Draft of a talk presented at the conference on
"The Chomskian Turn", Tel Aviv and Jerusalem,
April 11-14, 1988.

Rules and Representations: Chomsky and Representational Realism

Zenon Pylyshyn
Centre for Cognitive Science
University of Western Ontario,
London, Ontario, Canada

Introduction

Speaking as someone who has personally felt the influence of the "Chomskian Turn", I believe that one of Chomsky's most significant contributions to Psychology, or as it is now called, *Cognitive Science* was to bring back scientific realism. This may strike you as a very odd claim, for one does not usually think of science as needing to be talked into scientific realism. Science is, after all, the study of reality by the most precise instruments of measurement and analysis that humans have developed.

Yet in the human sciences, realism about theoretical (and especially mental) entities had fallen out of fashion in the middle third of this century. Positivism, and in particular the doctrine of Operationalism which inflicted on psychology Bridgeman's misunderstanding of what goes on in Physics, was one reason. There were other reasons as well. Some of them may simply have been matters of fashion or of the sociology of the field: when physics was demonstrating its power over the physical world, psychology felt an urgent need to demonstrate its own scientific prowess by making predictions and by controlling behavior. Without many of the intellectual tools needed to formulate precise mentalistic theories, and without some way to understand, at least in principle, how behavior could ensue from mental structures, Cognitivism — which is, after all, very close to folk psychology — may have been left feeling old fashioned and scientifically impotent. Whatever the reason, well established common sense notions, such as knowledge, beliefs and desires, were banished from psychology as merely an imprecise way of speaking.

Some Background History

It seems to me that there were two things that made the difference in bringing mentalism, or perhaps I should say cognitivism, back into cognitive science. One was the work that began with Hilbert and was developed by Turing and Church and Markov and others who formulated the abstract notions of mechanism and of what we now call "information processing." This is the lineage that led to Cybernetics and later to Artificial Intelligence, though a very large proportion of the field would now probably dissociate itself with that "logicist" part of the family tree, just as earlier Logicists like Frege dissociated themselves with psychological pursuits.

The other development that brought mentalism back was the discovery that it was possible to treat some aspects of the human capacity for language in a way that made it at least appear to be compatible with mechanism. These developments encouraged many people to hope that one day we might have an explanatory theory of some of the mechanisms of linguistic competence, not just a taxonomic description of a corpus of linguistic utterances. This was, of course, the beginning of what people here have dubbed the "Chomskian Turn". The specific results achieved in transformational grammar, coupled with the generative or procedural aspect of the theoretical mechanisms (which, after all, wore the formal garb of Post Production systems and of Markov Algorithms) gave us hope that we were on the track of a theory of language understanding and language production.

Well we were wrong about a lot of things, and especially about how a theory of grammar might be incorporated into a theory of comprehension/production (recall, for example, the decisive failure of the "derivational theory of complexity"). Many of the early ideas of psycholinguistics were later abandoned. What remained, however, was the basic belief that *rules*, which included "rules of grammar", would play a central role in the theory of not only comprehension, but also of cognition more generally. Moreover, ever since those developments in the late 50's and early 60's, talk about rules no longer meant we were describing a corpus of behavior; rather when we spoke of rules we were referring to an internal property of some system or mind. We sometimes even spoke of the rules as being "internally represented".

However, what was meant by the phrase "internally represented" was far from clear — even to those of us who spoke that way. And it did not get any clearer if one adopted Chomsky's way of putting it when, for example, he said that a theory of the speaker/hearer "involves rules", or that the theory postulates a certain rule R "as a constituent element of [the speaker's] initial state" or "attributes to …[the speaker] a mental structure … that includes the rule R and explains his behavior in terms of this attribution" (Chomsky, 1986a, p243); or when he says that a speaker is "equipped with a grammar" or "internalizes a system of rules". Yet, despite the uncertainties, none of us doubted that what was at stake in all such claims was nothing less than an empirical hypothesis about *how things really were inside the head of a human cognizer*. We knew that we were not speaking metaphorically nor were we in some abstract way describing the form of the data.

With the passage of time, and with an understanding of the computing as a concrete instance of a physical system that truly "represents", we refined and concretized our idea about what it might mean for something to be a "representation" and also what it might mean for something to "internalize a rule" —a refinement that for most Cognitive Scientists raised none of the problems that had perplexed Wittgenstein. For some of us this was the beginning of a new realism about the ontology of such mental constructs as rules and representations, as well as of such homely notions as knowledge and belief and goals. That's not to say that we all understood these notions in exactly the same way. Indeed I shall later discuss several ways in which some of us differ on these matters. But we all took it for granted that such mentalistic notions had precisely the same status in Cognitive Science as atoms and molecules have in physics. In both cases they constitute empirical hypotheses about the natural world, although of course the empirical status of the cognitive constructs is far more tentative at this time.

Chomsky was insistent on this realism from the start. He took the position that if one's best theory of some cognitive phenomena involved postulating, say, a representation X, then we have no more reason to doubt the existence of X on any a priori grounds, such as that we find it impossible to visualize how X could occur in the brain, than a physicist with comparably supportive evidence would have to doubt some particular physical hypothesis. Nor could we, in either domain, take the possibility that the phenomena *might* be compatible with some other, yet unformulated, hypothesis as an argument for some sort of inherent indeterminacy. Yet both of these avenues of resistance against realism have appeared —and are still counselled —in some quarters of Cognitive Science.

What I plan to do in this essay is to elaborate the idea of rules and representations along the lines of some of my own work, where I distinguish between "explicit representations" and what might be characterized as "implicit representations" (though in my own work I confine the term "representation" to the former case only). In the case of rules, this leads to distinguishing between rules that are "explicitly represented" or "explicitly encoded" and ones that merely describe constraints and regularities to which the system conforms, without doing so in virtue of explicitly encoded rules in the system. Since this is not a distinction that Chomsky has endorsed, I will attempt to defend this, perhaps even more radical realism, against some of his recently published views.

