Herbert A. Simon

THE SCIENCES OF THE ARTIFICIAL

reissue of the third edition with a new introduction by John E. Laird

The Sciences of the Artificial

The Sciences of the Artificial

Reissue of the third edition
with a new introduction by John E. Laird

Herbert A. Simon

The MIT Press
Cambridge, Massachusetts
London, England

This book was set in Sabon by Graphic Composition, Inc. Printed and bound in the United States of America.

Library of Congress Cataloging-in-Publication Data
Names: Simon, Herbert A. (Herbert Alexander), 1916-2001.
Title: The sciences of the artificial / Herbert A. Simon ; introduction by John E. Laird.
Description: Third edition [2019 edition]. | Cambridge, MA : The MIT Press, [2019] | "Reissue of the third edition with a new introduction by John Laird." | Third edition originally published in 1996. | Includes bibliographical references and indexes.
Identifiers: LCCN 2018058336 | ISBN 9780262537537 (pbk. : alk. paper)
Subjects: LCSH: Science--Philosophy.
Classification: LCC Q175 .S564 2019 | DDC 006.301--dc23
LC record available at https://lccn.loc.gov/2018058336

10 9 8 7 6 5 4 3 2 1

To Allen Newell
in memory of a friendship

Contents

Introduction

I've been away too long . . .

I first read *The Sciences of the Artificial* in the late 1970s as a graduate student studying artificial intelligence (AI) at Carnegie Mellon University with Allen Newell, Herb Simon's former student, close friend, and longtime colleague. At the time, I had a narrow view of science and was struggling to understand the field that would ultimately become my career. What was *artificial intelligence?* How did it compare to natural intelligence in humans? Was it really a science? Was it and other *artificial* sciences intrinsically different from natural sciences?

The Sciences of the Artificial provided many answers. I was intrigued by the idea that there could be a science, not just of natural phenomena, but also of what was artificial. Even though Simon's preferred name for AI was *complex information processing* or *simulation of cognitive processes*, he took seriously the idea that *artificial* had real semantic content, and he explored its meaning. For me, this exploration legitimized the term, and provided a solid foundation for AI as a scientific discipline. He began with a dictionary definition: "Produced by art rather than by nature; not genuine or natural; affected; not pertaining to the essence of the matter," choosing the first definition. He refined it and extended it to be systems designed by humans "in order to *attain goals*, and to *function,*" as well as to entities that adapt to goals and their environment.[1]

In the book, he brings into play his considerable and varied personal and academic expertise to explore the fundamental commonalities that transcend artificial systems including economic systems, the business firm, artificial intelligence, sophisticated engineering designs, and social plans.

1. Herbert A. Simon, *The Sciences of the Artificial*, 3rd ed. (Cambridge, MA: The MIT Press, 1996), 4.

His hypothesis is that designed systems are a valid subject of enquiry, and he proposes a science of design. He expands his analysis to complex systems in general, including biology, evolution, and the human mind.

This is an audacious agenda. But for Herbert Simon, it was a synthesis of his life's work, and the book tracks his evolution as a researcher and scholar. He is unique in not only the path he took in research but also in the ideas he developed through the years and in his ability to apply them to such disparate fields. The first edition, published in 1969, was a mere 118 pages, distilling his work in political science, economics, organizational theory, psychology, cognitive science, and computer science. In 1996, with the third edition, the book grew to 229 pages, as he extended and revised it to incorporate advances in his own research and in the many fields covered by the book.

Simon was originally trained in mathematics, logic, and economics but focused on political science in graduate school. One distinctive aspect of his approach to all these fields was that he set his "sights on the phenomena of human thinking and problem solving as the essential core of both organization theory and economics. . . . Organizations, it appeared, could be understood by applying to them what you knew about human behavior generally."[2]

His Ph.D. dissertation at the University of Chicago was a direct reflection of these ideas. In 1947, his thesis was published as *Administrative Behavior*: "It was built around two interrelated ideas that have been at the core of my whole intellectual activity: (1) human beings are able to achieve only a very bounded rationality, and (2) as one consequence of their cognitive limitations, they are prone to identify with subgoals."[3] Simon observed that humans are rarely, if ever, omniscient and completely rational, and have difficulty making optimal decisions that bring the greatest benefit possible.[4] Instead, they are limited in their available knowledge and instead *satisfice*, making decisions that achieve their most important goals, but may not be the absolute best solution. His development and exploration of the idea of bounded rationality helped establish the now flourishing field of behavioral economics, and it was the

2. Herbert A. Simon, *Models of My Life* (Cambridge, MA: The MIT Press, 1996), 56, 74.
3. Simon, *Models of My Life*, 88.
4. Herbert A. Simon, *Administrative Behavior: A Study of Decision-Making Processes in Administrative Organization* (New York: Macmillan, 1947).

cornerstone of his research that led to winning the 1978 Nobel Prize in Economics for ". . . his pioneering research into the decision-making process within economic organizations."

In 1949, Simon moved to Pittsburgh and helped establish the Graduate School of Industrial Administration, at what at the time was the Carnegie Institute of Technology. He continued to pursue his research on decision making in organizations through a combination of empirical and theoretical work. All of that changed in 1955 and 1956 through his work with Allen Newell and Cliff Shaw: "(I) transformed professionally into a cognitive psychologist and computer scientist, almost abandoning my earlier professional identity."[5] Simon was very much a *prepared mind* for understanding the potential that computers held, as he had already conceived of human decision making as a decomposable process open to scientific enquiry. He recognized that computers could not only process numbers but also perform symbolic computation and thus perform the same types of reasoning underlying human thought.

Simon, Newell, and Shaw's initial ideas came to life with two seminal developments, both related to symbolic computation. The first was the Logic Theorist, which was "the first computer program that solved non-numerical problems with selective search. It is for these two achievements that we are commonly adjudged to be the parents of artificial intelligence."[6] The Logic Theorist discovered proofs for theorems in *Principia Mathematica,* and was first presented at the now famous 1956 Dartmouth Summer Research Project on Artificial Intelligence. It was one of the few, if not the only, running computer programs presented at that meeting. The second development was the design and implementation of the Information Processing Languages (IPL-I through IPL-VI), which were used for developing LT and later AI systems. IPL-II was the first implemented list-processing language that directly supported symbolic computation and was a direct precursor to John McCarthy's LISP.

His collaboration with Newell and Shaw continued through the 1950s and 1960s. Together they developed the General Problem Solver (GPS).[7] GPS embodied their theory of problem solving based on means-ends

5. Simon, *Models of My Life*, 189.
6. Simon, *Models of My Life*, 189.
7. A. Newell, J. C. Shaw, and H. A. Simon. "Report on a General Problem-Solving Program," in *Proceedings of the International Conference on Information Processing* (Santa Monica, CA: RAND Corporation, 1959), 256–264.

analysis, and was the first AI program that separated task-specific knowledge, which consisted of the operators (the *means*), the goals (the *ends*), and the connections between them, from the problem-solving strategy (means-ends analysis). Thus, GPS provided a fixed, generic inference engine that could be used for many different problems. It set the stage for the development of domain-independent cognitive architectures, inspiring and directly contributing to my own work with the Soar architecture.[8] Simon's work with Newell culminated with the publication of *Human Problem Solving*,[9] a tome that set forth data, theory, and computer models of human goal-oriented reasoning across a wide range of tasks.

What attracted me to Simon and Newell's work was their strategy of developing artificial intelligence systems that closely modeled what we knew of human intelligence, and conversely expanded our knowledge of what was computationally possible for human reasoning. This methodology was their *secret weapon*, which they openly espoused, but it still provided them with unique and novel insights throughout their careers. Human behavior data was always an important component of their theories, possibly best illustrated by the voluminous protocols in *Human Problem Solving*. Their use of knowledge from other fields went beyond psychology, and Simon was unique in the expanse of disciplines he could draw from. But just as important as drawing from many fields was that their theories be *computational*, and implemented in running computer programs: "In the computer field, the moment of truth is a running program; all else is prophecy."[10]

In an interesting twist of fate, in 1975, three years before he won the Nobel Prize in Economics for the significant contributions of his *first* career, Simon together with Newell won the Turing Award—essentially the Nobel Prize of computer science, for the equally significant contributions of his *second* career: [for making] "basic contributions to artificial intelligence, the psychology of human cognition, and list processing." Newell and Simon demystified what it means to *think*, providing both

8. John E. Laird, *The Soar Cognitive Architecture* (Cambridge, MA: The MIT Press, 2012).
9. A. Newell and H. A. Simon, *Human Problem Solving* (Englewood Cliffs, NJ: Prentice-Hall, 1972).
10. Herbert A. Simon, *The Shape of Automation for Men and Management* (New York: Harper and Row, 1965), xv.

theory and empirical evidence as to many of the computational processes underlying thought.

