Kognitionswissenschaft (1999) 8: 1–4

Kognitions-
wissenschaft

Editorial

Cognitive modeling in perspective

Herbert A. Simon[1], **Dieter Wallach**[2]

[1] Psychology Department, Carnegie Mellon University, Pittsburgh, PA 15213 USA
[2] Institut für Psychologie, Universität Basel, Bernoullistrasse 16, CH-4056 Basel, Switzerland

Introduction

In *Elements of a theory of human problem solving*, the authors argued that "an explanation of an observed behavior of the organism is provided by a program of primitive information processes that generates this behavior" (Newell, Shaw & Simon, 1958, p. 151). Over the past four decades the view that human cognition can be conceptualized as computational processes operating on symbolic (and subsymbolic) representations has grown into the modeling approach which is one of the methodological hallmarks of cognitive science, an *Interdiscipline* (Tack, 1997) that seeks to understand human cognition by linking together empirical studies, theoretical analyses and computational models.

Cognitive modeling fills the "theoretical vacuum" (Miller, Galanter & Pribram, 1960, p. 11) between cognition and observable action by specifying a detailed mechanistic process that is actually sufficient to generate the phenomena under investigation. Cognitive simulation is thus a constructive-synthetic activity of designing and testing generative computational systems that are sufficient to perform the task whose performance they are explaining while constrained to fit the available empirical data on how humans perform the same task.

The computer programs that serve as cognitive simulation models may take the form of classical differential equations whose variables are numbers, or, more commonly, of difference equations whose variables are symbol structures constituted of words, mathematical expressions, diagrams or pictures.

Reconstructing and predicting in precise terms the trajectory of human thinking calls for temporally dense recordings of behavior as evidence for the hypothetical cognitive structures and processes that are postulated in the models, but that are not directly observable. Therefore, from its very beginnings (see Newell & Simon, 1972, p. 885), cognitive modeling has utilized such empirical data-gathering methods as think-aloud protocols (Ericsson & Simon, 1993) and eye-tracking data (Just & Carpenter, 1985) that provide fine-grained, rich streams of evidence for intermediate cognitive states during performance.[1]

Cognitive modeling is not an alternative to empirical studies – a substitute for experiments – but a powerful tool for formulating hypotheses so that they may be tested against temporally fine-grained experimental data.

The concept of a cognitive model

Cognitive models are deep models (Moravsik, 1980, p. 28) that refer to theoretical terms – hypothetical constructs – to explain human cognition. As such they necessarily go beyond what can be observed empirically, and their values of theoretical terms must be estimated indirectly from the empirical data by defining them as functions of observables (Simon, 1977). The presence of theoretical terms in the models, however, is not unique to cognitive modeling, for their presence is quite as essential to verbal psychological theories, and for that matter, to theories throughout all the sciences.

Physical theories, for example, of subatomic material structure were based on exceedingly indirect observations, without directly observing elementary particles. The "track" of a particle in a cloud chamber is not a sighting of the particle itself, but of the droplets of water it condenses as it (presumably) passes. Nor did Mendel, in formulating and testing basic genetic theory, ever see a chromosome, much less a gene.

The problem of theoretical terms may be avoided in research that produces mere descriptions of the end products of human cognition (e.g., button pushes or time latencies). Such phenomena can be observed directly. However, if the goal is to explain human cognition in terms of its underlying structures and processes, empirical data can constrain severely but not identify uniquely the "correct" model – a fact that the Moore Theorem in automata theory illuminated more than 40 years ago (Moore, 1956), and which was known to Hume several centuries earlier.

A cognitive simulation model (in a manner similar to every scientific theory) maps hypothesized components and processes of the human cognitive system S on the components and processes of a computational model M. Formally, this mapping can be described as a partial homomorphism

[1] This richness of information can be illustrated by comparing the data from a typical experimental trial with the content of a thinking aloud protocol. If an experimental trial requires a binary choice between two responses a, b it yields exactly one bit of information. In a protocol experiment that produces n responses per trial each having m values, $nlog_2(m)$ bits of data constrain the model (Anderson, 1987, p. 472).

from S onto M. The mapping should preserve the relations between components, but it is a partial mapping that does not necessarily comprise all components and relations of S.