I will introduce these questions by first discussing the issue of "indeterminism" which both Chomsky and I have attacked in somewhat different but related ways. This will lead to a discussion of a notion of "strong equivalence" among cognitive models, that will in turn serve to introduce several additional points regarding the interpretation of the notion of an explicitly encoded representation and of the conditions under which such representations need to be postulated.

Indeterminism and Strong Equivalence

Chomsky devotes considerable attention in several of his papers (e.g. Chomsky, 1984; 1986) to various indeterminism theses — from those which say that it is impossible to decide among extensionally equivalent grammars (e.g. Quine, Lewis, Dummett, and others), to those (which Chomsky calls "Wittgensteinian") that maintain that it is never possible to decide which rules a system is following — or, indeed, whether it is following rules at all. Chomsky quite rightly admonishes these critics for their lack of imagination in considering how one might go about empirically settling such questions. In the course of these debates the distinction that remains central is between the question of what an empirical claim *is*, or what its truth conditions are, and how one might in practice go about testing it. Even if at any point in time you have no idea how some claim could be tested (e.g. the appropriate experimental methods or tools have not yet been invented), this does not make the claim either vacuous or indeterminate. So long as there is something that is being claimed — so long as the claim itself is well defined and has truth conditions for its being true, it remains a perfectly sound scientific proposition.

I have also devoted considerable attention to pointing out the flaws in arguments put forward for various indeterminism theses by psychologists like John Anderson (Anderson, 1978; Pylyshyn, 1979). Now one might well wonder why anyone bothers to argue about such things: indeterminism claims are inherently boring (with the possible exception of ones that can be embedded within a theory as well developed as quantum mechanics — and even there they are only of marginal interest, in the view of many). After all who pays any attention to the crank who says that you can't tell whether the earth is round or flat?

The reason one bothers, I think, is that the difference between a view of cognition that leads naturally to an indeterminism thesis and one that does not goes quite deep, and exposing it reveals something important about alternative ways of understanding the phenomenon of mind. Apart from issues of parsimony and generality, we do not argue about which of two equally predictive formulations of classical mechanics provides the correct explanation of planetary motion (say one which expresses the invariance principles in the form of the Hamiltonian, or directly in the form of Newtonian axioms), nor do we conclude that the issue in indeterminate. The question simply does not arise because in mechanics no ontological claims are made about the notation in which the equations are cast. To put it another way, in mechanics the empirical consequence of the equations lie in their extension, not their form or their intension.[1]

The same would hold of cognitive science if that field were about predicting overt behavior, as the behaviorists insisted.[2] Of course, there is a sense in which all we have is behavior — i.e. empirical observations — but nothing follows from this fact alone since we have to interpret the behavior we observe before it becomes relevant to the task of deciding among theories. For example, if we interpret a certain "response" as a judgment (of, say, the grammaticality of a

---

1. Chomsky, 1986) has quite rightly played down the notion of extensional equivalence of theories by pointing out that what we really have in mind when we speak of "extensional equivalance" is not that two theories make all the same predictions, but that they coincide on some subset of the evidence. While this is both true and important to keep in mind, there is nonetheless a useful sense in which we can speak of two theories being extensionally equivalent; namely in cases where they are theories of a mechanism or a process. In that case it is useful to distinguish the input-output behavior of the process from the evidence that points to the detailed steps by which the process generates this input-output behavior. In cognitive science practice, the distinction between the two types of evidence usually is quite clear (although some people occasionally lose sight of it; e.g. Anderson, 1978). When two theories specify different mechanisms which produce the same input-output behavior I refer to them as "extensionally equivalent" or "weakly equivalent".

2. Chomsky was fond of pointing out to the behaviorists that the parallel in physics would have been to insist that physics was the science of meter readings!

sentence, or of whether a sentence correctly describes the contents of a picture), this piece of behavior no longer qualifies as *mere* behavior since it has a truth value — it is taken to mean something. Similarly, if we interpret reaction time as an index of computational complexity it too ceases to be mere behavior, to be accounted for in the same way as any other recorded response.

No psychologist would accept two theories of some process as equivalent just because they generated the same set of behavioral records,[3] say a list of pairs, the first of which was the predicted response of a subject and the second of which was a number representing the time at which the response occured. Suppose, for example, that two theories described behaviorally equivalent mechanisms, in that both correctly predicted the temporal pattern of a series of behavioral (i.e., input-output) events. If in one theory the time-of-occurence was merely calculated in some way (say, using an equation) while in the other theory the time corresponded to the number of operations performed, or some other natural function of the internal processing, it is clear that the second theory would be preferred. There is nothing unusual about this case, one would get pretty general agreement about which of two behaviorally equivalent process theories is to be preferred. Psychologists are fairly reasonable people when they refrain from offering philosophical opinions about the nature of their work.

Some notion of *strong equivalence* is implicit in cognitive science practice. In cognitive science there is a tacitly accepted notion (perhaps not very well understood at present) of *how* some behavior is arrived at; not just how it is neurophysiologically realized, but by what cognitive mechanisms it is carried out, what cognitive states it goes through, and what rules determine the sequence of cognitive states it undergoes. Elsewhere (Pylyshyn, 1984) I have tried to tease out the intuitions that the best practitioners in the field appear to share tacitly, and to attempt to sharpen and justify them. This analysis leads to a sharp distinction between what I call the functional architecture of the cognitive system, and the rules and representations it uses. This distinction will be important in later considerations of the notion of strong equivalence and will, I hope, serve to sharpen the question of what it is to postulate a strong sense of "rule governed" or "representation governed" process.

A Strong and Weak Sense of "Following a Rule"

This brings us back to the central idea I want to discuss, namely the idea of behavioral regularities being based on rules and representations. When people appeal to the existence of rules in order to explain generalizations in actual and potential behavior, they may have one of a number of ideas in mind. In particular, they might justifiably claim that their theory "involves rules" or that a person has "internalized a system of rules" while making quite different assumptions about which aspects of the rule-system are empirically significant, or about *how* the system of rules enters into the causation of behavior.

