Reprinted from AMERICAN PSYCHOLOGIST, Vol. 26, No. 2, February 1971
Printed in U. S. A.

HUMAN PROBLEM SOLVING:
THE STATE OF THE THEORY IN 1970[1]

HERBERT A. SIMON AND ALLEN NEWELL[2]
Carnegie-Mellon University

WHEN the magician pulls the rabbit from the hat, the spectator can respond either with mystification or with curiosity. He can enjoy the surprise and the wonder of the unexplained (and perhaps inexplicable), or he can search for an explanation.

Suppose curiosity is his main response—that he adopts a scientist's attitude toward the mystery. What questions should a scientific theory of magic answer? First, it should predict the performance of a magician handling specified tasks—producing a rabbit from a hat, say. It should explain how the production takes place, what processes are used, and what mechanisms perform those processes. It should predict the incidental phenomena that accompany the magic—the magician's patter and his pretty assistant—and the relation of these to the mystification process. It should show how changes in the attendant conditions—both changes "inside" the members of the audience and changes in the feat of magic—alter the magician's behavior. It should explain how specific and general magician's skills are learned, and what the magician "has" when he has learned them.

[1] The research reported here was supported in part by United States Public Health Service Research Grant MH-07722, from the National Institute of Mental Health.

[2] Since the Distinguished Scientific Contribution Award citation last year recognized that the work for which it was awarded was done by a team, rather than an individual, Dr. Simon thinks it appropriate that Allen Newell, with whom he has been in full partnership from the very beginning of the effort, should be enlisted into coauthorship of this report on it. Both authors would like to acknowledge their debts to the many others who have been members of the team during the past decade and a half, but especially to J. C. Shaw and Lee W. Gregg. This article is based on the final chapter of the authors' forthcoming book, *Human Problem Solving* (Englewood Cliffs, N. J.: Prentice-Hall, in press).

Requests for reprints should be sent to the authors, Graduate School of Industrial Administration, Carnegie-Mellon University, Pittsburgh, Pennsylvania 15213.

THEORY OF PROBLEM SOLVING—1958

Now I have been quoting—with a few word substitutions—from a paper published in the *Psychological Review* in 1958 (Newell, Shaw, & Simon, 1958). In that paper, titled "Elements of a Theory of Human Problem Solving," our research group reported on the results of its first two years of activity in programming a digital computer to perform problem-solving tasks that are difficult for humans. Problem solving was regarded by many, at that time, as a mystical, almost magical, human activity—as though the preservation of human dignity depended on man's remaining inscrutable to himself, on the magic-making processes remaining unexplained.

In the course of writing the "Elements" paper, we searched the literature of problem solving for a statement of what it would mean to explain human problem solving, of how we would recognize an explanation if we found one. Failing to discover a statement that satisfied us, we manufactured one of our own—essentially the paragraph I paraphrased earlier. Let me quote it again, with the proper words restored, so that it will refer to the magic of human thinking and problem solving, instead of stage magic.

> What questions should a theory of problem solving answer? First, it should predict the performance of a problem solver handling specified tasks. It should explain how human problem solving takes place: what processes are used, and what mechanisms perform these processes. It should predict the incidental phenomena that accompany problem solving, and the relation of these to the problem-solving process. . . . It should show how changes in the attendant conditions—both changes "inside" the problem solver and changes in the task confronting him—alter problem-solving behavior. It should explain how specific and general problem-solving skills are learned, and what it is that the problem solver "has" when he has learned them [p. 151].

A Strategy

This view of explanation places its central emphasis on process—on *how* particular human behaviors come about, on the mechanisms that enable them. We can sketch out the strategy of a research program for achieving such an explanation, a strategy that the actual events have been following pretty closely, at least through the first eight steps:

1. Discover and define a set of processes that would enable a system capable of storing and manipulating patterns to perform complex nonnumerical tasks, like those a human performs when he is thinking.

2. Construct an information-processing language, and a system for interpreting that language in terms of elementary operations, that will enable programs to be written in terms of the information processes that have been defined, and will permit those programs to be run on a computer.

3. Discover and define a program, written in the language of information processes, that is capable of solving some class of problems that humans find difficult. Use whatever evidence is available to incorporate in the program processes that resemble those used by humans. (Do not admit processes, like very rapid arithmetic, that humans are known to be incapable of.)

4. If the first three steps are successful, obtain data, as detailed as possible, on human behavior in solving the same problems as those tackled by the program. Search for the similarities and differences between the behavior of program and human subject. Modify the program to achieve a better approximation to the human behavior.

5. Investigate a continually broadening range of human problem-solving and thinking tasks, repeating the first four steps for each of them. Use the same set of elementary information processes in all of the simulation programs, and try to borrow from the subroutines and program organization of previous programs in designing each new one.