Throughout the rest of his career, he used his secret weapon to pursue research in AI and cognitive psychology that combined both empirical psychological research with computational modeling. He and his colleagues went after big, uncharted areas of cognition—those areas of thinking and learning where there was mystery—where there was rarely an established theory, much less computational models. His research with Ed Feigenbaum led to the first computational model of verbal learning, skill learning, and concept learning (EPAM),[11] which was subsequently extended by Fernand Gobet to model human chess players (CHREST). In a completely different direction, he explored the interactions between motivation and emotion, and cognition.[12] Together with Anders Ericsson,[13] he studied verbal protocols, a key methodology in cognitive science that Simon and Newell pioneered. He also did groundbreaking theoretical work with John Hayes that pushed beyond GPS and its ability to represent and solve multiple tasks, to the UNDERSTAND system,[14] which provided a theory of how new task representations can be learned. He led early work on theories of scientific discovery, working with Pat Langley, Jan Zytkow, and Gary Bradshaw. This work resulted in the development of the BACON[15] system and its many descendants. His research on learning by doing with Yuichiro Anzai[16] was also groundbreaking, as was his work on the role of diagrams in understanding with Jill Larkin.[17] According to Google Scholar, Herbert Simon's body of work is one of the most

11. Edward A. Feigenbaum and Herbert A. Simon, "EPAM-Like Models of Recognition and Learning," *Cognitive Science* 8, no. 4 (1984): 305–336.
12. H. A. Simon, "Motivational and Emotional Controls of Cognition," *Psychological Review* 74, no. 1 (1967): 29–39. http://dx.doi.org/10.1037/h0024127.
13. K. Anders Ericsson and Herbert A. Simon, *Protocol Analysis: Verbal Reports as Data*, rev. ed. (Cambridge, MA: The MIT Press, 1993).
14. John R. Hayes and Herbert A. Simon, "The Understanding Process: Problem Isomorphs," *Cognitive Psychology* 8 (1976): 165–190.
15. Pat Langley, Herbert A. Simon, Gary L. Bradshaw, and Jan M. Zytkow, *Scientific Discovery: Computational Explorations of the Creative Process* (Cambridge, MA: The MIT Press, 1987).
16. Y. Anzai and H. A. Simon, "The Theory of Learning by Doing," *Psychological Review* 86, no. 2 (1979): 124–140.
17. J. H. Larkin and H. A. Simon, "Why a Diagram Is (Sometimes) Worth Ten Thousand Words," *Cognitive Science* 11, no. 1 (1987): 65–100. doi:10.1111/j.1551-6708.1987.tb00863.x.

cited of all time, and as of 2016, he was the most cited person in AI and cognitive psychology.

* * *

In returning to this book after many years, it is a pleasure to see once again how Simon distilled the essential ideas from his lifelong pursuit of understanding organizations and the human mind. It is not just an introduction to those topics, but is his attempt at a comprehensive theory of the commonalities that arise across the seemingly disparate disciplines in which he finds ideas that transcend all of them. Every reader will find their favorite topics, most likely a reflection of their own disciplinary perspective, which is invariably much narrower than Simon's. One of the joys of reading this book is that you do get to see foundational ideas from your own discipline reflected in others, giving deeper insight into them. Simon's strength as a writer is that a reader does not have to be well versed in every field that he covers, nor have an advanced degree in any of those areas. He provides building blocks to understand fields as diverse as AI, design, psychology, and economics. My background is in computer science and artificial intelligence and to a lesser extent cognitive psychology, but I find all of the topics accessible (and fascinating), including organizational administration, economics, biology, evolution, and design. And although the book introduces a breadth of topics, it also provides depth. Within my own disciplines of AI and cognitive science, revisiting Simon's insights and perspective—many years since I originally encountered them—still enriches my understanding of familiar topics and reminds me of some of the beauty of how these pieces fit together.

If you have read the book before, it is worthwhile to come back and read it again. Simon significantly expanded the book in the second and third editions, incorporating new research and ideas, so if you read only the original edition, consider reading the book afresh. My expectation is that your experience will mirror my own, in that today, my mind is much better prepared to understand and appreciate the breadth and depth of his ideas and arguments. Each time I come back, I have changed and grown intellectually, and although there are many things I remember from my first reading, experiencing them again brings new insights not only to other disciplines but also to my own work.

If you have never read it, you are in for a treat, as you will experience a true master leading you through ideas that define the very foundations of artificial sciences and explain their connections with the natural sciences.

I will not attempt to give a comprehensive overview of the book—Simon provides that in the first chapter: "Understanding the Natural and the Artificial Worlds"—but I will explore one of the most memorable aspects of the book, his parable of the ant on the beach, because of its simplicity, and also because of the continued relevance of the underlying ideas it embodies.

In what has become known as "Simon's Ant," Simon observes that as an ant traverses a beach toward a destination, its path is not a straight line but is instead a complex set of climbs and turns as the ant navigates through its terrain. Simon's point is that the complexity of the ant's behavior comes not from any complex internal reasoning of the ant but from the complexity of its environment. Simon extends this to human behavior, arguing that the basic structures underlying human behavior are actually quite simple—we are just like the ant, pursuing our goals, attempting to find satisfactory solutions to our current problems, and that the complexity of our behavior is a reflection of the complexity of our tasks, our embodiment, and our environment: "Human beings, viewed as behaving systems, are quite simple. The apparent complexity of our behavior over time is largely a reflection of the complexity of the environment in which we find ourselves."[18]

My initial reaction was confusion. Aren't humans much more complicated than ants, and doesn't our difficulty in modeling and predicting human behavior come from the internal complexity we do not understand? But over the years, I've learned that any modeling of our behavior starts with understanding the environment and the task, and then the human's knowledge of the environment and task. Once we understand those things, predicting behavior becomes much simpler. As Simon stated, we *satisfice* by using our knowledge to choose actions that are most likely to help us achieve our goals. Our internal complexity is a reflection of the *models* of our environment that we have learned through experience. They are models of the objects, relations, and dynamics of our world, but also of how our actions affect the world and how our actions help us achieve our goals. The success of an organism depends on its ability to acquire these models, and they are where the bulk of complexity lies. Our success as a species is because we can learn better models of the complex relations and interactions in our environment, including models of each other, than other animals can. Beyond a theory of the mind, Simon's view

18. Simon, *Sciences of the Artificial*, 110.

of the world in terms of models was reflected in the title of many of his books and ultimately his autobiography: *Models of Man* (1957), *Models of Discovery* (1977), *Models of Thought* (1979 and 1989), *Models of Bounded Rationality* (1982), and *Models of My Life* (1991).

One of my surprises in revisiting the book was the recognition that the recent successes in game-playing programs are completely consistent with Simon's point of view. AlphaZero can learn to play either Go or chess at world-championship level.[19] Its learning mechanisms extract *models* of a game by playing millions of times against itself. It learns a symbolic model of the game states and their transitions, and a statistical model of the expected value of those states and transitions. Using the learned models, the program does a look-ahead search to evaluate alternative moves, picking the best one given its available computational resources. It does not attempt to find the optimal move, which for these games is almost always impossible, but instead it *satisfices*, picking the best move given its knowledge of its environment. The learning and search components are significant scientific advances, but the real complexity of the final program is in its model of the game. More generally, the recent advances in Machine Learning are tied to the development of conceptually simple learning mechanisms that exploit massive computational resources and data to build better and better models of their environments.

It was a real pleasure to return one more time to *The Sciences of the Artificial*. It is one of my all-time favorite books. Even though the world has changed in ways unimaginable over the last fifty years, this book still has ideas and analyses that seem to be fresh. I was surprised, after all these years, how much I learned from it and how much it inspired me to think about deep problems in AI, cognitive science, economics, and beyond.

Enjoy!

John Laird

19. David Silver, Thomas Hubert, Julian Schrittwieser, Ioannis Antonoglou, Matthew Lai, Arthur Guez, Marc Lanctot, et al., "A General Reinforcement Learning Algorithm That Masters Chess, Shogi, and Go through Self-Play," *Science* (December 7, 2018), 1140–1144.