---

3. Psychologists, like most practicing scientists, are quite sensible people when they are engaged in doing science. This doess not appear to be true, however, when they depart from this work in order to offer philosophical opinions about the science. It is a strange fact about the field of psychology, that whenever psychologists do meta-science they appear inevitably to revert to behaviorism. This is true of the recent discussions about the significance of Connectionism (see the discussion of this in Fodor & Pylyshyn, 1988).

There are many ways in which scientist might differ in the assumptions which they associate with a theory that contains a system of rules. These differ chiefly in terms of the intended grain of comparison between the theory, expressed in a certain standard form, and the empirical facts. For example, in the weakest case, it may be that the only empirical import of the set of hypothesized rules is their extension: *as a group* they account for regularities in behavior, for the structural relationships among behavioral elements, for judgments, and so on. The system of rules might meet what Chomsky has called "descriptive adequacy". In the next more detailed case, the set of rules may decompose into subsets, with empirically relevant relations among the subsets. In this case, for example, there may be tie-breaking principles, ordering relations, and other conditions that hold over the subset-types, thus reifying the types themselves. This is clearly the case for the distinction among, for example, phonological and morphological rules. In that case morphological rules provide the elements and structures over which phonological rules apply.

Going down to an even finer level of comparison, the theory may reify individual rules by claiming that each rule corresponds to some individual internal object or property. In this case, for example, the fact that there are 30, as opposed to 40 rules would be empirically significant, so that if the set of 40 rules were to be reduced to an extensionally equivalent set of 30 rules the resulting system would be empirically different —they would be correspond to different physical systems. Evidence for this level of ontology may, for example, come from observing that individual rules may be systematically added or deleted. That's the sort of evidence one might cite in the case of rules of etiquette or traffic rules. The reason for individuating rules may also rest on more subtle considerations, such as the fact that by individuating rules in a certain way we can show the operation of independent principles. (I shall later suggest that there is some reason to believe that this is the case for linguistic rules).

A further refinement would occur if we claimed that the individual terms and symbols in the canonical expression of each rule, as well as the structural relations among the symbols, correspond to distinct properties of the system, so that the *structure* of the expressions was empirically significant. This latter grain of comparison is precisely what I have in mind when I speak of rules, or other constructs, as being "explicitly" represented or encoded. In this case certain aspects of the actual notation used to describe what is represented in the mind or is "internalized" are assumed to map onto the empirical world.

Clearly it matters which of these senses of "involving rules" we intend. The motion of planets can also be described in terms of rules, but we don't want to ascribe any ontological status to the individual rules as expressed in some particular formulation of the theory, much less to the form that these rules take in some canonical notation. We don't intend the strongest of the above types of correspondence (the one I call "explicit encoding of rules"). That's why we would not say that planets behave the way they do because they *access* and *use* a representation of the rules. This is not a subtle point about how we should talk about our theories. It is a fundamental point about how we claim that some particular rule-governed system functions.

Perhaps the case of planets does not provide the best example because the "rules" in that case do not apply to representations, but to physical properties. A better example is provided by recent work on the computational processes in vision, where we see a clear distinction between

implicit and explicit representations. Consider first the case of what is referred to as "early vision" (Marr, 1982). The visual system is able to solve the following problem: given a 2-D image of a 3-D scene, recover the 3-D layout that gave rise to that image. This "inverse mapping" problem is underconstrained; there are an indefinite number of physically possible 3-D configurations that could have lead to any given 2-D image. Yet the visual system usually produces a unique interpretation, and moreover it is generally veridical. The resolution of this puzzle consists in recognizing that the visual system behaves as though it were making certain assumptions about the nature of the physical world — assumptions which are often, though not always, satisfied in our sort of world. The "assumptions", called "natural constraints" include such things as that most of an image consists of light reflected from surfaces, that the distance of the surfaces from the perceiver varies gradually in most of the image, that certain kinds of discontinuous visual features in the image usually arise from physical discontinuities on the surface of smooth rigid 3-D objects, that contour discontinuities usually arise from occluding edges, that the light illuminating the object usually comes from above, and so on. The visual system acts in a manner that is consistent with its having made assumptions such as those sketched above about the world it is perceiving. One might wish to say that it has "internalized" knowledge of certain constraints that hold in the physical world. However, nobody actually believes that the visual system "uses" an explicit representation of these constraints. The constraints are *implicit* in the structure of the perceptual system in the sense that only interpretations compatible with the constraints are attempted.

Matters are quite different when it comes to more interpretive stages of perception. Here the process is more like that carried out by Sherlock Holmes in his examination of clues at the scene of a crime. Each piece of information is weighed in relation to Holmes' beliefs about the crime and his understanding of what goes on in the criminal mind. Holmes explicitly "uses" his beliefs and assumptions in interpreting the scene, in a way that the visual system does not "use" its assumptions about natural constraints in the early stages of vision. The two cases can be separated empirically; for example, in Holmes' case the process is what I call "cognitively penetrable". What Holmes will think next can be rationally and predictably altered with changes in his ancillary beliefs, whereas the "assumptions" in the early vision case act in a fixed manner, enter into no processes other than early vision, and are imune from changes that are attributable to differences in ancillary beliefs about the scene being perceived. The important point here is that the two distinct cases are quite different and a theory that addresses the process must specify which one is being claimed when terms like "internallized" are invoked.

Whether it is true or not, a theory which claims that representations are "explicitly encoded" is making a much stronger claim than one which simply claims that the system has "internalized" the rules or representations. Similarly, a theory may make stronger or weaker claims about the *way* in which these representations generate behavior. Consider, for example, a theory that claims that under specified conditions the physical properties that instantiate the structures of symbolic codes cause the behavior in question (the conditions might, for example, specify that the physical codes must be in a certain relation to the system, e.g. located in a certain register or being in contact with a "read head", as in the Turing machine). Such a theory makes a much stronger claim than one which merely says that the behavior "involves" rules or representations. The difference between the weak and the strong versions of these theoretical claims is extremely

important in cognitive science. Indeed, there is good reason to think that the distinction between the weak and the strong claims marks one of the unique ways in which cognizing systems are organized, in contrast with other complex systems in nature.

