Understanding each other: Defining a conceptual space for cognitive modeling

Robert West (robert_west@carleton.ca) &
David Pierre Leibovitz (dpleibovitz@ieee.org)
Institute of Cognitive Science, 2201 DT, Carleton University
1125 Colonel By Drive, Ottawa, Ontario, K1S 5B6 Canada

Abstract

Cognitive modeling is a complex endeavor so it is not surprising that the goals and intentions of modelers are often misunderstood, even by other modelers. To try to clarify this we have attempted to map out the various philosophical and theoretical commitments that one makes when creating a cognitive model or architecture. The goal of this is to avoid misunderstandings between the adherents of different modeling systems and between cognitive modelers and the rest of the scientific community.

Keywords: cognitive modeling.

Introduction

In the 1990s there was movement to contrast mainstream cognitive modeling, which was labeled as *cognitivist*, with alternative approaches which were asserted to represent a fundamentally different paradigm. These alternatives included *situated* cognition, *distributed* cognition, *dynamicism*, *embodied* cognition and *subsumption* architectures. However, Vera and Simon (1993) argued that these theories represented progress and innovation but not an alternative approach. They did this by arguing that critics of the mainstream view had mistakenly assumed that the practices, strategies, and short cuts of mainstream modelers represented their actual philosophical and theoretical commitments. We argue that Vera & Simon's argument was a legitimate response and more generally that philosophical and theoretical commitments cannot be determined solely by analyzing systems and practices associated with a method of modeling. Full understanding requires explicating the philosophical and theoretical commitments of the modeler.

Today all of the alternatives to mainstream modeling discussed in Vera and Simon (1993) are accepted in the main stream. That is to say, they are broadly recognized as important contributions. At this time it would be fair to say that there is no widely accepted, "main stream" approach to modeling. However, despite efforts to understand different modeling systems as alternative approaches, each with their own strengths (e.g., McClelland, 2009) there are still numerous attempts to vilify modeling systems by critics who do not fully understand the goals and intentions of the system creators and users. In our view, the idea that a modeling system can be dismissed based on a principle or a philosophical argument is, in fact, a philosophical mistake.

We argue that pigeon holing modeling systems according to broad philosophical and theoretical distinctions (e.g., *cognitivist* vs. *anti-cognitivist*, *computational* vs. *non-computational*, *representational* vs. *anti-representational*, etc.) is misleading and counterproductive. In place of this, we advocate a multidimensional approach to characterize modeling systems along numerous dimensions, including the beliefs and motivations of the modeler. Thus in our system, two modelers can use the same computational code but actually have very little in common. Likewise two modelers could use very different codes (e.g., the ACT-R symbolic/subsymbolic code, J. R. Anderson & Lebiere, 1998; and the NENGO spiking neuron code, Eliasmith & Anderson, 2003) and still be completely on the same page. To illustrate this approach we will use the controversial example of the "symbol" throughout the paper, although each dimension can be applied to any modeling construct. Due to limited space we have focused on dimensions that we believe are important.

AI and Useful fictions

The strong AI hypothesis says that if the functions of the human mind can be correctly simulated on a computer then there will be no difference between the human mind and the computer mind. It is important to note that this hypothesis is silent about the level of abstraction or embodiment required for success. It could involve high level algorithms realized as software, or it could involve a brain made of highly realistic mechanical neurons embodied in a lifelike humanoid robot and raised as a human infant. Therefore, if we apply strong AI to symbols it means that the symbols in a model are a valid way of representing what is taking place in the brain. Alternatively, symbols can be viewed as *useful fictions*. That is, the brain does not process symbols but there is something about the processing of symbols that is analogous to how a brain works and it is therefore useful or expedient to model it in this way.

Metaphysics

In mainstream western philosophy there are substances and processes that act on the substances. For example, logical formalisms can be used to act on symbolic representations about the state of the world. We will refer to this as substance philosophy. In process philosophy (see Bickhard, 2010; Whitehead & Griffin, 1931) there are no modular substances, only interacting processes. What appear to be substances are temporary emergent properties of ongoing processes. According to Quantum Physics, process philosophy is true for physical objects. For example, the chair across the room is a temporary quantum process not an independent object. Likewise, when you close your eyes and

remember the chair, that memory is a temporary neural process not an object.