Preface to Third Edition

As the Earth has made more than 5,000 rotations since *The Sciences of the Artificial* was last revised, in 1981, it is time to ask what changes in our understanding of the world call for changes in the text.

Of particular relevance is the recent vigorous eruption of interest in complexity and complex systems. In the previous editions of this book I commented only briefly on the relation between general ideas about complexity and the particular hierarchic form of complexity with which the book is chiefly concerned. I now introduce a new chapter to remedy this deficit. It will appear that the devotees of complexity (among whom I count myself) are a rather motley crew, not at all unified in our views on reductionism. Various among us favor quite different tools for analyzing complexity and speak nowadays of "chaos," "adaptive systems," and "genetic algorithms." In the new chapter 7, "Alternative Views of Complexity" ("The Architecture of Complexity" having become chapter 8), I sort out these themes and draw out the implications of artificiality and hierarchy for complexity.

Most of the remaining changes in this third edition aim at updating the text. In particular, I have taken account of important advances that have been made since 1981 in cognitive psychology (chapters 3 and 4) and the science of design (chapters 5 and 6). It is gratifying that continuing rapid progress in both of these domains has called for numerous new references that record the advances, while at the same time confirm and extend the book's basic theses about the artificial sciences. Changes in emphases in chapter 2 reflect progress in my thinking about the respective roles of organizations and markets in economic systems.

This edition, like its predecessors, is dedicated to my friend of half a lifetime, Allen Newell—but now, alas, to his memory. His final book, *Unified Theories of Cognition,* provides a powerful agenda for advancing our understanding of intelligent systems.

I am grateful to my assistant, Janet Hilf, both for protecting the time I have needed to carry out this revision and for assisting in innumerable ways in getting the manuscript ready for publication. At the MIT Press, Deborah Cantor-Adams applied a discerning editorial pencil to the manuscript and made communication with the Press a pleasant part of the process. To her, also, I am very grateful.

In addition to those others whose help, counsel, and friendship I acknowledged in the preface to the earlier editions, I want to single out some colleagues whose ideas have been especially relevant to the new themes treated here. These include Anders Ericsson, with whom I explored the theory and practice of protocol analysis; Pat Langley, Gary Bradshaw, and Jan Zytkow, my co-investigators of the processes of scientific discovery; Yuichiro Anzai, Fernand Gobet, Yumi Iwasaki, Deepak Kulkarni, Jill Larkin, Jean-Louis Le Moigne, Anthony Leonardo, Yulin Qin, Howard Richman, Weimin Shen, Jim Staszewski, Hermina Tabachneck, Guojung Zhang, and Xinming Zhu. In truth, I don't know where to end the list or how to avoid serious gaps in it, so I will simply express my deep thanks to all of my friends and collaborators, both the mentioned and the unmentioned.

In the first chapter I propose that the goal of science is to make the wonderful and the complex understandable and simple—but not less wonderful. I will be pleased if readers find that I have achieved a bit of that in this third edition of *The Sciences of the Artificial.*

Herbert A. Simon
Pittsburgh, Pennsylvania
January 1, 1996

Preface to Second Edition

This work takes the shape of a fugue, whose subject and countersubject were first uttered in lectures on the opposite sides of a continent and the two ends of a decade but are now woven together as the alternating chapters of the whole.

The invitation to deliver the Karl Taylor Compton lectures at the Massachusetts Institute of Technology in the spring of 1968 provided me with a welcome opportunity to make explicit and to develop at some length a thesis that has been central to much of my research, at first in organization theory, later in economics and management science, and most recently in psychology.

In 1980 another invitation, this one to deliver the H. Rowan Gaither lectures at the University of California, Berkeley, permitted me to amend and expand this thesis and to apply it to several additional fields.

The thesis is that certain phenomena are "artificial" in a very specific sense: they are as they are only because of a system's being molded, by goals or purposes, to the environment in which it lives. If natural phenomena have an air of "necessity" about them in their subservience to natural law, artificial phenomena have an air of "contingency" in their malleability by environment.

The contingency of artificial phenomena has always created doubts as to whether they fall properly within the compass of science. Sometimes these doubts refer to the goal-directed character of artificial systems and the consequent difficulty of disentangling prescription from description. This seems to me not to be the real difficulty. The genuine problem is to show how empirical propositions can be made at all about systems that, given different circumstances, might be quite other than they are.

Almost as soon as I began research on administrative organizations, some forty years ago, I encountered the problem of artificiality in almost its pure form:

> . . . administration is not unlike play-acting. The task of the good actor is to know and play his role, although different roles may differ greatly in content. The effectiveness of the performance will depend on the effectiveness of the play and the effectiveness with which it is played. The effectiveness of the administrative process will vary with the effectiveness of the organization and the effectiveness with which its members play their parts. [*Administrative Behavior,* p. 252]

How then could one construct a theory of administration that would contain more than the normative rules of good acting? In particular, how could one construct an empirical theory? My writing on administration, particularly in *Administrative Behavior* and part IV of *Models of Man,* has sought to answer those questions by showing that the empirical content of the phenomena, the necessity that rises above the contingencies, stems from the inabilities of the behavioral system to adapt perfectly to its environment—from the limits of rationality, as I have called them.

As research took me into other areas, it became evident that the problem of artificiality was not peculiar to administration and organizations but that it infected a far wider range of subjects. Economics, since it postulated rationality in economic man, made him the supremely skillful actor, whose behavior could reveal something of the requirements the environment placed on him but nothing about his own cognitive makeup. But the difficulty must then extend beyond economics into all those parts of psychology concerned with rational behavior—thinking, problem solving, learning.

Finally, I thought I began to see in the problem of artificiality an explanation of the difficulty that has been experienced in filling engineering and other professions with empirical and theoretical substance distinct from the substance of their supporting sciences. Engineering, medicine, business, architecture, and painting are concerned not with the necessary but with the contingent—not with how things are but with how they might be—in short, with design. The possibility of creating a science or sciences of design is exactly as great as the possibility of creating any science of the artificial. The two possibilities stand or fall together.

These essays then attempt to explain how a science of the artificial is possible and to illustrate its nature. I have taken as my main examples the

fields of economics (chapter 2), the psychology of cognition (chapters 3 and 4), and planning and engineering design (chapters 5 and 6). Since Karl Compton was a distinguished engineering educator as well as a distinguished scientist, I thought it not inappropriate to apply my conclusions about design to the question of reconstructing the engineering curriculum (chapter 5). Similarily Rowan Gaither's strong interest in the uses of systems analysis in public policy formation is reflected especially in chapter 6.

The reader will discover in the course of the discussion that artificiality is interesting principally when it concerns complex systems that live in complex environments. The topics of artificiality and complexity are inextricably interwoven. For this reason I have included in this volume (chapter 8) an earlier essay, "The Architecture of Complexity," which develops at length some ideas about complexity that I could touch on only briefly in my lectures. The essay appeared originally in the December 1962 *Proceedings of the American Philosophical Society.*

I have tried to acknowledge some specific debts to others in footnotes at appropriate points in the text. I owe a much more general debt to Allen Newell, whose partner I have been in a very large part of my work for more than two decades and to whom I have dedicated this volume. If there are parts of my thesis with which he disagrees, they are probably wrong; but he cannot evade a major share of responsibility for the rest.

Many ideas, particularly in the third and fourth chapters had their origins in work that my late colleague, Lee W. Gregg, and I did together; and other colleagues, as well as numerous present and former graduate students, have left their fingerprints on various pages of the text. Among the latter I want to mention specifically L. Stephen Coles, Edward A. Feigenbaum, John Grason, Pat Langley, Robert K. Lindsay, David Neves, Ross Quillian, Laurent Siklóssy, Donald S. Williams, and Thomas G. Williams, whose work is particularly relevant to the topics discussed here.

Previous versions of chapter 8 incorporated valuable suggestions and data contributed by George W. Corner, Richard H. Meier, John R. Platt, Andrew Schoene, Warren Weaver, and William Wise.

A large part of the psychological research reported in this book was supported by the Public Health Service Research Grant MH-07722 from the National Institute of Mental Health, and some of the research on

design reported in the fifth and sixth chapters, by the Advanced Research Projects Agency of the Office of the Secretary of Defense (SD-146). These grants, as well as support from the Carnegie Corporation, the Ford Foundation, and the Alfred P. Sloan Foundation, have enabled us at Carnegie-Mellon to pursue for over two decades a many-pronged exploration aimed at deepening our understanding of artificial phenomena.

