc. Am. A*, Vol. 20, No. 7, July 2003; Thomas Dean, "Computational Models of the Neocortex," online article at http://www.cs.brown.edu/people/tld/projects/cortex/.

[16] Randall C. O'Reilly, "Biologically Based Computational Models of High-Level Cognition," *Science*, vol. 314, pp. 91-94, October 6, 2006.

In a paper about certain brain sub-systems, Richard Granger writes: "Together the system produces incrementally constructed and selectively reinforced hierarchical representations consisting of nested sequences of clusters."[17] Granger has also told me that "even in brains, many of us find it appropriate to include symbol-processing levels of description (though I should note that the science is evolving, and there are still those who would disagree)."[18]

In his "neural theory" of how the brain understands language, Jerome Feldman employs such computational level, symbolic constructs as "schema," "feature," and "value." He writes, "There is convincing evidence that people organize their perceptions and actions in terms of features and values."[19] Feldman stresses the importance of connecting computational level descriptions in his theory to "key neural properties, including massive parallelism, robustness, spreading activation, context sensitivity, and adaptation and learning."[20]

No doubt AI research will benefit greatly from what computational neuroscientists and cognitive scientists learn about how the brain works. But I don't think it will involve abandoning the computational metaphor.

Now, what about the idea that intelligence is "mindless"? Several examples of mindless processes are cited by adherents of this view. Here are some cited by Jordan Pollack,[21] who coined the word "ectomental" to describe them: The process of evolution, proceeding by random changes in the genome and selective survival of organisms that result from the genome, produced intelligent humans. (But *producing* an intelligent system is different from *being* an intelligent system.) Reinforcement learning produced a neural network that plays better backgammon than human experts. (Pollack failed to note that the inputs to the neural network were symbolic features of the backgammon board and that the best performance was obtained in combination with limited-look-ahead symbolic search.) The animal immune system can discriminate between self and non-self without "a central database listing which compounds are in or out." He concludes by writing that "dynamical processes, driven by accumulated data gathered through iterated and often random-seeming processes, can become more intelligent than a smart adult human, yet continue to operate on principles that don't rely on symbols and logical reasoning." So far, no such "dynamical processes" have produced systems that can prove theorems, make and execute plans, and summarize newspaper stories. And, when and if they ever do produce such systems, they will be best explained, I predict, as using "symbols and logical reasoning." Pollack's statement that "Most of what our minds are doing involves mindless chemical activity ..." is no more helpful than would be

[17] Richard Granger, "Essential Circuits of Cognition: The Brain's Basic Operations, Architecture, and Representations," in J. Moor and G. Cybenko (eds.), *AI at 50: The Future of Artificial Intelligence*, to be published. Available online at: http://www.dartmouth.edu/∼rhg/pubs/RHGai50.pdf

[18] E-mail communication, October 4, 2006.

[19] Jerome A. Feldman, *From Molecules to Metaphor: A Neural Theory of Language*, p. 140, Cambridge, MA: The MIT Press, 2006.

[20] *Ibid*, p. 142.

[21] Jordan B. Pollack, *op. cit.*

a statement like "Most of what airline reservations systems are doing involves mindless electronic currents."

Rodney Brooks has achieved a great deal of success in using his "nouvelle AI" ideas to program rather simple (one is tempted to say mindless) "creatures." Most of his systems lack complex representations, even though his "physical grounding hypothesis" doesn't explicitly disallow them. Nevertheless, the behaviors of these creatures are quite impressive and are described in his "Elephants Don't Play Chess" paper. But the title of that paper belies the difficulty. They don't, do they? Brooks attempts to deflect such criticism by writing "it is unfair to claim that an elephant has no intelligence worth studying just because it does not play chess." But I don't claim that elephant "intelligence" is not worth studying. I only claim that, whatever it is, it isn't human-level intelligence, and I think more complex representations, symbolically manipulated, will be needed for that.

In summary, I don't think any of the four different kinds of attacks on the PSSH diminishes the importance of symbolic processing for achieving human-level intelligence. The first attack is based on the erroneous claim that the PSSH lacks symbol grounding. By all means, let's have symbol grounding when needed. The second attack is based on the need for non-symbolic processing; let's have that too when needed. The third attack, based on the claim that the brain is not a computer, will vanish when people who study brains increasingly use computational concepts to understand brain function. And the fourth attack, based on the idea that brains are mindless, will vanish when it becomes evident that constructs best understood as being mindless achieve only mindless behavior.

So what does all this have to say about the status of the PSSH? Some might say that the PSSH's claim that a physical symbol system is "sufficient" for intelligent action is weakened by acknowledging that non-symbolic processing might also be necessary. Newell, however, seemed not to be willing to concede that point. In a 1988 book chapter, he wrote:[22]

> ... the concept of symbols that has developed in computer science and AI over the years is not inadequate in some special way to deal with the external world."
> ...
> For example, such symbols are used as a matter of course by the Navlab autonomous vehicle (a van that drives itself around Schenley Park next to Carnegie-Mellon), which views the road in front of it through TV eyes and sonar ears, and controls the wheels and speed of the vehicle to navigate along the road between the trees ... The symbols that float everywhere through the computational innards of this system refer to the road, grass, and trees in an epistemologically adequate, though sometimes empirically inadequate, fashion. These symbols are the symbols of the physical symbol system hypothesis, pure and simple.

[22] Allen Newell "Putting It All Together," Chapter 15, of D. Klahr and K. Kotovsky (eds.), *The Impact of Herbert A. Simon*, Hillsdale, NJ: Erlbaum and Associates, 1988.

I'll leave it at that. For those who would rather think about the perception and action routines of Navlab (and of Shakey and Stanley) in terms of signals rather than symbols, the "sufficiency" part of the PSSH is clearly wrong. But the "necessity" part remains uncontested, I think.

What about the future prospects for physical symbol systems in AI? Brooks's "Elephant" paper makes a proposal:

> Traditional [that is, symbolic] AI has tried to demonstrate sophisticated reasoning in rather impoverished domains. The hope is that the ideas used will generalize to robust behavior in more complex domains.
> Nouvelle AI tries to demonstrate less sophisticated tasks operating robustly in noisy complex domains. The hope is that the ideas used will generalize to more sophisticated tasks.
> Thus the two approaches appear somewhat complementary. It is worth addressing the question of whether more power may be gotten by combining the two approaches.

Here is my prediction about the future of physical symbol systems in AI: They will take on a partner (as Brooks proposes). AI systems that achieve human-level intelligence will involve a combination of symbolic and non-symbolic processing—all implemented on computers, probably networks of computers. Which parts are regarded as symbolic and which parts non-symbolic will depend on choices of the most parsimonious vocabulary and the most useful programming constructs—which, after all, are intimately linked. We will find it most convenient to describe some parts with equations involving continuous and discrete numbers. And, those parts will correspondingly be programmed using operations on continuous and discrete numbers. We will find it most convenient to describe the higher level operations in terms of non-numeric symbols. And, those parts will correspondingly be programmed using symbol-processing operations.

In the not-too-distant future, I hope, the controversies discussed in this paper will be regarded as tempests in a teapot.

Acknowledgments

I would like to thank Jerome Feldman, Richard Granger, Max Lungarella, and Yoav Shoham for their helpful suggestions.