chapter 1

Introduction: Cognitive Architecture and the Hope for a Science of Cognition*

Zenon W. Pylyshyn
zenon@ruccs.rutgers.edu
Center for Cognitive Science
Rutgers University

The research pursuits that are collectively referred to as *Cognitive Science*, or sometimes (unfortunately, in my view) as *The Cognitive Sciences*, are founded on certain (often tacit) assumptions or research-shaping hypotheses. One such assumption is that there are principles of functioning of the Cognizing Mind/Brain which, if they can be captured at all, will need to be stated in a vocabulary that is proprietary to Cognitive Science—that such principles cannot be stated in the vocabulary of biology or behaviorism, or any of the existing sciences like physics and chemistry; or technological pursuits such as engineering, communications theory, and so on. Another closely related assumption is that the understanding of mind is a task that is beyond the purview of any one of the existing disciplines—because of limitations of formal theoretical apparatus available to any one discipline, and also because relevant evidence is expected to come from a wide range of sources. A third assumption is that there are certain ancient puzzles about cognition and intelligent action that can no longer

*The conference and subsequent preparation of this volume was funded by grants from the National Science Foundation (Grant NSF BNS 91-11423), the Air Force Office of Scientific Research (Grant MIPR#91-0023) and the Office of Naval Research (Grant N-00014-91-J-1851), as well as the support of Rutgers, the State University of New Jersey, and its new Center for Cognitive Science on the New Brunswick campus. I wish to thank Carl Gillett and Paul Lodge for their help in preparing the papers for publication. Information on the Center may be obtained by writing to the director at Rutgers Center for Cognitive Science, New Brunswick, NJ 08903, or through the Internet under Rutgers University (info.rutgers.edu, or use the World Wide Web and access URL http://ruccs.rutgers.edu).

be swept under the rug: there are problems about meaning, about intentionality, and perhaps even about consciousness that either need to be addressed directly or at least circumnavigated (or gerrymandered) in a perspicuous and revealing way.

Historically, however, one additional idea brought researchers together under the new disposition and also provided the added hope for the rehabilitation of certain ancient puzzles. This is the idea of *cognition as computing*. Initially much was not clear about this union of a substantive empirical discipline with a formal (sometimes engineering) tool. The notion of computing itself was not well understood (and to some extent it is still an evolving idea today), and the correspondence between cognition and computation initially was assumed to be a weak one based on pragmatic considerations. The term often used to refer to the pursuit of the marriage between computing and cognition was "computer simulation of behavior." Slowly, however, it dawned on many researchers—I believe beginning with Newell and Simon—that more was at stake than the use of computers to execute models to derive predictions. What was special about computing is that it represented a new type of process: A process that could be described not only in abstract automata, symbol-manipulation, or information-processing terms, but at the same time had two additional properties that brought it into contact with long-standing deep problems in psychology and the philosophy of mind. In computing we had an instance of a process that could be described in terms of *rules and representations*—in terms of what its states were *about*, and at the same time avoiding the excesses of dualism by virtue of the fact that these processes were demonstrably instantiated in a physical system.

The recognition that a stronger level of correspondence between computing and cognition might exist has taken many years to dawn, and many scholars do not accept the general thesis of a deep equivalence between these two types of processes. Some of those who oppose the identification of computation and cognition continue to refer to themselves as Cognitive Scientists. Nonetheless I believe that the fundamental idea that unites cognitive scientists remains this recognition of cognition as a species of computation—at least as a working hypothesis (which is as certain as we get in science anyway). Certainly it is one of the main underlying assumptions of researchers at the Rutgers Center for Cognitive Science, where the papers included in this volume originated.

One of the reasons people are wary of accepting cognition literally as a computational process is that it seems clear that the mind/brain is quite different from typical commercial computers. This much is not controversial. But if the mind really was a computing system, the fact that it seems different from an IBM PC or other commercial type of machine could either mean that we have assumed the wrong level of abstraction for the

comparison, or it could mean that the mind is in fact quite a different type of computer. These two possibilities are closely related and this is where the notion of computational or cognitive architecture becomes central.

Complex systems can always be described at various levels of abstraction. However, separate laws or principles need not exist at all of these levels. Which levels correspond to real natural kinds is an empirical matter. Cognitive science and much of AI rests on the assumption that there is an independent level of organization, which I have called the "semantic level" and Newell has characterized as the "knowledge level." This is a level of organization at which semantic principles, such as rationality or at least plausible reasoning, apply. In the case of a computational process that purports to be a model of some cognitive process, only one level of the system's organization corresponds to what we call its cognitive architecture. That is the level at which the states (datastructures) being processed are the ones that receive a cognitive interpretation. To put it another way, it is the level at which the system is representational, and where the representations correspond to the objects of thought (including percepts, memories, goals, beliefs, and so on). Notice that many other levels of system organization may be below this, but these do not constitute different *cognitive* architectures because their states do not represent cognitive contents. Rather, they correspond to various kinds of implementations, perhaps at the level of some abstract neurology, which realize (or implement) the cognitive architecture. Similarly, various levels of organization may be above this, but they too do not constitute different cognitive architectures. They represent the organization of the cognitive process itself, say in terms of hierarchies of subroutines, not a different level of the system structure.