Finally, I am grateful to the Massachusetts Institute of Technology and to the University of California, Berkeley, for the opportunity to prepare and present these lectures and for the occasion to become better acquainted with the research in the sciences of the artificial going forward on these two stimulating campuses.

I want to thank both institutions also for agreeing to the publication of these lectures in this unified form, The Compton lectures comprise chapters 1, 3, and 5, and the Gaither lectures, chapters 2, 4, and 6. Since the first edition of this book (The MIT Press, 1969) has been well received, I have limited the changes in chapters 1, 3, 5, and 8 to the correction of blatant errors, the updating of a few facts, and the addition of some transitional paragraphs.

The Sciences of the Artificial

1

Understanding the Natural and the Artificial Worlds

About three centuries after Newton we are thoroughly familiar with the concept of natural science—most unequivocally with physical and biological science. A natural science is a body of knowledge about some class of things—objects or phenomena—in the world: about the characteristics and properties that they have; about how they behave and interact with each other.

The central task of a natural science is to make the wonderful commonplace: to show that complexity, correctly viewed, is only a mask for simplicity; to find pattern hidden in apparent chaos. The early Dutch physicist Simon Stevin, showed by an elegant drawing (figure 1) that the law of the inclined plane follows in "self-evident fashion" from the impossibility of perpetual motion, for experience and reason tell us that the chain of balls in the figure would rotate neither to right nor to left but would remain at rest. (Since rotation changes nothing in the figure, if the chain moved at all, it would move perpetually.) Since the pendant part of the chain hangs symmetrically, we can snip it off without disturbing the equilibrium. But now the balls on the long side of the plane balance those on the shorter, steeper side, and their relative numbers are in inverse ratio to the sines of the angles at which the planes are inclined.

Stevin was so pleased with his construction that he incorporated it into a vignette, inscribing above it

Wonder, en is gheen wonder

that is to say: "Wonderful, but not incomprehensible."

This is the task of natural science: to show that the wonderful is not incomprehensible, to show how it can be comprehended—but not to

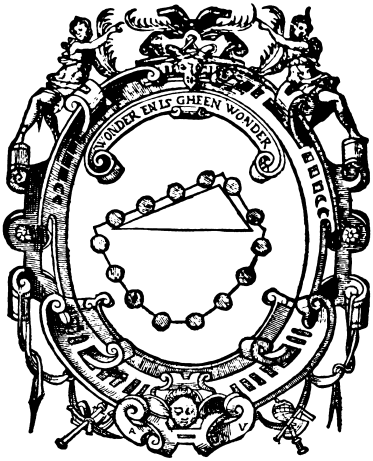

Figure 1
The vignette devised by Simon Stevin to illustrate his derivation of the law of the inclined plane

destroy wonder. For when we have explained the wonderful, unmasked the hidden pattern, a new wonder arises at how complexity was woven out of simplicity. The aesthetics of natural science and mathematics is at one with the aesthetics of music and painting—both inhere in the discovery of a partially concealed pattern.

The world we live in today is much more a man-made,[1] or artificial, world than it is a natural world. Almost every element in our environment shows evidence of human artifice. The temperature in which we spend most of our hours is kept artificially at 20 degrees Celsius; the humidity is added to or taken from the air we breathe; and the impurities we inhale are largely produced (and filtered) by man.

Moreover for most of us—the white-collared ones—the significant part of the environment consists mostly of strings of artifacts called "symbols" that we receive through eyes and ears in the form of written and spoken language and that we pour out into the environment—as I am now doing—by mouth or hand. The laws that govern these strings of

1. I will occasionally use "man" as an androgynous noun, encompassing both sexes, and "he," "his," and "him" as androgynous pronouns including women and men equally in their scope.

symbols, the laws that govern the occasions on which we emit and receive them, the determinants of their content are all consequences of our collective artifice.

One may object that I exaggerate the artificiality of our world. Man must obey the law of gravity as surely as does a stone, and as a living organism man must depend for food, and in many other ways, on the world of biological phenomena. I shall plead guilty to overstatement, while protesting that the exaggeration is slight. To say that an astronaut, or even an airplane pilot, is obeying the law of gravity, hence is a perfectly natural phenomenon, is true, but its truth calls for some sophistication in what we mean by "obeying" a natural law. Aristotle did not think it natural for heavy things to rise or light ones to fall (*Physics,* Book IV); but presumably we have a deeper understanding of "natural" than he did.

So too we must be careful about equating "biological" with "natural." A forest may be a phenomenon of nature; a farm certainly is not. The very species upon which we depend for our food—our corn and our cattle—are artifacts of our ingenuity. A plowed field is no more part of nature than an asphalted street—and no less.

These examples set the terms of our problem, for those things we call artifacts are not apart from nature. They have no dispensation to ignore or violate natural law. At the same time they are adapted to human goals and purposes. They are what they are in order to satisfy our desire to fly or to eat well. As our aims change, so too do our artifacts—and vice versa.

If science is to encompass these objects and phenomena in which human purpose as well as natural law are embodied, it must have means for relating these two disparate components. The character of these means and their implications for certain areas of knowledge—economics, psychology, and design in particular—are the central concern of this book.

The Artificial

Natural science is knowledge about natural objects and phenomena. We ask whether there cannot also be "artificial" science—knowledge about artificial objects and phenomena. Unfortunately the term "artificial" has a pejorative air about it that we must dispel before we can proceed.

My dictionary defines "artificial" as, "Produced by art rather than by nature; not genuine or natural; affected; not pertaining to the essence of the matter." It proposes, as synonyms: affected, factitious, manufactured, pretended, sham, simulated, spurious, trumped up, unnatural. As antonyms, it lists: actual, genuine, honest, natural, real, truthful, unaffected. Our language seems to reflect man's deep distrust of his own products. I shall not try to assess the validity of that evaluation or explore its possible psychological roots. But you will have to understand me as using "artificial" in as neutral a sense as possible, as meaning man-made as opposed to natural.[2]

In some contexts we make a distinction between "artificial" and "synthetic." For example, a gem made of glass colored to resemble sapphire would be called artificial, while a man-made gem chemically indistinguishable from sapphire would be called synthetic. A similar distinction is often made between "artificial" and "synthetic" rubber. Thus some artificial things are imitations of things in nature, and the imitation may use either the same basic materials as those in the natural object or quite different materials.

As soon as we introduce "synthesis" as well as "artifice," we enter the realm of engineering. For "synthetic" is often used in the broader sense of "designed" or "composed." We speak of engineering as concerned with "synthesis," while science is concerned with "analysis." Synthetic or artificial objects—and more specifically prospective artificial objects having desired properties—are the central objective of engineering activity and skill. The engineer, and more generally the designer, is concerned with how things *ought* to be—how they ought to be in order to *attain goals*,

2. I shall disclaim responsibility for this particular choice of terms. The phrase "artificial intelligence," which led me to it, was coined, I think, right on the Charles River, at MIT. Our own research group at Rand and Carnegie Mellon University have preferred phrases like "complex information processing" and "simulation of cognitive processes." But then we run into new terminological difficulties, for the dictionary also says that "to simulate" means "to assume or have the mere appearance or form of, without the reality; imitate; counterfeit; pretend." At any rate, "artificial intelligence" seems to be here to stay, and it may prove easier to cleanse the phrase than to dispense with it. In time it will become sufficiently idiomatic that it will no longer be the target of cheap rhetoric.

and to *function*. Hence a science of the artificial will be closely akin to a science of engineering—but very different, as we shall see in my fifth chapter, from what goes currently by the name of "engineering science."

With goals and "oughts" we also introduce into the picture the dichotomy between normative and descriptive. Natural science has found a way to exclude the normative and to concern itself solely with how things are. Can or should we maintain this exclusion when we move from natural to artificial phenomena, from analysis to synthesis?[3]

We have now identified four indicia that distinguish the artificial from the natural; hence we can set the boundaries for sciences of the artificial:

1. Artificial things are synthesized (though not always or usually with full forethought) by human beings.
2. Artificial things may imitate appearances in natural things while lacking, in one or many respects, the reality of the latter.
3. Artificial things can be characterized in terms of functions, goals, adaptation.
4. Artificial things are often discussed, particularly when they are being designed, in terms of imperatives as well as descriptives.

