1

What Is an "Explanation" of Behavior?

Herbert A. Simon

The cognitive "revolution" in psychology introduced a new concept of explanation and somewhat novel methods of gathering and interpreting evidence. These innovations assume that it is essential to explain complex phenomena at several levels, symbolic as well as physiological; complementary, not competitive. As with the other sciences, such complementarity makes possible a comprehensive and unified experimental psychology. Contemporary cognitive psychology also introduced complementarity of another kind, drawing upon, and drawing together, both the behaviorist and the Gestalt traditions.

I would like to begin with two comments on contemporary cognitive psychology—on where we stand. The evidence supporting these observations is so overwhelming that I will not bore you by rehearsing it. But we have some conventional, customary ways of talking about psychology that fly in the face of what I think are the facts, and I would like to distance myself from these ways of speaking, which I believe are harmful to the continued rapid progress of our science.

How often have you heard that "some day we will understand the mind," or that "the human brain is a great mystery that we must seek to solve"? In fact, psychology exists not in the future, but in the present. By any reasonable metric, we know more about the human mind and brain than geophysicists know about the plate tectonics that move the continents over the globe, far more than particle physicists know about elementary particles, or biologists about the processes that transform a fertilized egg into a complex multicellular organism.

We discount our knowledge because some of it is so commonplace, so familiar from our everyday acquaintance with ourselves and other people. We discount it also because it often is insufficient to permit predictions of behavior in important matters that concern us. The former is a great blessing to us, for it allows us to learn easily facts of sorts that other sciences have to tease out with great effort. The

latter is a true limitation that we share with meteorologists, evolutionary biologists, and all those physical or biological scientists who venture outside the laboratory into the complexity of real-world phenomena.

Hence, my first comment: In the year 1991, we know a great deal about human thinking, and especially about the symbolic processes, involving selective heuristic search and recognition of familiar cues, that people use to solve problems, to design artifacts and strategies, to make decisions, to communicate in natural language, and to learn. How people solve problems is no great mystery; we know enough about it to create computer programs that do it, and do it in a way that closely simulates human performance, step by step. By the same test, we know how people design strategies, and even how they learn language and make scientific discoveries.

In all these cases, we have examples of computer programs that perform these tasks in humanoid ways. If you want evidence for this claim, I can refer you, for starters, to standard sources like Anderson's (1990b) cognitive psychology textbook or the recent *Foundations of Cognitive Science*, edited by Posner (1989). Cognitive psychology is not some dream of the future; it exists, and it allows us to explain a vast range of phenomena. It is not a finished science, thank goodness (what science is?), but each year adds to its store of knowledge and understanding, and its powers of prediction.

My second comment: Histories of psychology are fond of talking about "schools of thought," and their rise and fall, attributing to the chronology of our field a circular course, rather than the helical one (at worst) attributed to other sciences. In the histories there is not just psychology, the science of human behavior; there is introspectionist psychology, and behaviorist psychology, and Gestalt psychology, and information processing psychology, and connectionist psychology—schools without end, and without cumulation, each school combating and destroying the previous one, to be consumed, in turn, by its successor.

This circular view of history is wholly counterfactual. The "cognitive revolution" (I even used the phrase in my opening summary) did not destroy either behaviorism or Gestalt psychology. It drew liberally upon both of them, both for experimental data and for concepts. The productions of information processing psychology are

natural descendants of the familiar stimulus-response links of behaviorism (though not identical with them). Means–ends analysis, central to information processing theories of problem solving, was explored by Duncker (1945), and by Selz (1913) before him. The neural nets of current connectionist models have their origins in physiological probings of the nervous system, via the "cell assemblies" of Hebb (1949), and in notions, traceable back to Aristotle, of the associative structure of memory.

In the course of this paper, I refer to another currently fashionable novelty in psychology, so-called situated action and situated learning, and show that its antecedents are also very familiar. Psychology is as progressive and cumulative as any of the sciences, and we can today cite experiments of Ebbinghaus (1964), or Wundt (1902), or Hovland (1951), or Skinner (1938) as major sources of empirical support for contemporary theories.[1]

In our generation, we have discovered a mode of psychological theorizing that has greatly facilitated, and will continue to facilitate, the cumulation of knowledge and theory in psychology. Today we build computer models of both symbolic and connectionist systems. Instead of constructing microtheories for each phenomenon we observe (e.g., theories of retrospective inhibition), or macrostatements that are too simplistic and general to explain much (e.g., "forgetting follows a power law"), we construct computer programs that can be given complex cognitive tasks, identical to those given to our human subjects, and that will predict the temporal path of human behavior on those tasks (Newell and Simon, 1972).

Some cognitive psychologists today aspire to build "unified" models of this kind: The SOAR (Newell, 1990), ACT* (Anderson, 1983), and PDP (Rumelhart and McClelland, 1986) systems are familiar examples. Others of us aim at models of middle range: a GPS (Newell and Simon, 1972) to account for problem-solving phenomena; an EPAM (Feigenbaum and Simon, 1984) to account for verbal learning processes; an ISAAC (Novak, 1977) to explain how people understand problems described in natural language text, construct mental representations of those problems, and go on to solve them; and an INTERNIST (Miller, Pople, and Myers, 1982) or a MYCIN (Shortliffe, 1976) to describe the processes of expert medical diagnosis.

Whether comprehensive or not, such models enable cognitive psychology to organize large bodies of data around the mechanisms

that produced them; and the availability and relevance of these large bodies of data provide powerful means to test adequacy of the models. This tying together and relating of disparate bodies of experimental data with hypotheses about the causal mechanisms greatly facilitates cumulation.

So much for these two debilitating myths: that the mind is something we will understand in the future and that the path of psychology is circular, each new "school" tearing down and replacing the one it succeeds. Neither myth bears the slightest resemblance to the true state of affairs, and it is time that we put them to rest and get on with advancing still further a science that has made great strides in this century.