6. After human behavior in several tasks has been approximated to a reasonable degree, construct more general simulation programs that can attack a whole range of tasks—winnow out the "general intelligence" components of the performances, and use them to build this more general program.

7. Examine the components of the simulation programs for their relation to the more elementary human performances that are commonly studied in the psychological laboratory: rote learning, elementary concept attainment, immediate recall, and so on. Draw inferences from simulations to elementary performances, and vice versa, so as to use standard experimental data to test and improve the problem-solving theories.

8. Search for new tasks (e.g., perceptual and language tasks) that might provide additional arenas for testing the theories and drawing out their implications.

9. Begin to search for the neurophysiological counterparts of the elementary information processes that are postulated in the theories. Use neurophysiological evidence to improve the problem-solving theories, and inferences from the problem-solving theories as clues for the neurophysiological investigations.

10. Draw implications from the theories for the improvement of human performance—for example, the improvement of learning and decision making. Develop and test programs of application.

11. Review progress to date, and lay out a strategy for the next period ahead.

Of course, life's programs are not as linear as this strategy, in the simplified form in which we have presented it. A good strategy would have to contain many checkpoints for evaluation of progress, many feedback loops, many branches, many iterations. Step 1 of the strategy, for example, was a major concern of our research group (and other investigators as well) in 1955–56, but new ideas, refinements, and improvements have continued to appear up to the present time. Step 7 represented a minor part of our activity as early as 1956, became much more important in 1958–61, and has remained active since.

Nor do strategies spring full-grown from the brow of Zeus. Fifteen years' hindsight makes it easy to write down the strategy in neat form. If anyone had attempted to describe it prospectively in 1955, his version would have been much cruder and probably would lack some of the last six steps.

The Logic Theorist

The "Elements" paper of 1958 reported a successful initial pass through the first three steps in the strategy. A set of basic information processes for manipulating nonnumerical symbols and symbol structures had been devised (Newell & Simon, 1956). A class of information-processing or list-processing languages had been designed and implemented, incorporating the basic information processes, permitting programs to be written in terms of them, and enabling these programs to be run on computers (Newell & Shaw, 1957). A program, The Logic Theorist (LT), had been written in one of these languages, and had been shown, by running it on a computer, to be capable of solving problems that are difficult for humans (Newell, Shaw, & Simon, 1957).

LT was, first and foremost, a demonstration of sufficiency. The program's ability to discover proofs for theorems in logic showed that, with no more capabilities than it possessed—capabilities for reading, writing, storing, erasing, and comparing

patterns—a system could perform tasks that, in humans, require thinking. To anyone with a taste for parsimony, it suggested (but, of course, did not prove) that only these capabilities, and no others, should be postulated to account for the magic of human thinking. Thus, the "Elements" paper proposed that *"an explanation of an observed behavior of the organism is provided by a program of primitive information processes that generates this behavior* [p. 151]," and exhibited LT as an example of such an explanation.

The sufficiency proof, the demonstration of problem-solving capability at the human level, is only a first step toward constructing an information-processing theory of human thinking. It only tells us that in certain stimulus situations the correct (that is to say, the human) gross behavior can be produced. But this kind of blind S-R relation between program and behavior does not explain the process that brings it about. We do not say that we understand the magic because we can predict that a rabbit will emerge from the hat when the magician reaches into it. We want to know how it was done—how the rabbit got there. Programs like LT are explanations of human problem-solving behavior only to the extent that the processes they use to discover solutions are the same as the human processes.

LT's claim to explain process as well as result rested on slender evidence, which was summed up in the "Elements" paper as follows:

> First, . . . (LT) is in fact capable of finding proofs for theorems—hence incorporates a system of processes that is sufficient for a problem-solving mechanism. Second, its ability to solve a particular problem depends on the sequence in which problems are presented to it in much the same way that a human subject's behavior depends on this sequence. Third, its behavior exhibits both preparatory and directional set. Fourth, it exhibits insight both in the sense of vicarious trial and error leading to "sudden" problem solution, and in the sense of employing heuristics to keep the total amount of trial and error within reasonable bounds. Fifth, it employs simple concepts to classify the expressions with which it deals. Sixth, its program exhibits a complex organized hierarchy of problems and subproblems [p. 162].

There were important differences between LT's processes and those used by human subjects to solve similar problems. Nevertheless, in one fundamental respect that has guided all the simulations that have followed LT, the program did indeed capture the central process in human problem solving: LT used heuristic methods to carry out highly selective searches, hence to cut down enormous problem spaces to sizes that a slow, serial processor could handle. Selectivity of search, not speed, was taken as the key organizing principle, and essentially no use was made of the computer's ultrarapid arithmetic capabilities in the simulation program. Heuristic methods that make this selectivity possible have turned out to be the central magic in all human problem solving that has been studied to date.