The Environment as Mold

Let us look a little more closely at the functional or purposeful aspect of artificial things. Fulfillment of purpose or adaptation to a goal involves a relation among three terms: the purpose or goal, the character of the artifact, and the environment in which the artifact performs. When we think of a clock, for example, in terms of purpose we may use the child's definition: "a clock is to tell time." When we focus our attention on the clock itself, we may describe it in terms of arrangements of gears and the

3. This issue will also be discussed at length in my fifth chapter. In order not to keep readers in suspense, I may say that I hold to the pristine empiricist's position of the irreducibility of "ought" to "is," as in chapter 3 of my *Administrative Behavior* (New York: Macmillan, 1976). This position is entirely consistent with treating natural or artificial goal-seeking systems as phenomena, without commitment to their goals. *Ibid.,* appendix. See also the well-known paper by A. Rosenbluth, N. Wiener, and J. Bigelow, "Behavior, Purpose, and Teleology," *Philosophy of Science, 10* (1943):18–24.

application of the forces of springs or gravity operating on a weight or pendulum.

But we may also consider clocks in relation to the environment in which they are to be used. Sundials perform as clocks *in sunny climates*—they are more useful in Phoenix than in Boston and of no use at all during the Arctic winter. Devising a clock that would tell time on a rolling and pitching ship, with sufficient accuracy to determine longitude, was one of the great adventures of eighteenth-century science and technology. To perform in this difficult environment, the clock had to be endowed with many delicate properties, some of them largely or totally irrelevant to the performance of a landlubber's clock.

Natural science impinges on an artifact through two of the three terms of the relation that characterizes it: the structure of the artifact itself and the environment in which it performs. Whether a clock will in fact tell time depends on its internal construction and where it is placed. Whether a knife will cut depends on the material of its blade and the hardness of the substance to which it is applied.

The Artifact as "Interface"

We can view the matter quite symmetrically. An artifact can be thought of as a meeting point—an "interface" in today's terms—between an "inner" environment, the substance and organization of the artifact itself, and an "outer" environment, the surroundings in which it operates. If the inner environment is appropriate to the outer environment, or vice versa, the artifact will serve its intended purpose. Thus, if the clock is immune to buffeting, it will serve as a ship's chronometer. (And conversely, if it isn't, we may salvage it by mounting it on the mantel at home.)

Notice that this way of viewing artifacts applies equally well to many things that are not man-made—to all things in fact that can be regarded as adapted to some situation; and in particular it applies to the living systems that have evolved through the forces of organic evolution. A theory of the airplane draws on natural science for an explanation of its inner environment (the power plant, for example), its outer environment (the character of the atmosphere at different altitudes), and the relation between its inner and outer environments (the movement of an airfoil

through a gas). But a theory of the bird can be divided up in exactly the same way.[4]

Given an airplane, or *given* a bird, we can analyze them by the methods of natural science without any particular attention to purpose or adaptation, without reference to the interface between what I have called the inner and outer environments. After all, their behavior is governed by natural law just as fully as the behavior of anything else (or at least we all believe this about the airplane, and most of us believe it about the bird).

Functional Explanation

On the other hand, if the division between inner and outer environment is not necessary to the analysis of an airplane or a bird, it turns out at least to be highly convenient. There are several reasons for this, which will become evident from examples.

Many animals in the Arctic have white fur. We usually explain this by saying that white is the best color for the Arctic environment, for white creatures escape detection more easily than do others. This is not of course a natural science explanation; it is an explanation by reference to purpose or function. It simply says that these are the kinds of creatures that will "work," that is, survive, in this kind of environment. To turn the statement into an explanation, we must add to it a notion of natural selection, or some equivalent mechanism.

An important fact about this kind of explanation is that it demands an understanding mainly of the outer environment. Looking at our snowy surroundings, we can predict the predominant color of the creatures we are likely to encounter; we need know little about the biology of the creatures themselves, beyond the facts that they are often mutually hostile, use visual clues to guide their behavior, and are adaptive (through selection or some other mechanism).

4. A generalization of the argument made here for the separability of "outer" from "inner" environment shows that we should expect to find this separability, to a greater or lesser degree, in *all* large and complex systems, whether they are artificial or natural. In its generalized form it is an argument that all nature will be organized in "levels." My essay "The Architecture of Complexity," included in this volume as chapter 8, develops the more general argument in some detail.

Analogous to the role played by natural selection in evolutionary biology is the role played by rationality in the sciences of human behavior. If we know of a business organization only that it is a profit-maximizing system, we can often predict how its behavior will change if we change its environment—how it will alter its prices if a sales tax is levied on its products. We can sometimes make this prediction—and economists do make it repeatedly—without detailed assumptions about the adaptive mechanism, the decision-making apparatus that constitutes the inner environment of the business firm.

Thus the first advantage of dividing outer from inner environment in studying an adaptive or artificial system is that we can often predict behavior from knowledge of the system's goals and its outer environment, with only minimal assumptions about the inner environment. An instant corollary is that we often find quite different inner environments accomplishing identical or similar goals in identical or similar outer environments—airplanes and birds, dolphins and tunafish, weight-driven clocks and battery-driven clocks, electrical relays and transistors.

There is often a corresponding advantage in the division from the standpoint of the inner environment. In very many cases whether a particular system will achieve a particular goal or adaptation depends on only a few characteristics of the outer environment and not at all on the detail of that environment. Biologists are familiar with this property of adaptive systems under the label of homeostasis. It is an important property of most good designs, whether biological or artifactual. In one way or another the designer insulates the inner system from the environment, so that an invariant relation is maintained between inner system and goal, independent of variations over a wide range in most parameters that characterize the outer environment. The ship's chronometer reacts to the pitching of the ship only in the negative sense of maintaining an invariant relation of the hands on its dial to the real time, independently of the ship's motions.

Quasi independence from the outer environment may be maintained by various forms of passive insulation, by reactive negative feedback (the most frequently discussed form of insulation), by predictive adaptation, or by various combinations of these.

Functional Description and Synthesis

In the best of all possible worlds—at least for a designer—we might even hope to combine the two sets of advantages we have described that derive from factoring an adaptive system into goals, outer environment, and inner environment. We might hope to be able to characterize the main properties of the system and its behavior without elaborating the detail of *either* the outer or inner environments. We might look toward a science of the artificial that would depend on the relative simplicity of the interface as its primary source of abstraction and generality.

Consider the design of a physical device to serve as a counter. If we want the device to be able to count up to one thousand, say, it must be capable of assuming any one of at least a thousand states, of maintaining itself in any given state, and of shifting from any state to the "next" state. There are dozens of different inner environments that might be used (and have been used) for such a device. A wheel notched at each twenty minutes of arc, and with a ratchet device to turn and hold it, would do the trick. So would a string of ten electrical switches properly connected to represent binary numbers. Today instead of switches we are likely to use transistors or other solid-state devices.[5]

Our counter would be activated by some kind of pulse, mechanical or electrical, as appropriate, from the outer environment. But by building an appropriate transducer between the two environments, the physical character of the interior pulse could again be made independent of the physical character of the exterior pulse—the counter could be made to count anything.

Description of an artifice in terms of its organization and functioning—its interface between inner and outer environments—is a major objective of invention and design activity. Engineers will find familiar the language of the following claim quoted from a 1919 patent on an improved motor controller:

> What I claim as new and desire to secure by Letters Patent is:
> 1 In a motor controller, in combination, reversing means, normally effective field-weakening means and means associated with said reversing means for

5. The theory of functional equivalence of computing machines has had considerable development in recent years. See Marvin L. Minsky, *Computation: Finite and Infinite Machines* (Englewood Cliffs, N.J.: Prentice-Hall, 1967), chapters 1–4.

rendering said field-weakening means ineffective during motor starting and thereafter effective to different degrees determinable by the setting of said reversing means . . . [6]

Apart from the fact that we know the invention relates to control of an electric motor, there is almost no reference here to specific, concrete objects or phenomena. There is reference rather to "reversing means" and "field-weakening means," whose further purpose is made clear in a paragraph preceding the patent claims:

The advantages of the special type of motor illustrated and the control thereof will be readily understood by those skilled in the art. Among such advantages may be mentioned the provision of a high starting torque and the provision for quick reversals of the motor.[7]

Now let us suppose that the motor in question is incorporated in a planing machine (see figure 2). The inventor describes its behavior thus:

Referring now to [figure 2], the controller is illustrated in outline connection with a planer (100) operated by a motor M, the controller being adapted to govern the motor M and to be automatically operated by the reciprocating bed (101) of the planer. The master shaft of the controller is provided with a lever (102) connected by a link (103) to a lever (104) mounted upon the planer frame and projecting into the path of lugs (105) and (106) on the planer bed. As will be understood, the arrangement is such that reverse movements of the planer bed will, through the connections described, throw the master shaft of the controller back and forth between its extreme positions and in consequence effect selective operation of the reversing switches (1) and (2) and automatic operation of the other switches in the manner above set forth.[8]

In this manner the properties with which the inner environment has been endowed are placed at the service of the goals in the context of the outer environment. The motor will reverse periodically under the control of the position of the planer bed. The "shape" of its behavior—the time path, say, of a variable associated with the motor—will be a function of the "shape" of the external environment—the distance, in this case, between the lugs on the planer bed.