Explaining a Conversation

The scene is a street in Singapore. A woman is talking to two other women, talking in Tamil, a Dravidian language that is spoken in a large region of southern India around Madras and in parts of Sri Lanka. We wish to explain her behavior.

What is there to explain? For one thing, why Tamil? Why not English, or Chinese, or Malayan, the predominant languages in Singapore? An explanation would describe the migrations that brought large numbers of Tamils from India to this distant port. This "simple" explanation still presupposes some vital theoretical underpinnings. It assumes that under some circumstances migrants will retain, for a generation or even beyond, the language of their ancestors. What are those circumstances, and what conditions in Singapore satisfied them? And when will this woman, probably multilingual, use Tamil, and when one of the other languages of Singapore?

The explanation by migration also assumes conditions that caused the migrants to leave their homeland, and historical "laws" that would explain migration as a response to such conditions. What were those conditions, and what is the nature of such laws?

Some social psychologists undertake to answer questions like these. For the rest, these questions are usually left to history, sociology, and the other social sciences. But insofar as they involve things stored in the human memory, they are also part of cognitive psychology. It is proper that they be welcomed back into our science, as

is being done by those who are now focusing on the psychology of everyday life.

The Structure of a Dynamic Explanation

Even answers to all of these questions will only begin to explain our Tamil woman's behavior, but before continuing, let us ask what has already been revealed.

Our explanation has the form of a fugue, with two intertwined themes. First, to explain an event, we refer to antecedent events—initial conditions. To explain a Tamil-speaking woman's presence in Singapore, we find a migration from Madras. But that poses the new question of explaining the presence of Tamil-speaking people in southern India. So explanation by antecedent events takes us back to the explanation of those antecedents. If the data were available (they are not), they could take us step by step in an almost infinite regress to the cosmological Big Bang and beyond.

But—the second theme of the fugue—explanation by antecedent events also requires general laws to explain how each situation causes the succeeding one. What causes of migration can take people from one land to another? What laws determine the language that a person will speak in an ethnically foreign land, and when?

The natural sciences commonly employ this fuguelike structure of explanation. The differential equations of physics describe mechanisms that determine the next movements of the stars and planets, given the initial conditions: their present positions and velocities. The laws of genetics and Darwinian selection explain how a community of organisms (the initial conditions) evolves over years or millennia into a new and different community.

For systems that change through time, explanation takes this standard form: Laws acting on the current state of the system produce a new state—endlessly. Such explanations can be formalized as systems of differential equations or difference equations.

Explaining by Simulation

We return to our Tamil women, whom we left talking on the street. To understand their conversation, we must have some knowledge of the lexicon and syntax of their language. Tamil is one of about 20 highly inflected Dravidian languages spoken throughout southern

India. To characterize its syntax, we build a computer program that parses the speaker's sentences. Such a parser is also a set of difference equations, playing the same role as the differential equations in physics.

But we might go even further in explaining Tamil. We might build a diachronic story—conceptually, another set of difference equations—to explain how the contemporary Dravidian languages evolved from some common ancestral base. This means postulating laws of linguistic transformation that cause language evolution. Since Chomsky's revolution, or even since Grimm's, explanation in linguistics has become another exercise in building and testing difference equations (Chomsky, 1957).

Another approach to these questions is to write computer programs that are capable of using and understanding, even learning, language. A computer program is (literally, not metaphorically) a system of difference equations. For each possible state of the computer, combined with the input at that instant, the program determines the next state of the computer. The computer's memory holds the initial conditions (the current state) and the laws of behavior (the program). Its input devices convey to it the external stimulus, which may, as in the case before us, take the form of sentences in a natural language.

Since a computer program is a system of difference equations, a properly programmed computer can be used to explain the behavior of the dynamic system that it simulates. Theories can be stated as computer programs.

Controlled experiments can be performed on computer programs, altering specified program components to determine how such changes affect the performance of tasks. The architecture can thereby be modified to simulate the human performance better.

There is no epistemological difference between using a program incorporating Newton's laws to explain the movements of Mars and using a program incorporating linguistic laws to explain how speech is generated or understood. But perhaps you are not familiar with the computer programs that have these linguistic capabilities. One example is ISAAC, written by Novak (1977), which reads the English language statements of problems in physics textbooks, forms internal representations ("mental pictures") of the problem situations, and then proceeds to derive the applicable equations and to solve them.

Another such program is ZBIE, written by Siklóssy (1972), which reads a simple sentence in a natural language at the same time it inputs a diagrammatic representation of the scene described by the sentence (e.g., "The dog chases the cat."). ZBIE learns the meanings of the words in the sentences it reads (i.e., learns what objects or relations in the diagrams the words denote) and analyzes their grammatical structure. When it is later confronted with a scene it has not seen before, but one composed of familiar kinds of objects in familiar relations, it constructs an appropriate and grammatically correct sentence to describe the scene.[2]

A remarkable feature of programs like ZBIE is that they not only explain how natural language is understood, they also understand it. The linguistic symbols are not translated into an esoteric formal language; hence, we do not have to numericize or otherwise encode the sentences whose production or understanding we wish to explain. The programs use symbol structures that are isomorphic to those the human subject uses. All information processing theories of cognition have this property: They actually perform the tasks whose performance they explain.

Programs that simulate cognitive processes describe the processes in symbolic languages isomorphic to those being modeled, and hence, actually execute the processes. Consequently, they provide a rigorous test of the sufficiency of the hypothesized processes to perform the tasks of interest.

Neurophysiological Explanation

Simulating language behavior with a computer teaches us the properties an architecture must possess if it is to speak and listen, and what processes are employed by its program. It allows us to test, at the level of symbolic behavior, how closely these processes resemble those of human speakers or listeners. It does not tell us how the same structural conditions and programs are realized by the biological components known as neurons and the assemblages of components that make up the biological brain.

Explanation of cognitive processes at the information processing (symbolic) level is largely independent of explanation at the

physiological (neurological) level that shows how the processes are implemented.