Note that there is no issue here with respect to the correctness of the rule-system as a description. It is not even a question of whether or not one is a realist about one's theoretical constructs. It is a question of the exact way in which the theory maps onto the world —of which aspects of the theoretical system are claimed to have ontological status. It is, to put it in terms more familiar to some people, a question of the truth conditions of the claim that a system "contains rules".

The claim that a system of rules is explicitly encoded in a certain form carries with it certain truth conditions. The rules are explicitly encoded if and only of there exists some mapping from the rules as inscribed in some canonical notation and the physical states of the system. We need not be commited to any particular form of this mapping, except insofar as we wish to claim that certain formal properties of the expression of the rules are empirically significant. If, for example, we claim that the structure of the rule (its individual parts and their relations to one another, for example their tree-structure) is empirically significant, then this structure must itself be preserved by the mapping.

The above ideas can be made mathematically precise. For example, the claim that the *form* of a symbolically encoded rule, or any other symbolic expression, is physically instantiated entails the existence of a mapping from the symbol structure to some physical properties of the machine (/brain), or to physical-state types, which is *structure preserving*. The formulation of such a mapping assumes the existence of a function from tokens of atomic symbols to distinct physical properties of the system. Then the mapping for complex elements which have a particular structural form is defined recursively in terms of this atomic-element mapping, together with the structure of the complex element. Thus the definition of the 'physical instantiation mapping' $\boldsymbol{F}$ for complex expressions would be given recursively in terms of $\boldsymbol{F}$, assumed to be defined for *atomic* symbols, together with the *form* of the complex expression. The relevant mapping in this case might state that for any expressions P and Q,

$$\boldsymbol{F}[\mathrm{P\&Q}] = \{\boldsymbol{F}[\mathrm{P}], \boldsymbol{F}[\mathrm{Q}]\} \qquad (1)$$

where is some physical relation that holds between physical properties $\boldsymbol{F}$[P] and $\boldsymbol{F}$[Q] and which thereby instantiates the '&' relation among the symbolic expressions. In this schema, P and Q are replaced by whatever symbol structure occur in the specified position of the original expression. For example, in establishing the mapping from an expressions such as "(A&B)&C", P and Q correspond to 'A&B' and 'C' respectively, so that mapping rule (1) would have to be applied a second time to pick the relevant physical structures.

Of course, not *all* aspects of the way the expressions are written are supposed to map onto the world. for example, the particular fonts used or even the left-right order of the expressions may not matter. On the other hand, they *might*: the theory that accompanies the notation must tell us which aspects are intended to be empirically significant. This interdependence between theory and notation does not diminish the importance of the distinction between systems in which nothing about the notation is significant and ones in which the structure of certain

expressions matters because the expressions are assumed to be instantiated in the system in a manner that preserves their structure.

Notice also that this particular level of correspondence, which I have referred to as "strong equivalence", does not just mean a more precise or a more detailed theory. It is a type of correspondence that has no precise parallel in physical theories[4], because it claims that *tokens* of symbols in an expression (in this case the expression of a particular rule) map onto some distinct properties of the system in such a way as to preserve (a) the *symbol types* and (b) the *structure* of the expressions. Formula (1) above illustrates how these conditions might be met in a particular case. Meeting these conditions is functionally equivalent to "writing down" the symbols in some physical form in the system (or the brain)[5].

Explanations that appeal to such "stored symbolic expressions" are not just theories that posit a more detailed microstructure. They are theories that posit a particular kind of microstructure; a *representational* microstructure. The terms of the theory not only designate properties of the system; they designate properties of the system that have representational content —— that have semantic properties.[6] The complex expression has a constituent structure that reflects the semantic structure of what it represents.

Although Chomsky has been one of the most vigorous exponents of a strong equivalence view, has spoken of linguistic processes as "computations", and has even emphasized the importance of the *form* in which rules are expressed, his writings have not acknowledged the distinction between the rules in a system being *explicitly encoded*, and a system merely implicitly *conforming to* rules —— i.e. behaving as if it were following rules even though the behavior may arise from unspecified causes. Of course Chomsky is correct when he says that all we can hope to do is find the theory that best accounts for all the evidence, and that if such a theory postulates rules, then we assume that the system does indeed "contain rules". That's not the issue: everyone agrees about that. The question is; exactly what do we take the claim of "containing rules" or "having internalized rules" to be: what the truth conditions of such a claim are. What does the claim say about the structure of the system and what does it commit us to (other than the obvious fact that by positing rules we can account for a certain range of phenomena). Some senses of "containing rules" entail consequences that other senses (equally compatible with the informal use of the phrase) do not.

---

4. Whether or not this makes representation-governed systems fundamentally different from other complex natural systems depends on what one takes to be fundamental. Certainly a science that deals with these systems is unlike physics, inasmuch as it is a science of a special part of the universe (it is what Fodor, 1976, calls a "Special Science"). Like the systems studied by other "Special Sciences" (such as biology or economics), representation-governed systems are natural systems that function according to the basic laws of nature (i.e. the laws of physics). But equally clearly, they involve other levels of organization; not just as approximations, but as genuine levels over which explanatory generalizations can be expressed.

5. There is nothing mysterious about this notion: It's exactly what can happen in a computer when the computer is correctly described as following a rule (in the strong sense). In the case where the rule is being executed "interpretively" —when the formulation of the rule constitutes some "executable code" —there really is an explicit physical encoding of the rule that meets the conditions I have been discussing, and therefore that functionally corresponds to the rule being "written down" in the system and "read" in the course of processing.

6. To claim that such symbol structures have semantic properties is not to claim that they must represent the content of thoughts. Some of the symbols may represent things that we would not want to count as actual *thought contents*. These include various kinds of "features", aspects of the control structure or markers which keep track of where the process has reached, and so on. They may also include aspects of grammatical rules. We may not want to call these "thought contents" simply because they only have a role within some narrow and highly encapsulated system, because they do not enter into general reasoning, because the semantics of the symbols in question does not lie in the domain of the thoughts that are taking place at the time, and so on. They may be what Dennett calls "subpersonal" contents. Nonetheless, these symbols do have a semantic content: they are not just the names of physical states that encode them (e.g NP refers to the class of noun phrases, VP refers to verb phrases, and so on).