Thus, in the domain of symbolic logic in which LT worked, obtaining by brute force the proofs it discovered by selective search would have meant examining enormous numbers of possibilities—10 raised to an exponent of hundreds or thousands. LT typically searched trees of 50 or so branches in constructing the more difficult proofs that it found.

Mentalism and Magic

LT demonstrated that selective search employing heuristics permitted a slow serial information-processing system to solve problems that are difficult for humans. The demonstration defined the terms of the next stages of inquiry: to discover the heuristic processes actually used by humans to solve such problems, and to verify the discovery empirically.

We will not discuss here the methodological issues raised by the discovery and certification tasks, apart from one preliminary comment. An explanation of the processes involved in human thinking requires reference to things going on inside the head. American behaviorism has been properly skeptical of "mentalism"—of attempts to explain thinking by vague references to vague entities and processes hidden beyond reach of observation within the skull. Magic is explained only if the terms of explanation are less mysterious than the feats of magic themselves. It is no explanation of the rabbit's appearing from the hat to say that it "materialized."

Information-processing explanations refer frequently to processes that go on inside the head—in the mind, if you like—and to specific properties of human memory: its speed and capacity, its organization. These references are not intended to be in the least vague. What distinguishes the information-processing theories of thinking and problem solving described here from earlier discussion of mind is that terms like "memory" and "symbol structure" are now pinned down and defined in

sufficient detail to embody their referents in precisely stated programs and data structures.

An internal representation, or "mental image," of a chess board, for example, is not a metaphorical picture of the external object, but a symbol structure with definite properties on which well-defined processes can operate to retrieve specified kinds of information (Baylor & Simon, 1966; Simon & Barenfeld, 1969).

The programmability of the theories is the guarantor of their operationality, an iron-clad insurance against admitting magical entities into the head. A computer program containing magical instructions does not run, but it is asserted of these information-processing theories of thinking that they can be programmed and will run. They may be empirically correct theories about the nature of human thought processes or empirically invalid theories; they are not magical theories.

Unfortunately, the guarantee provided by programmability creates a communication problem. Information-processing languages are a barrier to the communication of the theories as formidable as the barrier of mathematics in the physical sciences. The theories become fully accessible only to those who, by mastering the languages, climb over the barrier. Any attempt to communicate in natural language must perforce be inexact.

There is the further danger that, in talking about these theories in ordinary language, the listener may be seduced into attaching to terms their traditional meanings. If the theory speaks of "search," he may posit a little homunculus inside the head to do the searching; if it speaks of "heuristics" or "rules of thumb," he may introduce the same homunculus to remember and apply them. Then, of course, he will be interpreting the theory magically, and will object that it is no theory.

The only solution to this problem is the hard solution. Psychology is now taking the road taken earlier by other sciences: it is introducing essential formalisms to describe and explain its phenomena. Natural language formulations of the phenomena of human thinking did not yield explanations of what was going on; formulations in information-processing languages appear to be yielding such explanations. And the pain and cost of acquiring the new tools must be far less than the pain and cost of trying to master difficult problems with inadequate tools.

Our account today will be framed in ordinary language. But we must warn you that it is a translation from information-processing languages which, like most translations, has probably lost a good deal of the subtlety of the original. In particular, we warn you against attaching magical meanings to terms that refer to entirely concrete and operational phenomena taking place in fully defined and operative information-processing systems. The account will also be Pittsburgh-centric. It will refer mainly to work of the Carnegie-RAND group, although information-processing psychology enlists an ever-growing band of research psychologists, many of whom are important contributors of evidence to the theory presented here.

THEORY OF PROBLEM SOLVING—1970

The dozen years since the publication of the "Elements" paper has seen a steady growth of activity in information-processing psychology—both in the area of problem solving and in such areas as learning, concept formation, short-term memory phenomena, perception, and language behavior. Firm contact has been made with more traditional approaches, and information-processing psychology has joined (or been joined by) the mainstream of scientific inquiry in experimental psychology today.[3] Instead of tracing history here, we should like to give a brief account of the product of the history, of the theory of human problem solving that has emerged from the research.

The theory makes reference to an *information-processing system,* the problem solver, confronted by a task. The task is defined objectively (or from the viewpoint of an experimenter, if you prefer) in terms of a *task environment.* It is defined by the problem solver, for purposes of attacking it, in terms of a *problem space.* The shape of the theory can be captured by four propositions (Newell & Simon, in press, Ch. 14):

1. A few, and only a few, gross characteristics of the human information-processing system are invariant over task and problem solver.