There is nothing mysterious about explaining phenomena at different levels of resolution. It happens all the time in the physical and biological sciences. A theory of genetics need not (fortunately) rely on a knowledge of quarks. As a matter of history, the former theory preceded the latter by many years. The theory of genetic processes was developed by Mendel, using genes as abstract primitive "atoms." Fifty years later, a microscopic foundation was provided for the theory by locating the genes in visible chromosomes. After another half century, the structure of chromosomes was elucidated in terms of the combinatorics of DNA, strands of four complex submolecules, nucleotides. Two levels of reduction and still no quarks! And no need of them, although we surely believe that nucleotides are made of atoms, which are made of neutrons and protons, which are made of quarks.

Explanation on different levels does not deny the possibility of reduction. Higher level theories use aggregates of the constructs at lower levels to provide parsimonious explanations of phenomena without explicit reference to the microconstructs. The lower level details do not show through to the higher level.

Of course, the higher level mechanisms are reducible to those of the lower level (at least in principle, although the computations can actually be carried out only in the simplest situations). But we do not require the reduction in order to explain the aggregate events at the higher level. We can write the system of difference equations for this higher level independently of any lower level explanation. Cognitive psychology (fortunately) does not have to stand still with breathless expectation until neurophysiology completes its work. As cognitive psychology has been doing, it can proceed with its task of explaining thought processes at the level of symbol systems.

Partitioning explanation into levels also points to a strategy for neurophysiological research. Neuropsychology has two main tasks. It must explain electrochemically how neurons and simple organizations of neurons store and transmit information. It must also help build the bridge theory that shows how the symbol structures and symbol-manipulating processes that handle information at a more aggregated

level can be implemented by such neuronal structures and organizations. The bridge need not be built solely from one bank of the river; it can be constructed by cooperative effort of information processing psychologists with neuropsychologists. But if they are to cooperate, they must learn to read each other's blueprints.

This strategy relieves neuropsychology of the heroic, but impossible, task of climbing in a single step from neurons and nerve nets up to complex human behavior without inserting intermediate strata into the structure. Some neuropsychologists and connectionists do not yet accept the need for higher level aggregate theories, or the meaning of information processing programs as examples of such theories. Such misunderstanding forms a serious barrier to collaboration.

Nowadays, a discussion of neurophysiology necessarily raises the question of whether mental functions are to be modeled as parallel or serial systems. At the lowest level, the individual neuron demonstrably transmits signals longitudinally, in serial fashion. At the next level up, brain tissue forms a network of elements operating in parallel, and the same can be said of the eyes and ears. At the level of conscious reportable events, the bottleneck of attention and short-term memory again gives the mind the characteristics of a serial organization. It is worth pondering that the low-level anatomy of the conventional von Neumann "serial" computer looks every bit as parallel as a neural network; yet at the more aggregate, symbolic, level, it executes its processes sequentially, one or a few at a time.

From these observations, we can conclude, first, that at the level of the network of neurons, modeling will have to be largely parallel. It is not clear, as yet, how far we can abstract from the details of neural structure in our models, or how many structures the models will have to contain to simulate relevant events at this level.

Second, at the symbolic level—the level of events taking place in hundreds of milliseconds or more—modeling will continue to be largely serial, for the mind behaves like a serial system wherever the bottleneck of attention supervenes upon events. While most people can probably chew gum and walk at the same time, very few can carry on a technical conversation while maneuvering a car through heavy traffic.

Third, at the intermediate level of events milliseconds or tens of milliseconds in duration, the comparative advantages of parallel and

serial modeling are not yet clear. This is the level of the EPAM program (Feigenbaum and Simon, 1984), which simulates learning and perception at the symbolic level, and the level of most connectionist systems. It is also the foundation level of SOAR (Newell, 1990), a unified control structure for cognition. Teasing out the respective roles of parallel and serial processors and their interface at or near this level is a major contemporary task for cognitive research.

Static Theories

Concern with architecture reminds us that not all theories take the form of difference equations. In fact, theories in psychology have traditionally had a quite different form. Typically, they make assertions such as "If the independent variable, x, increases, the dependent variable, y, will also increase." Laws of this form are very weak. They are also merely descriptive, not explanatory.

Much stronger claims are made by laws of the form "$y = 80x + 300$," where the parameters, 80 and 300, were known or estimated prior to the current experiment. If, in addition, these parameters describe structural characteristics of the system (e.g., the speed at which it can store or access information), then the law begins to explain as well as to describe. Let us call laws of this kind, with the numerical parameters taken seriously, *models*.

For example, Baddeley (1981) showed that the contents of short-term memory can be retained for only about 2 s without overt or covert rehearsal. This finding implies that the maximum capacity of short-term memory is whatever content can be rehearsed in this time. Other experiments have shown that it takes about 300 ms to recover a familiar "chunk" (e.g., a familiar word or phrase) from long-term memory, and about 80 ms per syllable to pronounce it. From these facts, there follows the law: $2{,}000 = 300C + 80S$, where C and S are the numbers of chunks and syllables, respectively, in the longest strings that can be retained in short-term memory. The law can be tested using the standard immediate recall paradigm (Zhang and Simon, 1985).

Some of the properties of systems can be captured in static laws, preferably models, which specify the relations among variables, qualitatively or numerically.

Explaining Thinking

Our Tamil women are still talking on the street in Singapore. So far, we do not know what they are saying. When we eavesdrop, we find that the speaker is explaining to her companions how to solve the Tower of Hanoi puzzle![3]

By now, we know exactly how to theorize about this kind of behavior. We construct a set of difference equations (a computer program in a symbol-processing language) that simulates human behavior in solving the Tower of Hanoi problem. In fact, programs of this kind have existed for some years (Simon, 1975). Notice that I refer to "programs" in the plural, for different people may solve the problem in different ways, using different strategies.