In cognitive modeling, the mapping is generally carried down to the level of symbolic processes. Modeling at the neural level, at least for simulation of humans, is still a dream for the future, not a present reality[2].

The *Physical Symbol Systems* (PSS) Hypothesis, which characterizes a very broad class of successful cognitive models, states that a "physical symbol system has the necessary and sufficient means for general intelligent action" (Newell & Simon, 1976). It claims that a PSS can be programmed to think, and that a thinker, human or not, is a PSS[3].

A symbol system describes the human cognitive machinery in terms of patterns (symbol structures). It postulates a memory for storing symbols and operators for encoding, manipulating and transforming input symbols into output symbols. It provides for "branching", so that behavior can depend on what patterns are present. The human cognitive architecture is postulated to be mappable onto a physical symbol system. The hypothesis is tested by building such systems and comparing their behavior with human behavior on the same tasks.

The completely different physical devices used in computers built in 1960 and 1999, respectively, (vacuum tubes or mercury delay lines versus chips) do not prevent 1999 computers from running programs written in Fortran or Lisp for 1960 computers. By the same token, there is no intrinsic reason why a silicon PSS cannot model a PSS constructed of neurons.

As all stimuli that arouse the sense organs must be translated, once they are inside the head, into signals that can be transmitted by neurons and stored in neural tissue, the cognitive modeller seeks to construct the corresponding signalling system in the computer, using its physical structures to implement symbolic mechanisms and processes that parallel those of the human system. As the modeling is not carried below the symbolic level (in either human or computer), the radical difference in their hardware will not defeat the effort if the physical symbol system hypothesis is correct.

In our models, we need to distinguish between *positive analogies* (entities of S that are successfully captured by the mapping onto components in M) and *false analogies* (components of M that do not have counterparts in S). In statistical terms, *type I errors* (errors of omission) result when components of S that are relevant for the question under study are not modeled in M; *type II errors* (errors of commission) result when M incorporates entities that do not have counterparts in S. Type I errors miss relevant positive analogies, while type II errors embody false analogies. The central goals when evaluating cognitive models are to find empirical support for the positive analogies and to use evidence of false analogies as starting points for model revision.

[2] *Connectionist* and *neural-network* models seek to carry theory down to relatively fine perceptual structures and processes, but are still far from modeling identifiable neurons or neuron structures, and deal with phenomena at approximately the same "grain size" as symbolic models like EPAM (Richman & Simon, 1989; Richman, Staszewski & Simon, 1995).

[3] The term *symbol* here refers to any pattern, on any physical substrate that can be used to denote pictures, words, diagrams, numbers, and sensory patterns of any kind.

Methods of evaluation

Three decades ago, Fridja (1967, p. 65) pointed out (with perhaps some exaggeration) the need for more rigorous methods for analyzing computational theories of cognition and evaluating them empirically: "There is hardly any methodology existing here. As much ingenuity has been invested in the making of the programs, as little has been spent on the assessment of their value. Next to high precision there always seems to be spots of rough approximations which undercut this very precision. We are largely left to our subjective impressions of what we consider good or bad correspondence".

Today we still need considerable improvement in our methodologies for evaluating models, but many examples of careful evaluation can be found already. First of all, a running generative model provides an existence proof for the sufficiency of its structures and mechanisms for a given task, hence, a first strict falsification test for the model.

However, there are different degrees of sufficiency. Unless the tests for matching model against human behavior are sufficiently refined, a model may be able to match behavior closely at a temporal grain of 30 s, but be operating in a wholly different way from human subjects at grains of 1 s, or of 10 ms. (Of course this is equally possible for theoretical mechanisms in physics, chemistry or biology. Chemical reactions, for example, can be modeled at a wide range of temporal grains, nowadays down to nanoseconds.)