Often psychological and linguistic constructs are discussed as if they were actual objects, which leads to confusion because it could mean they are meant as objects (i.e., substance philosophy) or that they are meant as a simplification standing in for a process (i.e., process philosophy). For example, the use of symbols in a model could signify that symbols exist, or it could signify that there is a process that acts as though symbols exist. More generally, a process philosophy view implies that psychological or linguistic constructs in a model should be regarded as abstract proxies standing in for interacting processes – not informationally encapsulated modules as suggested by Fodor (1983). The issue then becomes the relative stability of the processes underlying psychological constructs. Process philosophy is silent on this - neural processes giving rise to psychological constructs could be very stable, resulting in a relatively crisp, well defined constructs; or they could be noisier, resulting in fuzzy and possibly temporary constructs. Determining this, in our opinion, is not a philosophical issue. However, deciding if constructs, such as symbols, actually exist is a philosophical issue.

Divide and conquer versus unification

The simplicity principle (Chater & Vitányi, 2003) refers to the idea that cognitive phenomena are best modeled in the simplest way. This goal, related to Occam's Razor, is not contentious but becomes less clear when the scope of the phenomena is considered. Newell, in his famous (1973) paper, argued that the phenomena to be explained is the whole brain. Newell (1990) distinguished between micro models (i.e., independent models of different phenomena) and architectural models (i.e., models constrained by the use of a cognitive architecture aimed at describing the whole brain). The goal with micro models is to make them as simple as possible but the goal for models built in an architecture is more complex. The model should be as simple as possible given the constraints of the architecture but the actual goal is to produce an architecture that is as simple as possible across numerous models of different phenomena (including neural phenomena, (J. R. Anderson, 2007a)).

Therefore, from an architectural point of view, *having lots of incommensurate micro models is not useful*, regardless of how simple they are individually. One way around this is to argue that the micro model represents a distinct cognitive/neural module that encapsulates and operates on a particular kind of information (Fodor, 1983). For example, by arguing that there is a distinct symbol based language module, one can ignore issues or problems concerning the viability of using symbols to model other functions of the brain. Therefore, the form of a particular model could reflect the goal of creating a unified architecture, of understanding a distinct module, or of creating the simplest model for a specific phenomenon.

Reverse and forward Engineering

Reverse engineering involves testing a system, the brain in this case, and working backwards to discover how it functions. Unfortunately, having a model that performs similarly to a human does not confirm that it is a valid model because other future tests may disconfirm it.

However, modeling can also move forward through forward engineering. Forward engineering involves designing a system with the goal of achieving certain functions. Therefore, the goal is to achieve the same functionality that humans have without worrying about doing it in the same way as the brain does. If successful the result would be a system that is roughly isomorphic to how the brain behaves, but does not give insight into how the brain does it. For example, the use of symbols in a model could be due to the belief that symbols behave isomorphically to neural representations.

The difference between backward and forward engineering is also important when evaluating how a model has been evaluated. Generally speaking, attempts to reverse engineer involve careful comparisons to experimental data while attempts to forward engineer involve showing that a model can produce certain functionality. It is important to note that a modeler may also iterate between reverse and forward engineering.

Epistemic commitments

Epistemic commitment refers to the mechanisms used to build the model. Specifically, we mean it to refer to a commitment to a particular way of understanding and modeling the brain. The debate between proponents of symbol systems and proponents of neural networks is an example of an epistemic debate. The idea motivating such debates is that it is necessary to first get the way of modeling right, otherwise the resulting models will be misleading and ultimately dead ends. This issue has fueled a lot of debate within cognitive science. Examples of different systems are: *symbol systems*, *neural networks*, *holographic systems*, *dynamic systems*, *spiking neuron models*, *Bayesian networks*, *logical systems*, *grammatical systems* etc. However, it is also possible to view these as tools rather than competing theories, in which case the choice of a particular way of modeling would reflect a pragmatic choice rather than a principled one. Another approach is to view different modeling systems as different lenses for viewing a phenomena in different ways (McClelland, 2009).

Related to this, it is important to note that the word *architecture* is used in two ways. As noted above, it can refer to a system meant to be a unified model of the whole brain, or it can refer to mechanisms for building models that are able to model the whole brain. For example, ACT-R (J. R. Anderson, 1993) is an attempt at creating a unified architecture but it is meant to be a hybrid system and therefore does not embody an epistemic commitment. ACT-R is often described as a production system but this is incorrect as ACT-R has numerous modules that use numerous mechanisms. The use of a production system

module to coordinate the other modules in ACT-R is a commitment to a theory about unification; it is not an epistemic commitment (i.e., a way of understanding the whole brain). In contrast, NENGO (Eliasmith & Anderson, 2003) is a system for building spiking neuron models according to a specific theory about spiking neurons, so the use of NENGO can be seen as an epistemic commitment (i.e., that the whole brain can be modeled in this way).