Heuristic Search

Heuristic search is too familiar to require lengthy description. Common to virtually all of the problem-solving strategies that people have been observed to use is a problem space and a search through this space until a solution is reached (Newell and Simon, 1972). The moves that change one situation into another in the Tower of Hanoi may be legal moves, as defined by the problem instructions, or they may be "wished-for" moves that change the current situation into a distant one in one step.

In some strategies, most of the problem solving takes place in the head, making use of symbolized goals and *mental models*, symbol structures describing the situation at each stage of the search. In other strategies, the subjects work directly from the physical Tower of Hanoi puzzle in front of them, using visual perception of the current arrangement of the disks to calculate a next move, and recording it by actually moving the disk. In currently fashionable terminology, the subjects who use the latter strategies are engaging in *situated action.*

There is a good deal of debate at present (under the rubric of situated action) as to whether problem solving requires the subject to create a mental problem space and to search in that space, or whether the search can be almost wholly external, with no significant problem representation in the head (Suchman, 1987; Winograd and Flores, 1986). Sometimes the debate is enlarged by challenging whether problem solving can be modeled at all by symbolic systems.

The best way to resolve the debate is to construct programs and observe what they can and cannot do. A running program is the moment of truth. This particular debate has been largely resolved by programs already written and tested. Some strategies that have been written for the Tower of Hanoi depend on search through an internal representation of the problem, or even initial search through an abstracted representation to find a plan for the more detailed search. Other strategies that have been written search externally, representing internally only the "affordances" provided by the external objects and their relations (Simon, 1975). Hence, it has been demonstrated constructively that both situated action and strategies requiring planning and internal representations are realizable by symbol-processing systems.

What has not been settled, and cannot be settled without extensive empirical study, is the extent to which, and the circumstances under which, human beings will use one or another kind of strategy. Our Tamil woman is not carrying a physical Tower of Hanoi puzzle with her. She has no alternative, if she is to explain the solution to her friends, but to form a mental representation of some sort—a problem space—and to describe the moves in that space. Her friends have no alternative for understanding the explanation but to translate the description into their own mental representations. If a physical Tower of Hanoi puzzle were present, matters might be quite different. But life does offer us a great deal of variety. So much for situated action.

Different people, or the same people in different situations, can employ different strategies for performing a given task. A theory of their performance would include a computer program describing the strategy they are using in a given instance together with a specification of the circumstance under which this particular strategy will be used. The specification can include a variety of elements, including the subjects' previous experience and learning.

Expert Behavior

Actually, I was joking about the Tower of Hanoi. That is not what the Tamil women are talking about at all. In fact, the speaker is telling about a new recipe she has learned; her friends regard her as an expert in preparing gourmet meals.

The conversation is not a monologue. The expert does most of the talking, but her friends ask frequent questions, and she usually replies promptly. One of them asks how long the dish should remain in the oven. The expert answers, then says, "Of course, I don't have any systematic rules for determining such things. I just use my intuition. It's all a matter of experience."

The expert has just stated, very succinctly, the theory of expert performance that has emerged in recent years from psychological research and modeling. In everyday speech, we use the word *intuition* to describe a problem-solving or question-answering performance that is speedy and for which the expert is unable to describe in detail the reasoning or other process that produced the answer. The situation has provided a cue; this cue has given the expert access to information stored in memory, and the information provides the answer. Intuition is nothing more and nothing less than recognition.

We do not have conscious access to the processes that allow us to recognize a familiar object or person. We recognize our friend, but we do not know what traits and features, what cues, enable that recognition to occur. Nor can we describe these traits and features to other people accurately enough to enable them to recognize the same person. We are aware of the *fact* of recognition, which gives us access to our knowledge about our friend; we are not aware of the *processes* that accomplish the recognition.

The process of recognition (i.e., intuition) is readily realized in computer programs by means of so-called *productions*. A production is an (if → then), or (condition → action), statement that, at least superficially, resembles a (stimulus → response) linkage. For our present purposes, we need note only that, while the stimuli of classical behaviorism are in the environment, not in the head, the conditions that have to be satisfied to trigger the action of a production may be (but need not be) symbol structures held in memory. Productions can implement either situated action or internally planned action, or a mixture of these.

Quite general programming languages (e.g., the language OPS5; Brownston, Farrell, and Martin, 1985) can be constructed entirely of productions. The execution of a production can be made to depend on a context by including among the conditions for execution one or more goal symbols. The production will then be activated only in

contexts where the appropriate goal is present. Conditions can also reflect other elements of contexts besides goals.

Consider a (simplified) expert modeled as a production system. Cues in the environment that the expert encounters trigger information in memory, hence, initiate actions appropriate to the situations marked by these cues. In its simplest form, the model produces situated action.

When the doctor notices some symptoms, a diagnosis is triggered, or, alternatively, information that is accessed indicates certain additional tests should be performed to reach a definitive diagnosis (a departure from pure situated action). When the doctor has reached a diagnosis, another production accesses information in memory about the prognosis and about appropriate courses of treatment.

Information organized in a production system of this kind—a sort of indexed encyclopedia—can produce expert behavior. Expert systems may, in addition, have some capabilities for means-ends analysis or other forms of reasoning and heuristic search, but at their core is a production system capable of recognizing appropriate cues, hence, capable of acting intuitively.

There is no incompatibility between intuition and analysis. A chess master in a tournament does a good deal of analysis, of look-ahead to possible continuations of the game. The same chess master, playing simultaneously a number of weaker players, moves quickly, hardly analyzing ahead at all but selecting moves almost wholly on intuition in the form of recognition of weaknesses created by the opponents. This rapid play is weaker than the more analytic play of the tournament, but only a little weaker.

A large part of the chess master's expertise lies in his or her intuitive (recognition) capabilities, based, in turn, on large amounts of stored and indexed knowledge derived from training and experience. Under the conditions of rapid play, the chess master's behavior is a form of situated action; under tournament conditions, it is more planful.

Similarly, our expert Tamil gourmet, after a quick inventory of her refrigerator and kitchen cabinet, can rustle up a presentable and tasty meal in a hurry, relying on intuition—experience encapsulated in memory and evoked by the sight of familiar items of food. Of

course, given some time to plan and prepare, she can usually produce an even more delicious meal.