I don't see how Chomsky can be noncommittal on this question, or how he can simply equate it with the question of whether the "best theory" posits rules. Before one can determine whether the best theory is justified in claiming that a system has internalized a rule we need to understand what such a claim means. As we have seen, there is sense in which it can mean that the system conforms to the set of rules, taken as a whole. But there is also a clear sense in which it can mean that the rule is explicitly encoded. This is the strongest sense of having "internalized the rule" for it claims the strongest degree of correspondence between the rules as formulated in the theory and the structure of the system. One can't be agnostic about this issue while being a realist about strong equivalence. As in the various debates in which Chomsky has so vigorously opposed the indeterminacy thesis, one must distinguish between *what* a theory claims (what the truth conditions of the theory are), and what the evidence for the theory is. It is common to have two theories that coincide on all available evidence, yet have different truth conditions — i.e., make *different claims*. [7]

Methodological Considerations

Having made a distinction between several ways that a system can have "internalized" a representation (including one strong sense that is of particular interest), the question immediately arises: How do we know whether some particular sense of internalization (say the "explicit encoding" sense) is warranted in a particular case? Putting aside the purely philosophical concerns raised in footnote 7, it may be of interest to inquire whether there are interesting cases in which it is reasonable to conclude that explicit representations are involved. As usual, we can't specify in advance precisely what evidence will be critical for such cases. However, the way that evidence bears on specific claims depends a great deal on our understanding of the claim. One way to try to get some insight into what the strong claim implies is to examine some kinds of evidence that has in the past led us to postulate "explicitly encoded" representations or rules, at least in certain clear cases.

In what follows I will consider several kinds of evidence that have been used in Cognitive Science to try to sort out the distinction raised above. In this discussion I do not distinguish between "rules" and other forms of representations[8]. In fact, the question will be whether certain states of the system must take the form of explicit symbol structures.

The first type of evidence I will consider is concerned with so-called "higher cognitive processes". The type of evidence that can be cited in this case is quite different from that to which we typically appeal in the case of language. In the case of higher cognitive processes there are strong general reasons for holding that requirements of expressive power, as well as the

---

7. One might ask, What if two theories coincide on <u>all possible evidence</u>? Could they still be be distinct theories? The answer is far from clear. In the debates over indeterminism, Chomsky has quite rightly denounced the very notion of "all possible evidence" by pointing out that this is not a well-defined class. There is no possible observation which can in principle be excluded in advance as irrelevant to some particular scientific hypothesis. That's what makes the problem of induction such a deep problem.

8. Indeed, in computer science the distinction between a rule and any other expression lies solely in what consequences follow from accessing it on specific occasions. In contemporary programming languages, such as *Prolog*, the distinction between a rule and an assertion does not even exist. (Although one could perhaps think of "rules" as those assertions that have variables which may get bound differently on different occasions. But this would be a rather unusual way to view the distinction between a rule and an assertion).

productivity and systematicity of representational capacities involved in reasoning demand symbol systems, and in particular that they require explicit structured representations. In addition, the plasticity and rationality of mental processes provides evidence for the ubiquity of explicit representation in reasoning generally. I will discuss examples drawn from the study of the "cognitive penetrability" of such processes as those involved in the use of mental imagery. Following this I will briefly raise the question of the status of linguistic representations, such as grammatical rules and the various representations of sentence structure that these rules define.

Cognitive Capacity and Cognitive Penetrability

In cognitive science, as in folk explanations, there appear to be two distinct kinds of explanatory principles; one that appeals to intrinsic functional properties or mechanisms, and one that appeals to the content of representations, such as knowledge, beliefs, goals and the like. This is pretty generally accepted in practice, if not in philosophical discussions. Thus nobody would think of trying to provide an explanation of why I am here and what I am doing at this moment which did not appeal to such facts as that I was invited to come here, that I was attempting to persuade you of certain propositions, and so on. This is not just a matter of convenience: the underlying empirical hypothesis is that there are certain regularities that can only be captured by the use of such a vocabulary. By contrast, nobody these days would think of giving an explanation for, say, the laws of perceived colour mixing by appeal to such things as beliefs, goals, intentions, utilities, and so on.

What is the difference between the two cases just mentioned (explaining why I am here and explaining the laws of colour mixing)? Let me just suggest one difference: whether perceived red and perceived yellow mix to form perceived orange is independent of what I believe I am looking at or what I believe about the psychophysics of colour mixing. It is a cognitively robust psychophysical regularity. By contrast, whether I will appear at a conference at some exotic location and carry on a discourse on representation *is* dependent on all sorts of other collateral beliefs, even though it may still be quite a robust regularity (judging by my recent travel itinerary). It is, in fact, a regularity that can be readily disrupted in a way that is both systematic and rational. For example, I would not have made the trip here if someone had called me up and persuaded me that the conference had been canceled, whether or not it actually was. In other words, the invitation-accepting regularity is cognitively penetrable, whereas the colour-mixing regularity is not. That's not the only criterion distinguishing the two cases, but it is one of the most important from my perspective, and I will return to it presently.

There is another way to look at this distinction that may be more revealing; namely, in terms of the distinction between cognitive capacity and representation-governed regularities. The reason that psychology cannot be viewed as concerned with predicting behavior or "accounting for variance" is that it matters a great deal *why* some particular regularity is manifested — whether it is because it is the only one possible, given the circumstances, or whether it is for some much more ephemeral reason — such as that the subject wishes to oblige, or understands it to be his task, or believes that it will serve his best interests, or just doesn't care enough to do

anything more than free associate or guess. Whenever we are concerned with *explaining* some regularity, as opposed to merely describing it, it is essential that we view the regularity against a background of what *might* have occurred given different circumstances[9].