Ontological commitments

Ontological commitment refers to the way a model is divided into functional parts and their connectivity. There are two reasons why ontological commitments are important. The first has to do with creating unified cognitive architectures. Simply put, a valid cognitive model of the whole brain requires that the functional parts of the model map onto the functional parts of the brain. Although the results of experimental psychology are good for testing models, they may be misleading in terms of telling us what the parts are because their ontologies are defined primarily to make experimentation possible on different psychological phenomena. Unfortunately this does not necessary tell us what the actual parts are. For example, we remember facts and we remember episodes and these can be treated separately for experimental purposes, but we still do not know if we have separate semantic and declarative memory systems or if they are both products of a single long term memory system.

Another very important issue related to ontologies is cognitive re-use (see M. L. Anderson, 2010). This refers to whether or not our cognitive ontology corresponds to a dedicated neural area. Much of the neural localization work taking place today explicitly or implicitly assumes that it does. However, the cognitive re-use hypothesis is that higher-level cognitive mechanisms and functions can be created by re-using and recombining lower level cognitive mechanisms and functions. If this is true then there are two important consequences: (1) specific brain areas are not dedicated to specific cognitive functions, and (2), the ontology that we should be looking for is at a lower level.

Therefore, modelers may believe that the modules of their system correspond to actual cognitive functions in the brain, and they may further believe that these functions map to dedicated areas of the brain. But having a module in a model does not necessarily mean that they believe either of these things. For example, following the cognitive re-use hypothesis, a module could also represent a function that is created through the interaction of lower level functions under specific conditions. Symbols, or any other construct, can be thought of in either way.

System levels

Allan Newell (1980) proposed that the brain is constructed in the way that computers are engineered, according to system levels. The reason why natural systems would develop distinct hierarchical levels was developed by Simon (1962) but a discussion of this is beyond the scope of this paper. A system level occurs when the behavior of a complex lower level system can be understood in terms of less complex higher level constructs. For example, in the theory of thermodynamics, the complex interactions of atomic particles can be understood through higher level concepts such as heat and pressure. So a systems level is a real thing (in as much as heat and pressure are real things) but it is important to note that a system level can be weak or strong depending on the relative reduction in complexity produced by the emergent level. A weak system level is leaky, meaning that it is sometimes affected be system levels below it (e.g., Saunders, Kolen, & Pollack, 1994).

The cognitive level is theorized to exist as a systems level above the neural level but there is considerable controversy over whether it exists and if it does, what form does it take? The symbol system hypothesis asserted that the cognitive level is based on processing symbols. Likewise, Chomsky (e.g., Chomsky, 1995) argued that for understanding language, symbols could be divorced from the underlying system that produces them. However, it is instructive look at exactly what was meant by, "symbol." For Chomsky a symbol is a word, but Newell defined a symbol in terms of distal access (Newell, 1990). Distal access refers to using information that is not local, i.e., information that is transported from another part of the brain. The form of the information or the way it is transferred is not important, therefore Newell's commitment to symbols is completely different from Chomsky's commitment to symbols.

Level of Analysis

Level of analysis is different from system level. Level of analysis refers to analyzing a system at a particular level (e.g., neural, neural groups, networks, symbols). Using a level of analysis may or may not indicate a belief that the level is a systems level. So the use of symbols in a model may indicate a commitment to the symbol system hypothesis, but it could also occur because that level of analysis is useful, without any commitment to the existence of an actual systems level. Also, it is possible to test a model constructed at a higher systems level using a lower level of analysis if there is a theory about how the lower level is related to the higher level. For example, ACT-R models can be tested using a neural level of analysis with an fMRI scan (J. R. Anderson, 2007b). Choice of a level of analysis reflects beliefs about the most effective way of testing a model.

Consciousness

Explaining consciousness is a special case of the strong AI issue that deserves its own section. The question is, could a properly constructed cognitive architecture actually have conscious experiences. From a strong AI point of view the answer is yes, but many people reject this position because they find it hard to imagine. This seems to be due mainly to our subjective experience of qualia.

Qualia refer to the various phenomenal feelings of our conscious experiences. From a modeling point of view

qualia creates a potential problem because thought, emotion, and different types of perception do not feel the same to us; they feel qualitatively different. However, from a cognitive science perspective, and a neuroscience perspective, all qualia arise from information processing that is ultimately realized through the firing of neurons. Since we do not understand what consciousness is or how it creates different qualia from the same underlying mechanism, most cognitive models simply ignore the issue or focus on the correlates of consciousness (e.g., awareness, wakefulness, report ability, etc.).