The device we have just described illustrates in microcosm the nature of artifacts. Central to their description are the goals that link the inner

6. U.S. Patent 1,307,836, granted to Arthur Simon, June 24, 1919.

7. Ibid.

8. Ibid.

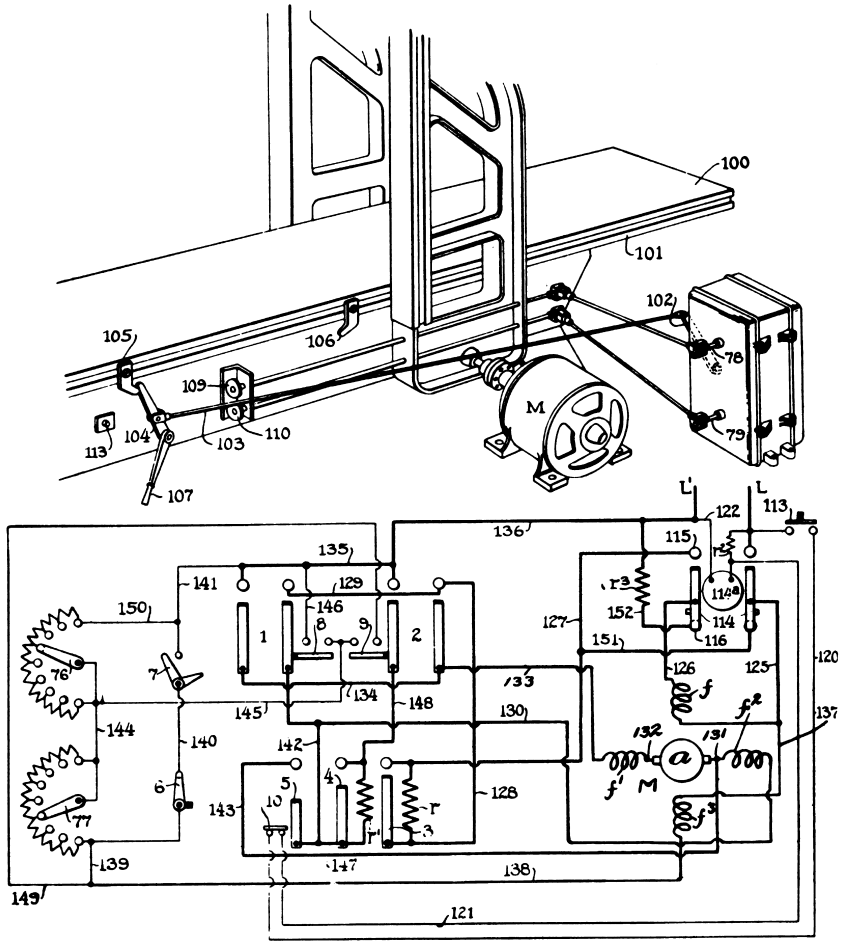

Figure 2
Illustrations from a patent for a motor controller

to the outer system. The inner system is an organization of natural phenomena capable of attaining the goals in some range of environments, but ordinarily there will be many functionally equivalent natural systems capable of doing this.

The outer environment determines the conditions for goal attainment. If the inner system is properly designed, it will be adapted to the outer environment, so that its behavior will be determined in large part by the

behavior of the latter, exactly as in the case of "economic man." To predict how it will behave, we need only ask, "How would a rationally designed system behave under these circumstances?" The behavior takes on the shape of the task environment.[9]

Limits of Adaptation

But matters must be just a little more complicated than this account suggests. "If wishes were horses, all beggars would ride." And if we could always specify a protean inner system that would take on exactly the shape of the task environment, designing would be synonymous with wishing. "Means for scratching diamonds" defines a design objective, an objective that *might* be attained with the use of many different substances. But the design has not been achieved until we have discovered at least one realizable inner system obeying the ordinary natural laws—one material, in this case, hard enough to scratch diamonds.

Often we shall have to be satisfied with meeting the design objectives only approximately. Then the properties of the inner system will "show through." That is, the behavior of the system will only partly respond to the task environment; partly, it will respond to the limiting properties of the inner system.

Thus the motor controls described earlier are aimed at providing for "quick" reversal of the motor. But the motor must obey electromagnetic and mechanical laws, and we could easily confront the system with a task where the environment called for quicker reversal than the motor was capable of. In a benign environment we would learn from the motor only what it had been called upon to do; in a taxing environment we would learn something about its internal structure—specifically about those aspects of the internal structure that were chiefly instrumental in limiting performance.[10]

9. On the crucial role of adaptation or rationality—and their limits—for economics and organization theory, see the introduction to part IV, "Rationality and Administrative Decision Making," of my *Models of Man* (New York: Wiley, 1957); pp. 38–41, 80–81, and 240–244 of *Administrative Behavior;* and chapter 2 of this book.

10. Compare the corresponding proposition on the design of administrative organizations: "Rationality, then, does not determine behavior. Within the area of rationality behavior is perfectly flexible and adaptable to abilities, goals, and

A bridge, under its usual conditions of service, behaves simply as a relatively smooth level surface on which vehicles can move. Only when it has been overloaded do we learn the physical properties of the materials from which it is built.

Understanding by Simulating

Artificiality connotes perceptual similarity but essential difference, resemblance from without rather than within. In the terms of the previous section we may say that the artificial object imitates the real by turning the same face to the outer system, by adapting, relative to the same goals, to comparable ranges of external tasks. Imitation is possible because distinct physical systems can be organized to exhibit nearly identical behavior. The damped spring and the damped circuit obey the same second-order linear differential equation; hence we may use either one to imitate the other.

Techniques of Simulation

Because of its abstract character and its symbol manipulating generality, the digital computer has greatly extended the range of systems whose behavior can be imitated. Generally we now call the imitation "simulation," and we try to understand the imitated system by testing the simulation in a variety of simulated, or imitated, environments.

Simulation, as a technique for achieving understanding and predicting the behavior of systems, predates of course the digital computer. The model basin and the wind tunnel are valued means for studying the behavior of large systems by modeling them in the small, and it is quite certain that Ohm's law was suggested to its discoverer by its analogy with simple hydraulic phenomena.

---

knowledge. Instead, behavior is determined by the irrational and nonrational elements that bound the area of rationality . . . administrative theory must be concerned with the limits of rationality, and the manner in which organization affects these limits for the person making a decision." *Administrative Behavior,* p. 241. For a discussion of the same issue as it arises in psychology, see my "Cognitive Architectures and Rational Analysis: Comment," in Kurt VanLehn (ed.), *Architectures for Intelligence* (Hillsdale, NJ: Erlbaum, 1991).

Simulation may even take the form of a thought experiment, never actually implemented dynamically. One of my vivid memories of the Great Depression is of a large multicolored chart in my father's study that represented a hydraulic model of an economic system (with different fluids for money and goods). The chart was devised by a technocratically inclined engineer named Dahlberg. The model never got beyond the pen-and-paint stage at that time, but it could be used to trace through the imputed consequences of particular economic measures or events—provided the theory was right![11]

As my formal education in economics progressed, I acquired a disdain for that naive simulation, only to discover after World War II that a distinguished economist, Professor A. W. Phillips had actually built the Moniac, a hydraulic model that simulated a Keynesian economy.[12] Of course Professor Phillips's simulation incorporated a more nearly correct theory than the earlier one and was actually constructed and operated—two points in its favor. However, the Moniac, while useful as a teaching tool, told us nothing that could not be extracted readily from simple mathematical versions of Keynesian theory and was soon priced out of the market by the growing number of computer simulations of the economy.