The notion of *Cognitive Architecture* in the context of the Computational Theory of Mind (CTM) comes directly from the notion of computer architecture in computer science, where it refers to the relatively fixed set of computational resources available to a programmer in designing a program for a given computer system. Among other properties, this includes the type of memory that the computer has, the way it encodes information (the system of symbolic codes or language it uses), the basic operations that are available, and the constraints on the application of these operations (e.g., serial vs. parallel sequencing). The architecture characterizes the computer system on which the program runs, but it may reflect its physical properties only indirectly since the architecture visible to the programmer might itself be simulated in software or firmware. For this reason it is sometimes referred to as the "functional architecture" or even as the "structure of the underlying virtual machine," instead of referring to it as the "hardware" as was sometimes done.

For purposes of cognitive science, the difference between cognitive architecture and other levels of system organization is fundamental. Archi-

tecture marks the boundary between processes that can be explained in biological terms (or other physical-science terms) and those that require appeal to representations or knowledge. A fundamental working hypothesis of Cognitive Science is that there exists an autonomous (or at least partially autonomous) domain of phenomena that can be explained in terms of representations (goals, beliefs, knowledge, perceptions, etc.) and algorithmic processes that operate over these representations. Another way to put this is to say that cognitive systems have a real level of organization at what Newell (1990) has called the knowledge level. Reasoning and rational knowledge-dependent principles apply at this level. Because of this, any differences in behavioral regularities that can be shown to arise from such knowledge-dependent processes do not reveal properties of the architecture, which remain invariant with changes in goals and knowledge. This observation leads to a novel methodological proposal; namely, that the effects of the architecture are cognitively impenetrable.

Another way to view this is to recognize that a system's repertoire of potential behavioral functions—those that remain possible without changing its inherent structure, which we might call its computational or cognitive "capacity"—is constrained by its architecture. In contrast, the different regularities that we may observe in different contexts or environments can then be attributed to differences in the goals, beliefs, strategies, or rules that it adopts.

COGNITIVE SCIENCE AND THE RECONCILIATION OF INTENTIONALITY AND NATURALISM

The idea that cognition is a species of computation helps us come to terms with one very important desideratum of a science of cognition—reconciling representation-governed processes with material causation. Computation offers at least the following hypothesis for how this might be possible. A computer is a physical device that shares with mind the property that some of its regularities can be stated in terms of the semantics of its representations. We can, for example, ask why a certain function continues to produce mathematically correct results or why certain operations over expressions continue to be truth-preserving (if the operations are valid and the expressions denote true states of affairs), and the answer has to make reference to semantical notions. Similarly when certain deviations from rationality or certain systematic semantical errors occur, we can sometimes explain these in terms of the way the semantics is encoded and in terms of the architecture of the system that operates on the encodings. So for example, systematic rounding errors in mathematical functions and the way in which the complexity of the process varies with the nature of the input (e.g., the size of the numerals) would be explained by adverting to

the system of codes—or to the symbol-level regularities. Similarly, further regularities require that we appeal to the physical properties of the system, particularly when the system fails in some way (e.g., the batteries are low or parts of the machine have been damaged).

This trilevel organization is precisely what we hypothesize to hold of the mind/brain. The proof of this assumption is a long-term project, but what success information-processing psychology has made—from studies of performance to studies of reasoning and psycholinguistics—is consistent with this general foundational assumption. And this, in turn, gives us some reason to hope that Cognitive Science may turn out to be a natural causal science in a certain sense which may not be true of other kinds of sciences. Perhaps this needs some clarification.