However, the concept of qualia is important for modeling because it cuts across the board and separates the issue of how information is processed from how it is subjectively experienced. By setting aside the issue of qualia we are implicitly adopting the view that qualia is an epiphenomena; that is, we can model the brain without considering qualia because qualia has no functional significance (Dennett, 1991). This is very convenient since it allows us to model all aspects of the brain as information processing and ignore or put off the problem of explaining why different brain functions feel qualitatively different from each other.

Alternatives to understanding consciousness as an epiphenomena arising from information processing are scarce. Searle (1980) makes his arguments against strong AI by arguing that it leads to absurd consequences or conclusions (the Chinese room is his most famous example), but he does not offer an alternative explanation. Hameroff & Penrose (1996) argue that normal information processing is inadequate to model human cognition and consciousness. They propose that the brain is capable of quantum computing and therefore a valid simulation would require a quantum computer. Although this view is not popular it should be noted as quantum computing is so far the only scientific alternative to normal computing, although, as Penrose concedes, it is still a type of information processing. Chalmers (2010) has argued that if you reject that consciousness arises from information processing, the only option is to adopt some form of dualism.

Philosophy of science

Some people define science with Popper's (1935) notion of falsifiability. However, although it is in theory possible to falsify cognitive models, it is often the case that the failure of a model leads to changes in the model rather than a rejection of the model. With unified architectures the problem of falsification is trickier because in order to test the architecture, it must be used to build a model of a task, therefore, if it fails, it is unclear if the architecture has been falsified or just the model. Newell (1990) realized this and argued that Lakatos' definition of science (1970) was more appropriate than Popper's for understanding architectures. Essentially, Lakatos defines science in terms of making progress over time, therefore if an architecture or model is improved through testing and refinement so that it explains more, it can be considered scientific (for a detailed discussion see Cooper, 2007).

It is interesting to note that although some of the criticism directed at testing models comes from Experimental Psychology, Experimental Psychology also fails to follow Popper's model. Specifically, most experiments in Experimental Psychology test for significant differences predicted by a theory, therefore, falsifying the theory would mean showing no significant difference, which would mean accepting the null hypothesis, which is not allowed in the ANOVA or t-test statistics that are generally used. Like modeling, theories in Experimental Psychology are generally altered and not rejected. In both cases it is possible to construe theories that are falsifiable; it is just not very common. The criticism of modeling coming from Experimental Psychology has more to do with statistics. Specifically, Experimental Psychology has a clear definition for defining when two conditions are significantly different. In contrast, the goal for a model is to show that it is significantly similar to a set of data and there is not an agreed upon standard for this (e.g., Roberts & Pashler, 2000), although there are statistical ways to tackle the issue (e.g., Stewart & West, 2010).

Another issue arises from comparisons with Computer Science or Engineering where it is common to evaluate algorithms against each other according to some clear criterion or test set. According to this approach, cognitive models should be compared to see which one explains the data best. This can be done when the models are specifically designed to model the same problem (see Erev et al., 2010 for an example). However, it is not commonly done because models are designed with different goals in mind, therefore a good test set for one might be an inappropriate or poor test set for another. It all depends on the goals and the theoretical framework of the modeler, which is why it is important to be clear about these.

Conclusion

We have outlined a number of dimensions on which modelers can take different views. Most of them are binary so it is possible to say agree, disagree, or agnostic. This list is not exhaustive, but being aware of where we stand on these issues can potentially avoid a lot of misunderstanding and provide a richer view of the whole modeling enterprise.

We have tried to be neutral in terms of laying out this list but we acknowledge that some people may feel that some of the choices we have presented are invalid. For, example, one could argue that there is no such thing as system levels in the brain. Our point is that we should separate that argument from the evaluation of modeling systems that appear to embody systems levels.

Another issue is the relationship between the different dimensions that we have laid out. People tend to associate sets of beliefs with the use of different modeling systems. Possibly some of the dimensions we described are correlated and logically go together. However, arguments about whether certain dimensions are conceptually related

or conceptually independent should be separated from subjective impressions concerning the co-occurrence of dimensions across the users of different modeling systems.

In order to progress in understanding the various cognitive modeling spaces, the impacts of these and other dimensions need to be further deliberated.

References

Anderson, J. R. (1993). *Rules of the Mind* (p. 336). Lawrence Erlbaum.