Simulation as a Source of New Knowledge

This brings me to the crucial question about simulation: *How can a simulation ever tell us anything that we do not already know?* The usual implication of the question is that it can't. As a matter of fact, there is an interesting parallelism, which I shall exploit presently, between two assertions about computers and simulation that one hears frequently:

1. A simulation is no better than the assumptions built into it.
2. A computer can do only what it is programmed to do.

I shall not deny either assertion, for both seem to me to be true. But despite both assertions simulation can tell us things we do not already know.

11. For some published versions of this model, see A. O. Dahlberg, *National Income Visualized* (N.Y.: Columbia University Press, 1956).

12. A. W. Phillips, "Mechanical Models in Economic Dynamics," *Economica*, New Series, *17* (1950):283–305.

There are two related ways in which simulation can provide new knowledge—one of them obvious, the other perhaps a bit subtle. The obvious point is that, even when we have correct premises, it may be very difficult to discover what they imply. All correct reasoning is a grand system of tautologies, but only God can make direct use of that fact. The rest of us must painstakingly and fallibly tease out the consequences of our assumptions.

Thus we might expect simulation to be a powerful technique for deriving, from our knowledge of the mechanisms governing the behavior of gases, a theory of the weather and a means of weather prediction. Indeed, as many people are aware, attempts have been under way for some years to apply this technique. Greatly oversimplified, the idea is that we already know the correct basic assumptions, the local atmospheric equations, but we need the computer to work out the implications of the interactions of vast numbers of variables starting from complicated initial conditions. This is simply an extrapolation to the scale of modern computers of the idea we use when we solve two simultaneous equations by algebra.

This approach to simulation has numerous applications to engineering design. For it is typical of many kinds of design problems that the inner system consists of components whose fundamental laws of behavior—mechanical, electrical, or chemical—are well known. The difficulty of the design problem often resides in predicting how an assemblage of such components will behave.

Simulation of Poorly Understood Systems

The more interesting and subtle question is whether simulation can be of any help to us when we do not know very much initially about the natural laws that govern the behavior of the inner system. Let me show why this question must also be answered in the affirmative.

First, I shall make a preliminary comment that simplifies matters: we are seldom interested in explaining or predicting phenomena in all their particularity; we are usually interested only in a few properties abstracted from the complex reality. Thus, a NASA-launched satellite is surely an artificial object, but we usually do not think of it as "simulating" the moon or a planet. It simply obeys the same laws of physics, which relate

only to its inertial and gravitational mass, abstracted from most of its other properties. It *is* a moon. Similarly electric energy that entered my house from the early atomic generating station at Shippingport did not "simulate" energy generated by means of a coal plant or a windmill. Maxwell's equations hold for both.

The more we are willing to abstract from the detail of a set of phenomena, the easier it becomes to simulate the phenomena. Moreover we do not have to know, or guess at, all the internal structure of the system but only that part of it that is crucial to the abstraction.

It is fortunate that this is so, for if it were not, the topdown strategy that built the natural sciences over the past three centuries would have been infeasible. We knew a great deal about the gross physical and chemical behavior of matter before we had a knowledge of molecules, a great deal about molecular chemistry before we had an atomic theory, and a great deal about atoms before we had any theory of elementary particles—if indeed we have such a theory today.

This skyhook-skyscraper construction of science from the roof down to the yet unconstructed foundations was possible because the behavior of the system at each level depended on only a very approximate, simplified, abstracted characterization of the system at the level next beneath.[13] This is lucky, else the safety of bridges and airplanes might depend on the correctness of the "Eightfold Way" of looking at elementary particles.

Artificial systems and adaptive systems have properties that make them particularly susceptible to simulation via simplified models. The characterization of such systems in the previous section of this chapter

13. This point is developed more fully in "The Architecture of Complexity," chapter 8 in this volume. More than fifty years ago, Bertrand Russell made the same point about the architecture of mathematics. See the "Preface" to *Principia Mathematica:* ". . . the chief reason in favour of any theory on the principles of mathematics must always be inductive, i.e., it must lie in the fact that the theory in question enables us to deduce ordinary mathematics. In mathematics, the greatest degree of self-evidence is usually not to be found quite at the beginning, but at some later point; hence the early deductions, until they reach this point, give reasons rather for believing the premises because true consequences follow from them, than for believing the consequences because they follow from the premises." Contemporary preferences for deductive formalisms frequently blind us to this important fact, which is no less true today than it was in 1910.

explains why. Resemblance in behavior of systems without identity of the inner systems is particularly feasible if the aspects in which we are interested arise out of the *organization* of the parts, independently of all but a few properties of the individual components. Thus for many purposes we may be interested in only such characteristics of a material as its tensile and compressive strength. We may be profoundly unconcerned about its chemical properties, or even whether it is wood or iron.

The motor control patent cited earlier illustrates this abstraction to organizational properties. The invention consisted of a "combination" of "reversing means," of "field weakening means," that is to say, of components specified in terms of their functioning in the organized whole. How many ways are there of reversing a motor, or of weakening its field strength? We can simulate the system described in the patent claims in many ways without reproducing even approximately the actual physical device that is depicted. With a small additional step of abstraction, the patent claims could be restated to encompass mechanical as well as electrical devices. I suppose that any undergraduate engineer at Berkeley, Carnegie Mellon University, or MIT could design a mechanical system embodying reversibility and variable starting torque so as to simulate the system of the patent.

The Computer as Artifact

No artifact devised by man is so convenient for this kind of functional description as a digital computer. It is truly protean, for almost the only ones of its properties that are detectable in its behavior (when it is operating properly!) are the organizational properties. The speed with which it performs it basic operations may allow us to infer a little about its physical components and their natural laws; speed data, for example, would allow us to rule out certain kinds of "slow" components. For the rest, almost no interesting statement that one can make about an operating computer bears any particular relation to the specific nature of the hardware. A computer is an organization of elementary functional components in which, to a high approximation, only the function

performed by those components is relevant to the behavior of the whole system.[14]

Computers as Abstract Objects

This highly abstractive quality of computers makes it easy to introduce mathematics into the study of their theory—and has led some to the erroneous conclusion that, as a computer science emerges, it will necessarily be a mathematical rather than an empirical science. Let me take up these two points in turn: the relevance of mathematics to computers and the possibility of studying computers empirically.

Some important theorizing, initiated by John von Neumann, has been done on the topic of computer reliability. The question is how to build a reliable system from unreliable parts. Notice that this is not posed as a question of physics or physical engineering. The components engineer is assumed to have done his best, but the parts are still unreliable! We can cope with the unreliability only by our manner of organizing them.

To turn this into a meaningful problem, we have to say a little more about the nature of the unreliable parts. Here we are aided by the knowledge that *any* computer can be assembled out of a small array of simple, basic elements. For instance, we may take as our primitives the so-called Pitts-McCulloch neurons. As their name implies, these components were devised in analogy to the supposed anatomical and functional characteristics of neurons in the brain, but they are highly abstracted. They are formally isomorphic with the simplest kinds of switching circuits—"and," "or," and "not" circuits. We postulate, now, that we are to build a system from such elements and that each elementary part has a specified probability of functioning correctly. The problem is to arrange the elements and their interconnections in such a way that the complete system will perform reliably.

The important point for our present discussion is that the parts could as well be neurons as relays, as well relays as transistors. The natural laws governing relays are very well known, while the natural laws governing

14. On the subject of this and the following paragraphs, see M. L. Minsky, *op. cit.;* then John von Neumann, "Probabilistic Logics and the Synthesis of Reliable Organisms from Unreliable Components," in C. E. Shannon and J. McCarthy (eds.), *Automata Studies* (Princeton: Princeton University Press, 1956).

neurons are known most imperfectly. But that does not matter, for all that is relevant for the theory is that the components have the specified level of unreliability and be interconnected in the specified way.

This example shows that the possibility of building a mathematical theory of a system or of simulating that system does not depend on having an adequate microtheory of the natural laws that govern the system components. Such a microtheory might indeed be simply irrelevant.