2. These characteristics are sufficient to determine that a task environment is represented (in the information-processing system) as a problem space,

[3] The authors have undertaken a brief history of these developments in an Addendum to their book, *Human Problem Solving* (Newell & Simon, in press).

phenomena once explained. Those who have the instincts and esthetic tastes of scientists presumably will not be disappointed. There is much beauty in the superficial complexity of nature. But there is a deeper beauty in the simplicity of underlying process that accounts for the external complexity. There is beauty in the intricacy of human thinking when an intelligent person is confronted with a difficult problem. But there is a deeper beauty in the basic information processes and their organization into simple schemes of heuristic search that make that intricate human thinking possible. It is a sense of this latter beauty—the beauty of simplicity—that we have tried to convey to you.

REFERENCES

BARTLETT, F. C. *Thinking*. New York: Basic Books, 1958.

BAYLOR, G. W., JR., & SIMON, H. A. A chess mating combinations program. *AFIPS Conference Proceedings*, 1966 Spring Joint Computer Conference, Boston, April 26–28, 28, 431–447. Washington, D. C.: Spartan Books, 1966.

DE GROOT, A. *Thought and choice in chess*. The Hague: Mouton, 1965.

DUNCKER, K. On problem solving. *Psychological Monographs*, 1945, 58(5, Whole No. 270).

ERNST, G. W., & NEWELL, A. *GPS: A case study in generality and problem solving*. New York: Academic Press, 1969.

HILGARD, E. R., & BOWER, G. W. *Theories of learning*. (3rd ed.) New York: Appleton-Century-Crofts, 1966.

HILGARD, E. R., & ATKINSON, R. C. *Introduction to psychology*. (4th ed.) New York: Harcourt, Brace & World, 1967.

KATONA, G. *Organizing and memorizing*. New York: Columbia University Press, 1940.

KLAHR, D., & WALLACE, J. G. Development of serial completion strategy: Information processing analysis. *British Journal of Psychology*, 1970, 61, 243–257.

NEWELL, A. Studies in problem solving: Subject 3 on the cryptarithmetic task: DONALD + GERALD = ROBERT. Pittsburgh: Carnegie-Mellon University, 1967.

NEWELL, A., & SHAW, J. C. Programming the Logic Theory Machine. *Proceedings of the 1957 Western Joint Computer Conference*, February 26–28, 1957, 230–240.

NEWELL, A., SHAW, J. C., & SIMON, H. A. Empirical explorations of the Logic Theory Machine: A case study in heuristics. *Proceedings of the 1957 Western Joint Computer Conference*, February 26–28, 1957, 218–230.

NEWELL, A., SHAW, J. C., & SIMON, H. A. Elements of a theory of human problem solving. *Psychological Review*, 1958, 65, 151–166.

NEWELL, A., SHAW, J. C., & SIMON, H. A. The processes of creative thinking. In H. E. Gruber & M. Wertheimer (Eds.), *Contemporary approaches to creative thinking*. New York: Atherton Press, 1962.

NEWELL, A., & SIMON, H. A. The Logic Theory Machine: A complex information processing system. *IRE Transactions on Information Theory*, 1956, IT-2, 3, 61–79.

NEWELL, A., & SIMON, H. A. *Human problem solving*. Englewood Cliffs, N. J.: Prentice-Hall, 1971, in press.

PAIGE, J. M., & SIMON, H. A. Cognitive processes in solving algebra word problems. In B. Kleinmuntz (Ed.), *Problem solving*. New York: Wiley, 1966.

SIMON, H. A. Scientific discovery and the psychology of problem solving. In R. C. Colodny (Ed.), *Mind and cosmos: Essays in contemporary science and philosophy*. Pittsburgh: University of Pittsburgh Press, 1966.

SIMON, H. A. An information-processing explanation of some perceptual phenomena. *British Journal of Psychology*, 1967, 58, 1–12.

SIMON, H. A. *The sciences of the artificial*. Cambridge: M.I.T. Press, 1969.

SIMON, H. A., & BARENFELD, M. Information-processing analysis of perceptual processes in problem solving. *Psychological Review*, 1969, 76, 473–483.

SIMON, H. A., & KOTOVSKY, K. Human acquisition of concepts for sequential patterns. *Psychological Review*, 1963, 70, 534–546.

WILLIAMS, D. S. Computer program organization induced by problem examples. Unpublished doctoral dissertation, Carnegie-Mellon University, 1969.

WILLIAMS, T. G. Some studies in game playing with a digital computer. Unpublished doctoral dissertation, Carnegie Institute of Technology, 1965.