When evaluating the empirical adequacy of computational models we compare the observable trace of subjects' behaviors when performing a task with the performance or trajectory of a model. Depending on the richness of the data available we must use criteria at different levels of resolution for the comparison (Wallach, 1998):

- *Product correspondence:* Similarity of the final performances (such as success in solving a problem or classes of problems) on a specified scale.
- *Correspondence of intermediate steps* toward problem solution (problem solving strategies, verbal statements, eye movements).
- *Temporal correspondence:* Latencies of S and M that fall into a comparable range, so that the temporal trajectories have a similar profile (Pylyshyn, 1984).
- *Error correspondence:* Comparability of the numbers and kinds of errors, on some specified scale.
- *Correspondence of context dependency:* Comparability in sensitivity to degradation by interfering contextual factors.
- *Learning correspondence:* Comparability in rate of improvement of performance with practice in the same learning environments.

These criteria are not mutually independent but can be combined for *convergent validation* of a model in the sense of Garner, Hake and Eriksen (1956). Of course, the question as to what qualifies as a good approximation has only pragmatic answers.

None of these questions is in any way unique to cognitive modeling; they arise with all deep models that contain theoretical terms, and which hence provide only indirect evidence for these unobservable entities. The procedures for

dealing with the problem are the same in all sciences: what we seek is a high level of goodness of fit relative to the number of degrees of freedom that are available for estimating free parameters. The fewer degrees of freedom, and the better the fit, the more reason we have for taking the model seriously until a better one comes along, or new evidence (possibly based on new methods of observation) undermines it.

To whatever extent and in whatever degree a model does not fit precisely, the specific pattern of deviation from the empirical data provides a powerful means to guide model revision. On the other hand, to the extent that the model matches the available empirical data, a useful next step is to test the model on related phenomena in order to explore the generality of the postulated mechanisms (Richman et al., 1995).

By altering the model in various respects, we can investigate the effects of experimental manipulations, and thereby derive empirical predictions for empirical studies. A computational model can thus give rise to new theoretically motivated experiments, and to reexamination of existing experimental data – further examples of the complementary relation between cognitive modeling and empirical studies[4].

There is a great need for more extensive and systematic comparison of alternative cognitive models capable of performing the same tasks. Such comparison should not be aimed at selecting "winners". Rather it should aim to explicate the central structures and mechanisms of alternative models, find functional identities among components, and compare their power to explain wide ranges of phenomena: "This involves laying bare the theory's principles and their entailments, showing how each .... in the context of its interactions with the others, increases empirical coverage, reduces tailorability, or improves the adequacy of the theory in some other way. With such developments, the theorist provides explanations of why the particular principles were chosen" (VanLehn, Brown & Greeno, 1984, 240f).

Comparison of alternative theories allows a discussion of underlying processes and mechanisms, for no matter how difficult it is to discover what is going on from moment to moment inside the mind of a human subject, there is no similar difficulty in finding out what is going on inside the computer program that is simulating the human performance. Its mechanisms can be ascertained by examining the code of the program, and its performance by tracing the program. The superiority of fit to the human data of one program relative to another can then be attributed with some accuracy to differences in their mechanisms and organization.

Analysis of the principles underlying a model is a powerful complement to empirical evaluation when judging its scientific contributions and relations to alternative models. It also includes sensitivity analysis to assess the contributions of model components to overall performance (Richman & Simon, 1989)[5]. Carpenter and Just (1999) have coined the term *cognitive lesioning* for an analysis of how withdrawing particular structures or resources affects the performance of a computational model.

If two models propose different sets of mechanisms, both sufficient to perform a task, comparative analysis may determine which mechanisms have better theoretical and empirical support.

Models in discovery

In considering the role of modeling in cognitive science, we should not limit ourselves to questions of theory verification. At least as important to a science as verification is the discovery of new phenomena and new theories to explain them.

An old recipe for rabbit stew begins with the instruction, "First catch a rabbit". In the same way, before we can explain phenomena, we must observe them, and before we can test a theory, we must create it. Both finding new phenomena and designing new theories are inductive activities involving exploratory problem solving. Attempting to model a phenomenon can suggest the need for novel mechanisms, and the incorporation of such mechanisms in the model can then suggest new empirical explorations to seek evidence of their presence in the human behavior that is being modeled.

De Groot (1946) and his students drew upon Selz's theories of problem solving to explain the phenomena of expert chess memory they had observed, and then invented a new experimental design (memory for random arrangements of pieces) to test predictions of the theory. The initial version of Feigenbaum's EPAM model of perception and memory was tested by searching the psychological literature of the first half of the 20th century for experimental phenomena that could verify or contradict the model's predictions (Feigenbaum, 1961).