The core of an expert system, in human or computer, is a system of productions that operates like an indexed encyclopedia. Cues in the situation (external or imagined) are recognized by the conditions of productions, triggering the actions associated with these conditions. The case in which the cues are predominantly external is sometimes called situated action.

The production system of an expert is generally associated also with reasoning (search) capabilities that support an adaptive system of analytic and intuitive responses.

Adaptivity of Behavior

The human mind is an adaptive system. It chooses behaviors in the light of its goals, and as appropriate to the particular context in which it is working. Moreover, it can store new knowledge and skills that will help it attain its goals more effectively tomorrow than yesterday: It can learn.

As a consequence of the mind's capacities for adaptation and learning, human behavior is highly flexible and variable, altered by both circumstances and experience.

Scientific laws, whether descriptive or explanatory, are supposed to capture the invariants of the phenomena, those underlying regularities that do not change from moment to moment. How does one find laws to describe or explain the behavior of an adaptive system?

The shape of a gelatin dessert cannot be predicted from the properties of gelatin, but only from the shape of the mold into which it was poured. If people were perfectly adaptable, psychology would need only to study the environments in which behavior takes place. Some of this viewpoint is reflected in the *affordances* of Gibson's (1979) theories of perception, and in the rational adaptation models of my colleague Anderson (1990a, 1991).

In its extreme form, this position eliminates the need to run laboratory experiments or to observe people. Merely examine the shape of the mold: Analyze the environment in which the behavior is to take place and the goals of the actor, and from these deduce

logically and mathematically what the optimal behavior (and hence the actual behavior) must be.

Nowhere has this method of explaining human behavior been carried further than in modern neoclassical economics. The neoclassical theories also show the severe limits of the approach. First, the scheme works only if the actor's goals and the alternative behaviors available for choice are known in advance. Change either the goals or the alternatives and the optimal decision may change (Simon, 1991). Do we think that we can predict what the menu will be in the Singapore apartment tonight without knowing what is in the refrigerator, or what some of our gourmet's favorite recipes are? Can we predict it from a book on nutritionally optimal diets?

In most real choice situations, there is a multiplicity of goals, often partly conflicting and even incommensurable. A simple example is the trade-off between speed and accuracy: Unless we know their relative importance, we cannot select an optimal behavior.

Nor are the alternatives from which the actor might choose usually known in advance (even to the experimenter). Human beings spend much of their time inventing or discovering actions that fit the circumstances. The whole vast collection of human activities known as design—whether in architecture or engineering, or painting, or management—is aimed at synthesizing appropriate actions. In explaining or predicting behavior, whether optimal or not, we must know not only the design product (the alternative finally chosen) but the design process as well (Simon, 1981, chaps. 5 and 6).

The process of design is highly dependent on history and experience. Before Newton, designers did not use the calculus, and undoubtedly reached different solutions than in later ages when the calculus was available. So choice is always relativized to the current state of knowledge, and inventing new alternatives or even new processes for generating alternatives is very different from choosing among available and known alternatives.

Design does not aim at optimization. Almost always, the process must be halted and a solution selected long before all alternatives have been generated and compared. Even the idea of generating "all" alternatives is usually chimerical. Limits on human (and computer) calculation and incomplete information foreclose finding the best: Most often, a stop rule halts the search when a satisfactory alternative

is found—one that meets a variety of criteria but maximizes none. So we should not expect the recipes of our expert gourmet cook to be optimal; but if she invites us to dinner at her Singapore home, the meal will be delicious; it will "satisfice."

The nonoptimality of behavior is obvious even in the simple Tower of Hanoi task discussed earlier. Many different strategies can be used to solve the problem; and even in identical circumstances, different subjects use different strategies, not all of which can be optimal. There is substantial empirical evidence that subjects also adopt a wide range of strategies, most of them suboptimal, in solving cryptarithmetic problems (Newell and Simon, 1972).

In complex adaptive behavior, the link between goals and environment is mediated by strategies and knowledge discovered or learned by the actor. Behavior cannot be predicted from optimality criteria alone without information about the strategies and knowledge agents possess and their capabilities for augmenting strategies and knowledge by discovery or instruction.

What constitutes an available alternative depends on the capabilities of the actor: such things as visual acuity, strength, short-term memory, reaction times, and speed and limits of computation and reasoning—to say nothing of expertise based on stored knowledge and skill. Before the exercise of optimizing can be carried out, all of these side conditions must be nailed down: goals, knowledge of immediately available alternatives, means for generating new alternatives, knowledge for predicting the outcomes these alternatives will produce, and limits on the ability of the actor to hold information in memory and to calculate.

The predictions of an optimizing theory depend as much on the postulated side conditions as on the optimization assumption. In fact, in most cases, if the correct side conditions are foreseen and predicted, the behavior can usually be predicted without any strict assumption of optimality; the postulate that people satisfice, look for "good enough" answers, is usually adequate to anticipate behaviors.

There is no way to determine a priori, without empirical study of behavior, what side conditions govern behavior in different circumstances. Hence, the study of the behavior of an adaptive system like the human mind is not a logical study of optimization but an empirical study of the side conditions that place limits on the approach

to the optimum. Here is where we must look for the invariants of an adaptive system like the mind.

But does the point need to be belabored? Optimization is an ideal that can be realized only in (a) extremely simple worlds (if offered the choice, take a $10 bill in preference to a $1 bill) and (b) worlds having strong and simple mathematical structures that admit the computations required for optimization (e.g., worlds that can be described in terms of a linear objective function and linear constraints, so that solutions can be found by linear programming algorithms). These are not the worlds in which most human life is lived.

We would not think of trying to predict where the moon will be at midnight tomorrow night without knowing where it is tonight. In the same way, we should not presume to predict how a human being will solve a problem or learn a new skill without knowing what that human being already has stored in memory by way of relevant information and skills. Changing the information and skills will change the behavior. This principle is the basis for all of the differences observed between experts and novices.