I would like to dwell on this point a bit since it connects my ideas on strong equivalence with the requirements for explanatory adequacy that Chomsky has so forcefully articulated over the past 25 years. I said that explaining some regularity requires that we be concerned not only with the occurrence of instances of the regularity, but also with the range of circumstances under which the regularity will remain fixed and the range under which it will vary, and in particular, with the *way* the regularity might be modulated by differences in circumstances. This is crucial: much depends on how the counterfactuals turn out. If in circumstances that differ only in terms of what a person believes or in terms of what the person's utilities are, we find quite different regularities, where the difference bears some logical relation to the difference in beliefs or goals, then we know that the regularities in question are not attributable to the person's cognitive capacity. Cognitive capacity may change, but not in ways that can be explained as rational responses to what the person believes: in other words, cognitive capacities are not cognitively penetrable[10].

Strong equivalence of processes, as I have interpreted it, is closely tied to this distinction. In order for two processes to be strongly equivalent they must not only exhibit the same behavior (that would be extensional or weak equivalence), but they must generate the behaviors by the same cognitive mechanisms. In other words, they must have the same capacities.

The analogy here with computer algorithms is very close. In order for two computer systems to be strongly equivalent they must not only exhibit the same input-output behavior, but they must do so by means of the same algorithm and data structure —— which is to say they must also have the same functional architecture, since identity of algorithm implies an identical set of basic operations[11].

The notion of capacity I have tried to sketch above is closely tied to the distinction between the two senses of representation for which I have been arguing. If some particular cognitive regularity is part of the capacity of the system —— if the system could behave in no other way over a certain range of counterfactual circumstances —— then it is at least possible, barring other sorts of evidence to the contrary (some of which are sketched below), that we might get away without

---

9. The importance of considering a set of behaviors against a background of possibilities is important for other reasons as well. For example, it is only when a particular piece of behavior —e.g. the particular behavior I am emitting at this very moment —is viewed as a member of a certain equivalence class, that it comes possible to explain it. Thus, if my present behavior is viewed as a member of a class of bodily movements, it calls for a different explanation than if it is viewed as a member of a class of utterances. Chomsky (1986) has made a similar point against Wittgenstein's argument that since it is meaningless to ascribe rules to a person considered in isolation, then rule-following must be a conventional description used to predict the behaviors of members of a certain social community. Chomsky correctly pointed out that if a person is viewed as an individual whose behavior completely unique, no scientific claims at all can be made. Theories, quite generally, apply to behaviors taken under a certain description, which means that they are viewed as non-unique both with respect to occasions and with respect to individuals.

10. Note, by the way, that this is not a deep point. It simply affirms that if you want to explain some phenomenon by appealing to the way certain mechanism are used to process information, you can't then turn around and claim that the phenomenon is caused by the way the mechanisms themselves change. If that were true it would be a different explanation from the one you claimed you were giving. Capacities are supposed to be just those mechanisms whose behavior need not be explained in terms of rules and representations.

11. Of course it's always possible for one of the systems to explicitly emulate the functional architecture of the other and then execute the algorithm using the primitive operations provided by the emulated architecture. In that case, however, strong equivalence would only hold for the emulated system, not the original one. For more on the technical notion of strong equivalence, see Pylyshyn (1984).

positing that the rules in question are actually represented. A perfectly acceptable interpretation *might be* (though it needn't be) that the system only behaves *in accordance with* rules or behaves *as if* it had certain beliefs — i.e. in a way that is consistently described in terms of the rules or beliefs. On the other hand, if the regularity was only one of many that were compatible with the system's capacity, *and* if it followed the other ones when it was rational for it to do so given the information at hand, then we need some account of how mere differences in beliefs or utilities could make that difference.

The problem to be explained is how systematic changes can occur that are attributable to differences in beliefs and utilities. More generally, it is to show how reasoning, leading to semantically characterizable plasticity, can occur in a physical system. Fodor and I have discussed this problem at length in connection with our analysis of the inadequacy of Connectionist models in accounting for reasoning (Fodor & Pylyshyn, 1988). The basic argument is not unlike that given by Chomsky many years ago for the need for a generative grammar. The argument appeals to the fact that the capacity for reasoning, and for representing beliefs and other propositional attitudes in general, is both productive and systematic: in an intelligent system, the capacity to represent certain states of affairs almost never occurs in isolation from the capacity to represent systematically related and novel states of affairs, and the capacity to draw certain kinds of inferences always occurs together with the capacity to draw other kinds of inferences. This pattern of capacitites is natural and involves no additional assumptions in systems that encode beliefs using a combinatorial system of codes, much as occurs in natural language. In systems that do not encode beliefs in this way, this kind of systematicity need not hold, so the pattern of capacities remains a mystery.

Thus in any system that represents beliefs by encoding them (or "writing them") in a system of symbolic codes with a combinatorial syntax and semantics (i.e in a "language of thought"), it must be the case that if the system is capable of representing the situation P&Q it will also be capable of representing the situation P and the situation Q. Just as with natural language (which presumably encodes thoughts) if a member of a linguistic community is able to assert, say, that it is warm and sunny, he will in general also be able to assert that it is warm and he will be able to assert that it is sunny. The exceptions would be noteworthy. These include novel phrases memorized from a phrase book or idiomatic expressions that do not derive their meaning in the usual way from the meaning of their constituents.

Similarly, reasoning typically involves the application of rule schemas. Because of this we do not in general find that people are able to infer P from P&Q&R, but are unable to infer P from P&Q (e.g. are able to infer that John went to the store from knowing that John and Mary and Fred went to the store together, but are unable to infer that John went to the store from knowing that John and Mary went to the store together). A natural explanation for this regularity in the inferencing capacity is that classes of inference, such as those in the example, involve the application of a common rule schema. However, such a rule schema can only work if there are articulated symbolic expressions to which it can apply — if the beliefs in question are explicitly encoded as symbol structures. Furthermore, the rule itself — whether or not it is itself explicitly encoded in its entirety — must at least provide variables that can be bound to constituent parts of particular belief encodings. This sort of systematicity has been used by Fodor & Pylyshyn

(1988) to argue that in general beliefs must be encoded by systems of symbols which have a constituent structure that mirrors the constituent structure of the situation being represented.