Computers as Empirical Objects

We turn next to the feasibility of an *empirical* science of computers—as distinct from the solid-state physics or physiology of their componentry.[15] As a matter of empirical fact almost all of the computers that have been designed have certain common organizational features. They almost all can be decomposed into an active processor (Babbage's "Mill") and a memory (Babbage's "Store") in combination with input and output devices. (Some of the larger systems, somewhat in the manner of colonial algae, are assemblages of smaller systems having some or all of these components. But perhaps I may oversimplify for the moment.) They are all capable of storing symbols (program) that can be interpreted by a program-control component and executed. Almost all have exceedingly limited capacity for simultaneous, parallel activity—they are basically one-thing-at-a-time systems. Symbols generally have to be moved from the larger memory components into the central processor before they can be acted upon. The systems are capable of only simple basic actions: recoding symbols, storing symbols, copying symbols, moving symbols, erasing symbols, and comparing symbols.

Since there are now many such devices in the world, and since the properties that describe them also appear to be shared by the human central nervous system, nothing prevents us from developing a natural history of them. We can study them as we would rabbits or chipmunks and discover how they behave under different patterns of environmental stimulation. Insofar as their behavior reflects largely the broad functional

15. A. Newell and H. A. Simon, "Computer Science as Empirical Inquiry," *Communications of the ACM, 19*(March 1976):113–126. See also H. A. Simon, "Artificial Intelligence: An Empirical Science," *Artificial Intelligence, 77*(1995): 95–127.

characteristics we have described, and is independent of details of their hardware, we can build a general—but empirical—theory of them.

The research that was done to design computer time-sharing systems is a good example of the study of computer behavior as an empirical phenomenon. Only fragments of theory were available to guide the design of a time-sharing system or to predict how a system of a specified design would actually behave in an environment of users who placed their several demands upon it. Most actual designs turned out initially to exhibit serious deficiencies, and most predictions of performance were startlingly inaccurate.

Under these circumstances the main route open to the development and improvement of time-sharing systems was to build them and see how they behaved. And this is what was done. They were built, modified, and improved in successive stages. Perhaps theory could have anticipated these experiments and made them unnecessary. In fact it didn't, and I don't know anyone intimately acquainted with these exceedingly complex systems who has very specific ideas as to how it might have done so. To understand them, the systems had to be constructed, and their behavior observed.[16]

In a similar vein computer programs designed to play games or to discover proofs for mathematical theorems spend their lives in exceedingly large and complex task environments. Even when the programs themselves are only moderately large and intricate (compared, say, with the monitor and operating systems of large computers), too little is known about their task environments to permit accurate prediction of how well they will perform, how selectively they will be able to search for problem solutions.

Here again theoretical analysis must be accompanied by large amounts of experimental work. A growing literature reporting these experiments is beginning to give us precise knowledge about the degree of heuristic power of particular heuristic devices in reducing the size of the problem spaces that must be searched. In theorem proving, for example, there has

16. The empirical, exploratory flavor of computer research is nicely captured by the account of Maurice V. Wilkes in his 1967 Turing Lecture, "Computers Then and Now," *Journal of the Association for Computing Machinery, 15*(January 1968):1–7.

been a whole series of advances in heuristic power based on and guided by empirical exploration: the use of the Herbrand theorem, the resolution principle, the set-of-support principle, and so on.[17]

Computers and Thought

As we succeed in broadening and deepening our knowledge—theoretical and empirical—about computers, we discover that in large part their behavior is governed by simple general laws, that what appeared as complexity in the computer program was to a considerable extent complexity of the environment to which the program was seeking to adapt its behavior.

This relation of program to environment opened up an exceedingly important role for computer simulation as a tool for achieving a deeper understanding of human behavior. For if it is the organization of components, and not their physical properties, that largely determines behavior, and if computers are organized somewhat in the image of man, then the computer becomes an obvious device for exploring the consequences of alternative organizational assumptions for human behavior. Psychology could move forward without awaiting the solutions by neurology of the problems of component design—however interesting and significant these components turn out to be.

Symbol Systems: Rational Artifacts

The computer is a member of an important family of artifacts called symbol systems, or more explicitly, physical symbol systems.[18] Another important member of the family (some of us think, anthropomorphically, it is the *most* important) is the human mind and brain. It is with this family

17. Note, for example, the empirical data in Lawrence Wos, George A. Robinson, Daniel F. Carson, and Leon Shalla, "The Concept of Demodulation in Theorem Proving," *Journal of the Association for Computing Machinery, 14*(October 1967):698–709, and in several of the earlier papers referenced there. See also the collection of programs in Edward Feigenbaum and Julian Feldman (eds.), *Computers and Thought* (New York: McGraw-Hill, 1963). It is common practice in the field to title papers about heuristic programs, "Experiments with an *XYZ* Program."

18. In the literature the phrase *information-processing system* is used more frequently than symbol system. I will use the two terms as synonyms.

of artifacts, and particularly the human version of it, that we will be primarily concerned in this book. Symbol systems are almost the quintessential artifacts, for adaptivity to an environment is their whole *raison d'être*. They are goal-seeking, information-processing systems, usually enlisted in the service of the larger systems in which they are incorporated.

Basic Capabilities of Symbol Systems

A physical symbol system holds a set of entities, called symbols. These are physical patterns (e.g., chalk marks on a blackboard) that can occur as components of symbol structures (sometimes called "expressions"). As I have already pointed out in the case of computers, a symbol system also possesses a number of simple processes that operate upon symbol structures—processes that create, modify, copy, and destroy symbols. A physical symbol system is a machine that, as it moves through time, produces an evolving collection of symbol structures.[19] Symbol structures can, and commonly do, serve as internal representations (e.g., "mental images") of the environments to which the symbol system is seeking to adapt. They allow it to model that environment with greater or less veridicality and in greater or less detail, and consequently to reason about it. Of course, for this capability to be of any use to the symbol system, it must have windows on the world and hands, too. It must have means for acquiring information from the external environment that can be encoded into internal symbols, as well as means for producing symbols that initiate action upon the environment. Thus it must use symbols to *designate* objects and relations and actions in the world external to the system.

Symbols may also designate processes that the symbol system can interpret and execute. Hence the programs that govern the behavior of a symbol system can be stored, along with other symbol structures, in the system's own memory, and executed when activated.

Symbol systems are called "physical" to remind the reader that they exist as real-world devices, fabricated of glass and metal (computers) or flesh and blood (brains). In the past we have been more accustomed to thinking of the symbol systems of mathematics and logic as abstract and disembodied, leaving out of account the paper and pencil and human minds that were required actually to bring them to life. Computers have

19. Newell and Simon, "Computer Science as Empirical Inquiry," p. 116.

transported symbol systems from the platonic heaven of ideas to the empirical world of actual processes carried out by machines or brains, or by the two of them working together.

Intelligence as Computation

The three chapters that follow rest squarely on the hypothesis that intelligence is the work of symbol systems. Stated a little more formally, the hypothesis is that a physical symbol system of the sort I have just described has the necessary and sufficient means for general intelligent action.

The hypothesis is clearly an empirical one, to be judged true or false on the basis of evidence. One task of chapters 3 and 4 will be to review some of the evidence, which is of two basic kinds. On the one hand, by constructing computer programs that are demonstrably capable of intelligent action, we provide evidence on the sufficiency side of the hypothesis. On the other hand, by collecting experimental data on human thinking that tend to show that the human brain operates as a symbol system, we add plausibility to the claims for necessity, for such data imply that all known intelligent systems (brains and computers) are symbol systems.

Economics: Abstract Rationality

As prelude to our consideration of human intelligence as the work of a physical symbol system, chapter 2 introduces a heroic abstraction and idealization—the idealization of human rationality which is enshrined in modern economic theories, particularly those called neoclassical. These theories are an idealization because they direct their attention primarily to the external environment of human thought, to decisions that are optimal for realizing the adaptive system's goals (maximization of utility or profit). They seek to define the decisions that would be substantively rational in the circumstances defined by the outer environment.

Economic theory's treatment of the limits of rationality imposed by the inner environment—by the characteristics of the physical symbol system—tends to be pragmatic, and sometimes even opportunistic. In the more formal treatments of general equilibrium and in the so-called "rational expectations" approach to adaptation, the possibilities that an information-processing system may have a very limited capability for

adaptation are almost ignored. On the other hand, in discussions of the rationale for market mechanisms and in many theories of decision making under uncertainty, the procedural aspects of rationality receive more serious treatment.

In chapter 2 we will see examples both of neglect for and concern with the limits of rationality. From the idealizations of economics (and some criticisms of these idealizations) we will move, in chapters 3 and 4, to a more systematic study of the inner environment of thought—of thought processes as they actually occur within the constraints imposed by the parameters of a physical symbol system like the brain.