Awareness is increasing today of the fruitfulness and significance of exploratory research, of both empirical and model-building varieties. Hence, any balanced account of methodology must give major attention to the constant interplay of observation and modeling in the discovery process, and cannot restrict itself to verification. Experiments may be undertaken in order to discover new phenomena as profitably as to test hypotheses. Models may be designed to explore the range of mechanisms needed to produce phenomena, even in the absence of current evidence that such mechanisms exist.

Perspectives of cognitive modeling

Cognitive modeling provides a powerful approach to developing complex and rigorously precise models of cognition with strong theoretical foundations. Cognitive modeling has greatly advanced our understanding of the mind and its symbolic processes, and has raised the bar of what we qualify as an exacting scientific explanation.

One of the main advantages that cognitive modeling offers is the potential for truly integrated approaches to cognition. Anderson, Bothell, Lebiere, and Matessa (1998) present an example for this unifying force with an ACT-R model (Anderson & Lebiere, 1998) that accounts for a large range of phenomena in the list-learning paradigm. This work is one

[4] It is generally argued that a model is especially strongly supported if effects that were predicted in experimental model manipulations are confirmed in subsequent empirical investigations (Kieras, 1985). For a more symmetric view of the relation between experiment and theory, see Simon (1955).

[5] For a functional decomposition of connectionist models see Schneider (1988).

valuable illustration of the scope and practicality of Newell's plea (1973) for a broad computational theory that can tie together and relate disparate bodies of data with precise hypotheses about underlying structures and mechanisms.

Ohlsson (1988) emphasizes that cognitive modeling enforces comprehensive models that view human beings as complete agents: "There cannot be a theory of problem solving, or of memory, or of perception, because such a theory cannot be tested against data: only complete systems, which have the entire range of capabilities (albeit, perhaps in simplified form), and which show explicitly the interactions between them, can be meaningfully compared against data" (Ohlsson, 1988, p. 12).

One might demur that this claim is somewhat extreme in that, as in the other sciences, particular sets of phenomena can usually be isolated to a degree that permits them to be studied in at least temporary isolation from others[6]. However, the drive towards completeness can reveal gaps in our understanding of the mind and thereby reveal the need for well-directed empirical studies, and new instruments and observational methods for conducting them. However far short we are today of having "unified theories of cognition" (Newell, 1990) in any comprehensive sense, the past 20 years have nevertheless produced half a dozen cognitive models that can claim to unify important subcontinents of the entire cognitive world.

The objection against cognitive modeling heard most often is that it involves many parameters and thus many degrees of freedom in constructing a model. Of course alternative approaches to the study of the mind face the same problem, simply concealing it by lower standards of precision. In the final event, the complexity of an empirical theory must be determined by the complexity of the phenomena it seeks to explain. Molecular genetics does not have the simplicity of classical mechanics.

Cognitive modeling typically uses rich and dense streams of empirical data to constrain and test models. Our confidence in a model can grow with the increase in the ratio of the number of data points explained to the number of degrees of freedom in the model (cf. Simon, 1998). Cognitive architectures severely restrict their degrees of freedom when they apply the same core set of justified constructs across an entire range of phenomena, a fundamentally important goal of the drive toward unified theories.

Computational models of cognition have accomplished a major advance in our understanding of the mind over the past four decades. This special issue of *Kognitionswissenschaft* presents four papers on cognitive modeling that sample the scope and variety of modeling approaches in the field. It is our hope that the exciting work reported in them will help to pave the way for an even much wider attention to and application of the cognitive modeling methodology.

References

Anderson, J.R. (1987). Methodologies for the study of human knowledge. *Behavioral and Brain Sciences*, *10*, 467-505.

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

Anderson, J.R., Bothell, D., Lebiere, C. & Matessa, M. (1998). An integrated theory of list memory. *Journal of Memory and Language*, *38*, 341-380.

Carpenter, P.A. & Just, M.A. (1999). Computational modeling of high-level cognition versus hypothesis testing. In: R.J. Sternberg (ed.), *The nature of cognition* (pp. 245-294). Cambridge, MA: MIT Press.