Anderson, J. R. (2007a). Using Brain Imaging to Guide the Development of a Cognitive Architecture. In W. D. Gray (Ed.), *Integrated Models of Cognitive Systems* (pp. 49-62). New York, NY: Oxford University Press.

Anderson, J. R. (2007b). *How Can the Human Mind Occur in the Physical Universe? Science* (p. x-290). Oxford University Press.

Anderson, J. R., & Lebiere, C. (1998). *The atomic components of thought* (p. 504). NJ: Erlbaum.

Anderson, M. L. (2010). Neural reuse: A fundamental organizational principle of the brain. *Behavioral and brain sciences*, *33*(4), 245-66.

Bickhard, M. H. (2010). Does Process Matter? An Introduction to the Special Issue on Interactivism. *Axiomathes*, *21*(1), 1-2.

Chalmers, D. J. (2010). *The Character of Consciousness*. *Consciousness and Cognition* (p. xxvii-596). Oxford University Press.

Chater, N., & Vitányi, P. (2003). Simplicity: a unifying principle in cognitive science? *Trends in cognitive sciences*, *7*(1), 19-22.

Chomsky, N. (1995). Language and Nature. *Mind*, *104*(413), 1-61.

Cooper, R. P. (2007). The Role of Falsification in the Development of Cognitive Architectures: Insights from a Lakatosian Analysis. *Cognitive Science*, *31*(3), 509-533.

Dennett, D. C. (1991). *Consciousness Explained* (p. 528).

Eliasmith, C., & Anderson, C. H. (2003). *Neural Engineering: Computation, Representation, and Dynamics in Neurobiological Systems* (p. 376). Cambridge, MA: MIT Press.

Erev, I., Ert, E., Roth, A. E., Haruvy, E., Herzog, S. M., Hau, R., Hertwig, R., et al. (2010). A choice prediction competition: Choices from experience and from description. *Journal of Behavioral Decision Making*, *23*(1), 15-47.

Fodor, J. A. (1983). *The Modularity of Mind: An Essay on Faculty Psychology* (p. 154).

Hameroff, S., & Penrose, R. (1996). Orchestrated reduction of quantum coherence in brain microtubules: A model for consciousness. *Mathematics and Computers in Simulation*, *40*(3-4), 453-480.

Lakatos, I. (1970). Falsification and the methodology of scientific research programs. In I. Lakatos & A. Musgrave (Eds.), *Criticism and the Growth of Knowledge* (pp. 91-196). Cambridge, UK: Cambridge University Press.

McClelland, J. L. (2009). The Place of Modeling in Cognitive Science. *Topics in Cognitive Science*, *1*(1), 11-38.

Newell, A. (1973). You Can't Play 20 Questions with Nature and Win: Projective Comments on the Papers of this Symposium. In W. G. Chase (Ed.), *Visual Information Processing: Proceedings of the 8th Symposium on Cognition* (pp. 283-308). New York: Academic Press.

Newell, A. (1980). Physical Symbol Systems. *Cognitive Science*, *4*(2), 135-183.

Newell, A. (1990). *Unified Theories of Cognition* (pp. 1-530). Cambridge, MA: Harvard University Press.

Popper, K. R. (1935). *The logic of scientific discovery* (English tr.). New York, NY: Basic Books.

Roberts, S., & Pashler, H. (2000). How persuasive is a good fit? A comment on theory testing. *Psychological Review*, *107*(2), 358-367.

Saunders, G. M., Kolen, J. F., & Pollack, J. B. (1994). The Importance of Leaky Levels for Behavior-Based AI. In D. Cliff, P. Husbands, J.-A. Meyer, & S. W. Wilson (Eds.), *From Animals to Animats 3: Proceedings of the Third International Conference on Simulation of Adaptive Behavior (Complex Adaptive Systems) SAB94*. MIT Press.

Searle, J. R. (1980). Minds, brains, and programs. *Behavioral and Brain Sciences*, *3*(3), 417-457.

Simon, H. A. (1962). The Architecture of Complexity. *Proceedings of the American Philosophical Society*, *106*(6), 467-482.

Stewart, T. C., & West, R. L. (2010). Testing for Equivalence: A Methodology for Computational Cognitive Modelling. *Journal of Artificial General Intelligence*, *2*(2), 69-87.

Vera, A. H., & Simon, H. A. (1993). Situated Action: A Symbolic Interpretation. *Cognitive Science*, *17*(1), 7-48.

Whitehead, A. N., & Griffin, D. R. (1931). Process and Reality. *Economica*, (32), 251.