To some extent, we can finesse this requirement for our research by restricting our study to the ubiquitous college sophomore, assuming that all college sophomores know roughly the same things, at least those that are relevant to the mainly contentless tasks we confront them with. When we want to go further to study individual differences in task performance or to study the effects of previous knowledge and skill on performance, we must face up to the boundary conditions outlined above.

Cognitive and Social Psychology

Since adaptive behavior is a function of strategies and knowledge, both largely acquired from the social environment, there can be no sharp boundary between cognitive psychology and social psychology. The context in which knowledge is acquired and used, an exogenous variable in cognitive psychology, provides the endogenous variables for social psychology and sociology.

Studying expert behavior immediately begins to dissolve the boundary between cognitive psychology, on the one side, and social

psychology (to say nothing of social and intellectual history), on the other. It is not an accident that histories of science provide an important part of the data used to test cognitive theories of scientific discovery (Langley et al., 1987). The histories do not draw a boundary around individual investigators, but encompass the sources of an investigator's knowledge and, more broadly, the social processes that direct the production of scientific knowledge and its communication.

But we have already seen this point illustrated in the simple interaction among the Tamil women—their choice of language, their very presence in Singapore, the influence of their experience (itself a product of social environment) on what they can do and like to do.

As another example of this intermingling of the social with the cognitive, communication between different communities of experts involves translation, that is, understanding by members of one group of the language and concepts of the other. As Voss and his associates have shown, we can study one aspect of this phenomenon by observing how experts from different communities attack the same problem in quite different ways (Voss, Tyler, and Vengo, 1983). Another aspect, not yet much studied, would tell us how experts learn to translate from foreign dialects.

The flow between cognitive and social runs in both directions. Social psychologists have long been interested in how people form beliefs, or models, about other persons. Theories of person perception need to be integrated with cognitive theories about knowledge acquisition and formation of representations. There is no a priori reason to suppose that different processes are involved in the two cases.

Divide and Conquer

In trying to understand the behavior of three women on a street in Singapore, we have already set a dizzying array of tasks for psychology: to explain the migrations of peoples; the origins and changes in their languages; their development as individuals in society; their gradual acquisition of values, skills (including skills of social interaction), knowledge, and attitudes; the adaptation of their behavior to their goals; and the physiological underpinnings of all of these

processes. It appears that we are going to have to build computer programs, systems of difference equations, of immense complexity to explain such behavior.

Forms of Subdivision

Fortunately, we do not have to explain everything at once, or within the boundaries of a single program. We have already seen that complex phenomena can usually be segmented into levels from macroscopic to microscopic, separated by both the spatial and the temporal scales of the events they describe. Provided that the phenomena are roughly hierarchical in structure, as most natural phenomena are, we can build explanatory theories at each level, and then bridging theories that link the aggregated physiological behavior to the units of explanation at the symbolic level just above.

Above the symbolic level, we can study more comprehensive social phenomena on a different time scale, without serious interaction between our theories of social history, say, and our theories of problem solving. Only aggregative properties of the symbolic processes will enter into the explanation of the larger scale social phenomena (Simon and Ando, 1961).

We can divide up the task of explanation in other ways. Difference equations explain actions and their consequences as functions of the initial conditions; they explain the moment after in terms of the moment before. For many purposes, we can take the system's initial conditions, the contents and organization of memory when our observations begin, as given, and leave to another day and another theory the explanation of how those initial conditions came about.

Thus, we can study the behavior of an accomplished expert and compare it with the behavior of a novice, while putting aside the explanation of how the expert became so. We can study how different strategies—plans versus situated action, say—lead to different behaviors, but study separately how strategies are acquired.

Similarly, we can factor, if only incompletely, the syntax of language from its semantics, and thereby study how speech strings are processed more or less independently of our study of how large structures of knowledge are organized when they are stored in the human brain.

Unified Theories

In pointing to the virtues and even necessities of the divide-and-conquer strategy, I am not denigrating the efforts of others to build unified theories of cognition: Anderson's (1983) ACT*, Newell's (1990) SOAR, or Rumelhart and McClelland's (1986) connectionist systems—just to mention the efforts of some colleagues. But we must understand the goal of those efforts. The goal is not to erect a single system representing the "whole man." Rather, it is to show how a single control structure can handle all of the cognitive processes of which the human mind is capable.

Perhaps the activity would be better understood if it were labeled "unified theories *of the control* of cognition." In any event, the effort to build such comprehensive control structures does not in any way make otiose or superfluous efforts to build explanatory theories of components of cognitive performances, and to build them at various levels of aggregation.

For a realistic conception of what *unified* might mean, we need to look over our shoulders at that most unified and parsimonious of sciences, physics, with its hundreds of pages of theory of specific phenomena at various levels of detail and resolution, all bound together rather shakily into the broader structures of quantum mechanics, relativity theory, and the still somewhat visionary unified field theories.

And if a look at physics does not persuade us that unified theories tell only a small part of the story, we can inspect chemistry, and biology, and geology, and genetics, where the point is even more glaringly obvious.

Methods for the Study of Behavior

Our methods for gathering data to test our theories must fit the formal shapes of the theories. I limit my remarks to theories of symbolic cognitive processes. What are appropriate methods for testing the fit of computer programs (difference equations) to human behavior? The programs predict the next action a system will take as a function of its present state and current input; that is to say, they predict what production will fire at each successive moment. The fineness of resolution of symbolic programs is of the order of tens or

hundreds of milliseconds: The programs predict what the subject will do each few hundreds of milliseconds.

Contemporary technology largely limits us to observing subjects' visible and audible behaviors, and the richest streams of such behaviors are verbalizations and eye movements. Under most circumstances, we do not yet know how to interpret in detail the information we get from electrical measurements on the scalp.