There is another way of viewing the need for drawing a distinction between regularities that arise from explicitly encoded representations and those that arise from the intrinsic capacity of a system (or properties of its functional architecture). As every psychologist knows, when you are interested in explanatory power the first thing you have to do is minimize the number of free empirical parameters at your disposal. In higher cognitive processes, rules and representations function rather like free empirical parameters, inasmuch as they can vary from situation to situation with few independent constraints. By contrast, cognitive capacities remain more-or-less fixed, except for certain specified variations directly attributable to biological causes. Variations in capacity are conditioned by laws of growth, neural arborization, laws of chemistry and endocrinology, and the like. On the other hand, variations in rule-governed behavior, at least in the case of central cognitive processes, are approximately as broad as the set of computable functions. Therefore one must endeavor to attribute as much as possible to the *capacity* of the system or, in my terms, to properties of the *functional architecture*. Put another way, one must find the least powerful functional architecture compatible with the range of variation observed (or observable under some relevant condition). This is exactly the goal Chomsky declared many years ago for linguistics: find the least powerful formalism for expressing the range of human languages and you will have a formalism that you can view as intensionally (as opposed to merely extensionally) significant.

Some Proposed Capacities

Let me turn now to the question of whether certain particular cognitive phenomena can be ascribed to the intrinsic capacity of the cognitive system (i.e. its functional architecture) or whether they should be viewed as governed by representations.

Very little is known about the capacity of the central cognitive system. There is every reason to believe that it imposes strong constraints, as Chomsky has always claimed. Surely not every logically possible thought can be thought by humans. In addition, it is even more apparent that not all cognitive processes are equally complex — by whatever measure of complexity one might wish to use. Differences in complexity very likely reflect properties of the functional architecture. Indeed, measuring processing complexity using such techniques as reaction times has been used with considerable success in validating, in the sense of strong equivalence, computational models of small scale cognitive processes. Thus they provide one methodological route into the nature of cognitive capacity. Nonetheless, virtually every proposal for functions attributable to the cognitive capacity of the central cognitive system, or to its functional architecture, fails to stand up under scrutiny. Thus, for example, the laws of classical and operant conditioning appear to be cognitively penetrable (see the review by Brewer, 1974) — i.e., they can be altered in a rational way be providing the subject with information (e.g. showing him how the apparatus is wired up, or explaining the reinforcement contingencies). Similarly various proposals for memory storage and retrieval mechanisms, such as holograms or quality spaces, are also inadequate because the processes being modeled are demonstrably penetrable by

beliefs (e.g. the confusability profile of a set of stimuli is sensitive to what the stimuli are perceived as being, which in turn depends upon goals and beliefs). Gibson's direct perception thesis also falters on (inter alia) the facts of cognitive penetrability of much of the later stages of perception (what we see things *as*).

One of my favorite targets has been the various regularities observed in experiments in which subjects use mental imagery. Examples include the increase in time required to report a feature in a mental image when that image is made smaller, or as the location of the feature is made further away from the point on the image where one is currently focused (for other such examples, see Kosslyn, 1980). Such regularities all appear to be cognitively penetrable — i.e. they can be changed in a rational way by changing what a subject takes the task to be, or what he believes would happen in the imagined situation. For example (cf Pylyshyn, 1984) if a subject is instructed to imagine a situation that he believes would involve an instantaneous switch of attention from one point of an image to another, the linear increase in reaction time with distance can be made to disappear. The conclusion in each of these cases is that the regularities obtain because the subject understands the task to be to reproduce what he believes would have happened had he been observing a real situation — and this he does very nicely because he often knows what would have happened, or can deduce it from general principles.

There have been a few interesting proposals for properties attributable to the central functional architecture. For example, there are proposals for the structure that can exist among attainable concepts — e.g. proposals by Osherson (1978) and by Keil (1979). There are even a few proposals for mechanisms involved in reasoning and problem solving, such as Newell's proposal for the primacy of the recognize-act cycle or some proposals for memory retrieval or property inheritance mechanisms. There are also proposals for resource-limited bottlenecks caused by the architecture, such as limits on working memory or on the number of internal referencing tokens available (cf Pylyshyn, forthcoming). On the whole, however, nothing comparable to Universal Grammar has been proposed for the central processing system. Whether this is because we are missing a critical idea or methodology, or because our search for mechanisms has been conditioned too much by current computers, or because, as Chomsky suggests, the problem really is beyond our capacity to solve, I can't say. I can say, however, that the enterprise of taking strong equivalence seriously in Cognitive Science — which, in my view, is tantamount to the assumption that there exists something that deserves to be called Cognitive Science — is very much dependent on finding such properties; of discovering cognitively impenetrable basic capacities.

Rules and Representation in Cognitively Impenetrable Modules

Representations of beliefs, goals and other propositional attitudes is only one of two areas where explicit symbolic encodings have been postulated. The other area is in the study of such modular processes as those involved in language processing. In this case much of the process appears to be cognitively impenetrable, and therefore one of the principle reasons for inferring that they involve reasoning is not available. However, as I said earlier, cognitive penetrability,

because it entails reasoning processes, is a necessary but not a sufficient reason for inferring that a process involves the manipulation of explicit representations. There are other kinds of representation-governed processes besides reasoning; and for these we require other sources of evidence.

The outstanding example of an impenetrable (or "encapsulated") process is the stage of language processing known as parsing — i.e. the stage at which only grammatical rules are brought to bear to extract the thematic structure (or the "logical form") of a sentence. At this stage the process involves operations that analyze a sentence in accordance with the structures given by a grammar. Some grammars describe the structures in terms of rules (such as phrase structure rules) whereas other describe them in terms of constraints and conditions. In either case, the question remains: need we claim that the rules or principles are represented, in the sense of being explicitly encoded and accessed in the course of parsing.