De Groot, A.D. (1946). *Het Denken van den Schaker*. Amsterdam, Neth: North Holland Publishing Company.

Ericsson, K.A. & Simon, H.A. (1993). *Protocol analysis* (2nd edn). Cambridge, MA: MIT Press.

Feigenbaum, E. (1961). The simulation of verbal learning behavior. *Proceedings of the Western Joint Computer Conference*, *19*, 121-132.

Fridja, N.H. (1967). Problems of computer simulation. *Behavioral Science*, *12*, 59-67.

Garner, W.R., Hake, H.W. & Eriksen, C.W. (1956). Operationism and the concept of perception. *Psychological Review*, *63*, 149-159.

Just, M.A. & Carpenter, P.A. (1985). Cognitive coordinate systems: accounts of mental rotation and individual differences in spatial ability. *Psychological Review*, *92*, 137-172.

Kieras, D. (1985). The why, when, and how of cognitive simulation: a tutorial. *Behavior Research Methods, Instruments & Computers*, *17 (2)*, 279-285.

Miller, G.A., Galanter, E. & Pribram, K.H. (1960). *Plans and the structure of behavior*. New York: Holt, Rinehart & Winston.

Moore, E.F. (1956). Gedanken-experiments on sequential machines. In: C.E. Shannon & J. McCarthy (eds.), *Automata studies* (pp. 129-153). Princeton, NJ: Princeton University Press.

Moravsik, J.M. (1980). Chomsky's radical break with modern tradition. *Behavioral and Brain Sciences*, *3*, 28-29.

Newell, A. (1973). You can't play twenty question with nature and win (pp. 283-310). In: W.C. Chase (ed.). *Visual information processing*. New York: Academic.

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

Newell, A. & Simon, H.A. (1972). *Human problem solving*. Englewood Cliffs, NJ: Prentice-Hall.

Newell, A. & Simon, H.A. (1976). Computer science as empirical inquiry: Symbols and search. *Communications of the ACM*, *19 (3)*, 113-126.

Newell, A., Shaw, J.C. & Simon, H.A. (1958). Elements of a theory of human problem solving. *Psychological Review*, *65*, 151-166.

Ohlsson, S. (1988). Computer simulation and its impact on educational research and practice. *International Journal of Educational Research*, 12, 5-34.

Pylyshyn, Z. (1984). Computation and cognition. Cambridge, MA: Bradford Books.

Richman, H.B. & Simon, H.A. (1989). Context effects in letter perception: comparison of two theories. *Psychological Review*, *96*, 417-432.

Richman, H.B., Staszewski, J.J. & Simon, H.A. (1995). Simulation of expert memory using EPAM IV. *Psychological Review*, *102*, 305-330.

Schneider, W. (1988). Sensitivity analysis in connectionist modeling. *Behavior Research Methods, Instruments & Computers*, *20 (2)*, 282-288.

Simon, H.A. (1955). Prediction and hindsight as confirmatory evidence. *Philosophy of Science*, *22*, 227-230.

Simon, H.A. (1977) Identifiability and the status of theoretical terms. In: R.E. Butts & J. Hintikka (eds.), *Basic problems in methodology and linguistics*. Dordrecht, Neth: D. Reidel Publishing Company.

Simon, H.A. (1996). *The Sciences of the Artificial* (3rd edn). Cambridge, MA: MIT Press.

Simon, H.A. (1998). What is an "explanation" of behavior? In: P. Thagard (ed.). *Mind Readings* (pp. 1-28). Cambridge, MA: MIT Press.

Tack, W.H. (1997). Kognitionswissenschaft: Eine Interdisziplin. *Kognitionswissenschaft*, *6*, 2-8.

VanLehn, K., Brown, J.S. & Greeno, J. (1984). Competitive argumentation in computational theories of cognition. In: W. Kintsch, J.R. Miller and P.G. Polson (eds.), *Methods and tactics in cognitive science* (pp. 235-262). Hillsdale, NJ: Erlbaum.

Wallach, D. (1998). *Komplexe Regelungsprozesse*. Wiesbaden: DUV.

[6] See Simon (1997) on nearly-decomposable systems.