We can obtain data for analyzing the behavior of the Tamil women because one of them, not wanting to miss any of the details of the recipes, is tape-recording their conversation. Unfortunately, the available technology does not permit us to record eye movements on a street in Singapore.

Data on eye movements and verbalizations are still too coarse to capture all the behavior at the symbolic level. In eye movements, we may detect a new saccade every 1/3 or 1/2 s. In verbalization, subjects may utter a clause or phrase equivalent to a proposition every 2 or 3 s, at best. Much of our inference from behavior to the underlying program has to be indirect.

But that is no cause for dismay. In this regard, cognitive psychology is not different from the other sciences, which are always inferring underlying theoretical processes from gross observed events. At that future time when we shall obtain direct evidence, say, electrochemical evidence, identifying precisely the sequence of processes being executed, the game will be over and we will need to look for new domains of research. But we need not hold our breaths while waiting for that to happen.

We now know the difference between verbal protocols, interpreted as behavior, and introspection (Ericsson and Simon, 1984). Over the past quarter century, we have gathered vast experience in encoding verbal protocols and eye movement records at a level of detail that permits us to test what productions are being executed. We should strive to improve these methodologies, and they will continue to improve, but we do not need to be unhappy with our current ability to test our theories of cognition.

Along one dimension at least, considerable unhappiness is still expressed. How can we test the significance of the discrepancies we find between our models and the observed human behavior? Computer programs are complex, having many degrees of freedom. By

taking advantage of this freedom, cannot we simply adjust the program ad hoc to fit any data?

A sound caution underlies this objection. Our confidence in a theory grows, and should grow, with increase in the ratio of the number of data points explained to the number of degrees of freedom in the theory. A theory expressed as a computer program has many degrees of freedom. But a human thinking-aloud protocol, or a set of such protocols, contains a great many data points. It is the ratio that counts, and that ratio can be very large.

Standard procedures for evaluating the fit of computer programs to data are lacking today. The familiar tests of statistical significance are inappropriate. The percentage of variance explained is more useful, but does not take into account the number of degrees of freedom. I have no precise solution to offer to the problem, but the direction in which we should look for one is obvious.

Search for alternative ways of testing our theories brings us back to more conventional psychological experiments. Conventionally, we observe a few behaviors (latencies, accuracies) over some minutes, then average the data over tasks and subjects, then compare the averaged numbers between control and experimental conditions. While this standard procedure is often useful and valuable, it also suffers from severe limitations. Its temporal resolution is very low; it can seldom be used to study individual events of a few seconds' duration.

More serious, conventional experimental methods do not deal with the serial dependency of events on this temporal scale. Since the execution of each production of the cognitive system can change memory contents, hence, change the conditions that determine what production will fire next, it is hard to test an explanation of the behavior unless this temporal dependency can be captured in the data. In particular, averaging over subjects is bound to destroy sequential contingencies. Verbal protocols and eye movement records are almost the only forms of data that give us any means for capturing these contingencies.

A principal means for testing theories of cognition at the level of elementary symbolic processes is to compare the successive behaviors the theories predict with the successive behaviors of subjects revealed by thinking-aloud protocols and eye movement records. The procedures for testing goodness of fit are not yet standardized, but the

underlying principle is to demand a high ratio of data points to numbers of productions in the simulation programs.

Conclusion

We have left our Tamil women standing on the street in Singapore, but I am sure that they will finish their conversation and return home before the heavy afternoon shower drenches them and refreshes the city. They have given us some hope that their behavior, as an example of the general run of human behavior, is explainable, and that today we already possess many important pieces of that explanation at the level of symbolic processes.

By way of summary, I recall here the main generalizations we reached along the way:

Computer Programs as Theories

For systems that change through time, explanation takes the form of laws acting on the current state of the system to produce a new state—endlessly. Such explanations can be formalized with differential or difference equations.

A properly programmed computer can be used to explain the behavior of the dynamic system that it simulates. Theories can be stated as computer programs.

Controlled experiments can be performed on computer programs to determine how such changes affect the performance of tasks. The programs can then be modified to simulate the human performance better.

Programs that simulate cognitive processes describe these processes in symbolic languages and actually execute the processes. Consequently, they test the sufficiency of the theory to perform the tasks.

Symbolic and Physiological Explanation

Explanation of cognitive processes at the information processing (symbolic) level is largely independent of explanation at the physiological (neurological) level.

Explanation on different levels does not deny the possibility of reduction. Higher level theories use aggregates of the constructs at

lower levels. The lower level details do not show through to the higher level.

Some of the properties of systems can be captured in static laws that specify the relations among variables, qualitatively or numerically.

Dependence of Behavior on Knowledge

Different people, or the same people in different situations, can employ different strategies for performing a given task. A theory of their performance would describe their strategies and specify the circumstance under which each strategy will be used.

The core of an expert or expert system is a system of productions that operates like an indexed encyclopedia. External or imagined cues are recognized by the conditions of productions, triggering the associated actions. The case in which the cues are predominantly external is sometimes called situated action.

The production system of an expert is associated also with reasoning (search) capabilities that support an integrated system of analytic and intuitive responses.

Adaptive Systems

The human mind is an adaptive system that chooses behaviors in the light of its goals, and as appropriate to context. Moreover, it can store new knowledge and skills: It can learn.

The link between goals and environment is mediated by learned strategies and knowledge. Behavior cannot be predicted from optimality criteria without information about the strategies and knowledge agents possess or acquire.

The study of the behavior of an adaptive system is not a logical study of optimization but an empirical study of the side conditions that place limits on the approach to the optimum.

Cognitive and Social Psychology

Since strategies and knowledge are both largely acquired from the social environment, there can be no sharp boundary between cognitive psychology and social psychology. The context in which knowledge is acquired and used, an exogenous variable in cognitive psychology, provides the endogenous variables for social psychology and sociology.

Verbal Protocols as Data

Theories of cognition can be tested by comparing the behaviors they predict with the successive behaviors of subjects revealed by thinking-aloud protocols and eye movement records. Strictness demands a high ratio of data points to numbers of productions in the programs.