There are two extremes in styles of science in the social-biological sciences (and of course all grades in between these extremes). One style is exemplified in botany and history (which is sometimes viewed as a science). In this style, we collect humanly interesting facts and place them in taxonomies that reveal some local patterns, thereby enlightening the domain of the science. This kind of pursuit is sometimes called the "natural history" approach. But there is another style of science that attempts to find broad underlying "*causal*" principles. Because in the end, causal principles are materialist, this kind of science is committed to making contact with physics and chemistry eventually. This does not make it a reductionist pursuit since along the way these sciences may evolve a whole new set of principles based on their own vocabularies—providing the world is truly so organized. But in the end, sciences practicing this style are committed to the Unity of Science. Physiology is an example of a science that aspires to be a causal science. And so, some of us believe, is the computational heartland of Cognitive Science, as exemplified in the study of perception for example. That is why Cognitive Science pays attention to physical constraints on transduction on both the input and output side of the organism, rather than redefining the nature of the input in nonphysical terms as Gibson tried to do (and a science based on describing the input in terms of ecological categories such as affordances might well have worked after a fashion, though it would have violated the Causal Unity criterion).

We have no right to assume that all the questions of interest concerning Cognition that arise are questions that will fall under a causal-science explanation. And not all of it can be a computational science (as I tried to argue in my "Computation and Cognition"). This is not such a terrible indictment: There are plenty of very useful noncausal sciences. Linguistics, for example, is a science with a rich deductive structure and beautiful deep principles that constrain a causal science such as psycholinguistics. And there are excellent and useful taxonomic sciences. Most of social, educational, personality, and clinical studies are like that (though because the

word *science* has such strong positive connotations they would deny being akin to natural history). The end-product of studies in most of psychology is a *collection* of generalizations, sometimes connected by a very loose tissue of just-so stories that serve mostly as a mnemonic framework. Psychoanalysis is such a framework, but so is almost everything in sociology and, for that matter, in economics as well (despite the high level of mathematics in the latter discipline—which just shows that formalization does not guarantee a causal science, you can axiomatize almost anything, including psychoanalysis).

There is no a priori guarantee that any particular problem, or class of problems associated with cognition will fall under a causal theory of cognition. For example, much of what goes on in cognition is what we might call common-sense reasoning and, alas, we know almost nothing about this process (by almost nothing I mean that when calibrated against what grandmother knew there has been little progress). And that's why, in my view, long-term memory has not produced a lasting scientific research program in psychology. I would not be surprised to find that molecular biology or neuroscience uncovers some useful mechanisms of LTM, but the reconstructive part of memory—the part involving reasoning that Bartlett studied—awaits an entirely new idea about reasoning. The same applies to learning, personality, and many other parts of traditional psychology: The part that lies within Cognitive Science, contrary to our earlier beliefs, either may be nonexistent or may fall under the general-reasoning problem that we don't know how to begin to analyze.

If cognitive science is to be a science like other causal explanatory sciences (and that's a big "if"—we may one day discover that all we can get out of cognitive science is a descriptive taxonomic discipline like botany) then we will have to be realists about our constructs. If we claim that the mind is computational then we have committed ourselves to a program of research whose goals are to specify what kind of computer the mind is, and what representations and codes it computes over. If that sounds like a tall order it should. We have been trying to make sense of mental activity for at least 3000 years, and this is the first time that we have any idea—however embryonic—how mindlike behavior might be produced by a material object. It may be a small and tenuous step, but it is, as Jerry Fodor would put it, the only game in town or the only straw afloat, depending on how optimistic you happen to be feeling.

THE PAPERS

The papers collected in the anthology initially arise from a historical event: The inauguration of the Rutgers Center for Cognitive Science (RuCCS) in New Brunswick, New Jersey in October 1991. A number of the present authors were in attendance at that event and presented papers. The papers contained herein, however, were prepared much later and many of the authors of these papers were not present at the conference. Rather they are individuals who are in one way or another associates or supporters of the Rutgers Center. Taken as a whole the papers represent a range of cognitive activities that are typical of the discipline. Moreover they are in one way or another concerned with Cognitive Architecture, the original theme of the conference, and provide a variety of views about the current and potential architectures that may shape the future progress in the field.

THEORETICAL ISSUES IN COGNITIVE SCIENCE

Zenon Pylyshyn, Series Editor

Constraining Cognitive Theories: Issues and Options
edited by Zenon Pylyshyn, 1998
From Models to Modules: Studies in Cognitive Science
edited by I. Gopnik and Myrna Gopnik, 1986
Language Learning and Concept Acquisition: Foundational Issues
edited by William Demopoulos and Ausonio Marras, 1986
Meaning and Cognitive Structure: Issues in the Computational Theory of Mind
edited by Zenon Pylyshyn and William Demopoulos, 1986
The Robot's Dilemma: The Frame Problem in Artificial Intelligence
edited by Zenon Pylyshyn, 1989
The Robot's Dilemma Revisited
edited by Kenneth Ford and Zenon Pylyshyn, 1996

Constraining Cognitive Theories
Issues and Options

edited by
Zenon Pylyshyn
Rutgers University
New Brunswick, NJ

1998

Ablex Publishing Corporation
Stamford, Connecticut