In summarizing at this high level of abstraction, I have left out all of the rich detail of the behavior we can explain: chess playing, medical diagnosis, problem solving in physics and mathematics, the use of diagrams in thinking, scientific discovery—yes, and even the Tower of Hanoi, and a conversation about cookery on a street in Singapore.

Acknowledgments

This article is in substance the author's Keynote Address to the Third Annual Convention of the American Psychological Society, Washington, DC, June 1991.

The research was supported by the Personnel and Training Programs, Psychological Sciences Division, Office of Naval Research, under Contract No. N00014-86-K-0768, and by the Defense Advanced Research Projects Agency, Department of Defense, ARPA Order 3597, monitored by the Air Force Avionics Laboratory under Contract F33615-81-K-1539. Reproduction in whole or in part is permitted for any purpose of the U.S. government. Approved for public release; distribution unlimited.

Notes

1. For example, see the uses of Ebbinghaus (1964) and Hovland (1951) in testing the EPAM theory of verbal learning (Feigenbaum and Simon, 1984).

2. Because the words in the sentences have denotations in the diagrams, ZBIE has a genuine understanding of the sentences it reads and those it constructs. It anticipates fully, and by a decade, the objections against machine understanding raised by Searle (1984) in his Chinese Room parable—and answers these objections decisively.

3. It is widely believed on the Carnegie-Mellon campus that I cannot give a talk without mentioning the Tower of Hanoi within the first 15 minutes. I contribute this new evidence to support that belief.

References

Anderson, J. R. (1983). *The architecture of cognition.* Cambridge, MA: Harvard University Press.

Anderson, J. R. (1990a). *The adaptive character of thought.* Hillsdale, NJ: Erlbaum.

Anderson, J. R. (1990b). *Cognitive psychology and its implications* (3rd ed.). New York: Freeman.

Anderson, J. R. (1991). The place of cognitive architectures in a rational analysis. In K. VanLehn (Ed.), *Architectures for intelligence* (pp. 1–24). Hillsdale, NJ: Erlbaum.

Baddeley, A. D. (1981). The concept of working memory. *Cognition, 10,* 17–23.

Brownston, L., Farrell, R., Kant, E., and Martin, N. (1985). *Programming expert systems in OPS5.* Reading, MA: Addison-Wesley.

Chomsky, N. (1957). Syntactic structures. The Hague: Mouton.

Duncker, K. (1945). On problem solving. *Psychological Monographs, 58* (Whole No. 270).

Ebbinghaus, H. (1964). *Memory: A contribution to experimental psychology.* New York: Dover.

Ericsson, K. A., and Simon, H. A. (1984). *Protocol analysis.* Cambridge, MA: MIT Press.

Feigenbaum, E. A., and Simon, H. A. (1984). EPAM-like models of recognition and learning. *Cognitive Science, 8,* 305–336.

Gibson, J. J. (1979). *The ecological approach to visual perception.* Boston: Houghton-Mifflin.

Hebb, D. O. (1949). *The organization of behavior.* New York: Wiley.

Hovland, C. I. (1951). Human learning and retention. In S. S. Stevens (Ed.), *Handbook of experimental psychology* (pp. 613–689). New York: Wiley.

Langley, P., Simon, H. A., Bradshaw, G., and Zytkow, J. (1987). *Scientific discovery.* Cambridge, MA: MIT Press.

Miller, R. A., Pople, H. E., and Myers, M. D. (1982). INTERNIST-1, an experimental computer-based diagnostic consultant for general medicine. *New England Journal of Medicine, 307,* 468–476.

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

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

Novak, G. (1977). Representation of knowledge in a program for solving physics problems. *Proceedings of the Fifth International Joint Conference on Artificial Intelligence,* pp. 286–291.

Posner, M. I. (Ed.). (1989). *Foundations of cognitive science.* Cambridge, MA: MIT Press.

Rumelhart, D. E., and McClelland, J. L. (1986). *Parallel distributed processing.* Cambridge, MA: Bradford Books.

Searle, J. R. (1984). *Minds, brains, and science.* Cambridge, MA: Harvard University Press.

Selz, O. (1913). *Ueber die Gezetze des Geordneten Denkverlaufs.* Stuttgart: Spemann.

Shortliffe, E. H. (1976). *Computer-based medical consultations: MYCIN.* New York: American Elsevier.

Siklóssy, L. (1972). Natural language learning by computer. In H. A. Simon and L. Siklóssy (Eds.), *Representation and meaning* (pp. 288–328). Englewood Cliffs, NJ: Prentice-Hall.

Simon, H. A. (1975). The functional equivalence of problem solving skills. *Cognitive Psychology,* 7, 268–288.

Simon, H. A. (1981). *The sciences of the artificial* (2nd ed.). Cambridge, MA: MIT Press.

Simon, H. A. (1991). Cognitive architectures and rational analysis: Comment. In K. VanLehn (Ed.), *Architectures for intelligence* (pp. 25–39). Hillsdale, NJ: Erlbaum.

Simon, H. A., and Ando, A. (1961). Aggregation of variables in dynamic systems. *Econometrica, 29,* 111–138.

Skinner, B. F. (1938). *The behavior of organisms.* New York: Appleton-Century-Crofts.

Suchman, L. A. (1987). *Plans and situated actions.* Cambridge, England: Cambridge University Press.

Voss, J. F., Tyler, S. W., and Vengo, L. A. (1983). Individual differences in the solving of social science problems. In R. Dillon and R. Smech (Eds.), *Individual differences in cognition* (pp. 205–232). New York: Academic Press.

Winograd, T., and Flores, F. (1986). *Understanding computers and cognition.* Norwood, NJ: Ablex.

Wundt, W. (1902). *Grundriss der psychologie.* Leipzig: Wilhelm Engelmann.

Zhang, G.-J., and Simon, H. A. (1985). STM capacity for Chinese words and idioms. *Memory and Cognition, 13,* 193–201.