The answer is far from obvious. There are some reasons for thinking that the rules are at least individuated, if not explicitly encoded. This is the argument made by Pinker and Prince (1988), in their critical analysis of a Connectionist model for the acquisition of past test morphemes, a model which quite deliberately eschews individual rules. As Pinker & Prince (1988) put it,

> … rules are generally invoked in linguistic explanations in order to *factor* a complex phenomenon into simpler components that feed representations into one another. … Rules are individuated not only because they compete and mandate different transformations of the same input structures (such as *break — breaked/broke*), but because they apply to different *kinds* of structures, and thus impose a factoring of a phenomenon into distinct components. Such factoring allows orthogonal generalizations to be extracted out separately, so that observed complexity can arise through the interaction and feeding of independent rules and processes…(p84)

The need to individuate rules has also been defended on a number of other grounds. For example, there have been based on the need to explain why there is a rough synchrony in the cross-over from the acquisition of rules in comprehension to their use in production, based on the convergence of rules inferred from judgments and those that appear to be needed for parsing or production, or (as suggested by Fodor, Bever and Garrett, 1974) based on the observation that people can sometimes fail to follow some particular rules, based on the fact that certain universals appear to be stateable only over rules of a certain form, or that the parameters which specify particular languages are specific to certain formulations of grammar.

To my knowledge, however, there have been no arguments that the rules of language (i.e. rules of phonology, morphology, or syntax) are explicitly encoded. Indeed, the recent trend in linguistics has been to play down the role of rules in favor of general principles. As I have suggested, individuating rules or even principles is a different matter from claiming that they are explicitly encoded. The latter assumes that the symbolic expression of the rule or principle in

some canonical notation is empirically significant — e.g., that the expression itself is mapped onto the physical system in a way that preserves its structure.[12]

Although there may be some doubt as to whether grammatical rules are explicitly encoded, there appears to be good evidence that both the output of the analysis (i.e. LF) and certain intermediate states in the parsing *are* so encoded. These have to do with the fact that certain universal properties of the language faculty appear to be stateable only in terms of certain properties of the parse tree. For example, the various constraints on movement are stated in relation to certain properties of the analysis of the sentence, and thus imply that such an analysis is actually available to the system in the course of parsing and/or generation. Attempts to design parsing systems have also suggested that not only the logical form itself, but also various intermediate stages of the grammatical analysis may be explicitly encoded. In other words it is likely that parts of the analyzed structure of the sentence appears as a symbolic code, although the rules themselves may not appear in such a form. In computer jargon, although the rules may simply be compiled into the functional architecture of the system and not accessed in interpreted mode, the data structures to which these rules apply are explicitly represented and their form is empirically significant.

Conclusion: "knowledge" of rules

Finally, I want to comment on the use of the term "knowledge", especially in connection with Chomsky's use of the phrase "knowledge of the rules of grammar". Nobody has a proprietary right to the term. In fact it is widely used in a variety of ways. For example, it is frequently used as synonymous with "belief." Some people in Artificial Intelligence (e.g. Fahlman, 1981) even speak of "implicit knowledge" in referring to properties of the functional architecture. This would be harmless if we still had a way of distinguishing between the strong and the weak sense of rule-governed or representation-governed processes. If when we say that someone *knows rule R* we do not imply that he has an encoded representation of R (i.e. the strong sense of rule following) we will have to invent a term meaning "has an encoded representation of R and uses it to generate instances of the behavioral regularity in question". My sense is that this is what the term means in the vernacular and I would make a plea to confine its use to this strong sense. Thus I would prefer not to speak of "knowledge of the rules of grammar," or "knowledge of the principals of U.G.," though I *would* speak of "knowledge of what a person meant (in uttering a particular sentence or sentence fragment)", or the knowledge that certain coreferential relations hold between parts of a sentence.

It seems to me that when I say that someone *knows P* or *believes P*, I intend you to understand that the same person, with exactly the same cognitive capacity, might not have known P, but might have believed something else, say Q which might even entail not-P, and furthermore that it is even possible that he might yet come to believe Q under the right set of

12. Note that the question of whether the rules and/or principles are explicitly encoded is an empirical one, not one of principle. There is no problem in imagining a system which explictly encodes, say the principles of GB theory, and carries out the parsing by referring to this encoding. Indeed, my colleague Ed Stabler has designed a system that runs on a computer and does exactly that for Chomsky's (1986) syntactic theory (Stabler, in press).

circumstances (i.e. given the right data). Of course we all know that people's beliefs cannot be changed willy-nilly, but it is part of our understanding of human nature and of rationality that under ideal information conditions people like you and I could come to believe pretty near the same things that I and you believe, respectively. Hope in the human race, you see, springs eternal in each of us.

References

Anderson. J.R. (1978). Arguments concerning representations for mental imagery. *Psychological Review, 85*, 249-277.

Brewer, W.F. There is no convincing evidence for operant or classical conditioning in adult humans. In W.B. Weimer and D.S. Palermo, eds. *Cognition and the Symbolic Processes*. Hillsdale, N.J.: Erlbaum, 1974.

Chomsky, N. (1986a) *Knowledge of Language: Its Nature, Origin, and Use*. New York: Praeger.

Chomsky, N. (1986b). *Barriers*. Cambridge, Mass: MIT Press.

Chomsky, N. (1984) Changing perspectives on knowledge and use of language. ms.

Fahlman, S.E. (1981). Representing implicit knowledge. In G.E. Hinton & J.A. Anderson (eds), *Parallel Models of Assocoateive Memory*. Hillsdale, N.J.: Erlbaum.

Fodor, J.A., Bever, T., and Garrett, M. *The Psychology of Language*. New York: McGraw-Hill, 1974.

Fodor, J.A. (1976) *The Language of Thought*. Sussex: Harvester Press.

Fodor, J.A. and Pylyshyn, Z.W. (1988). Connectionism and cognitive architecture: A critical analysis. *Cognition, 28*, 3-71.

Keil, F. (1979). *Semantic and Conceptual Development: An Ontological Perspective*. Cambridge, Mass: Harvard University Press.

Kosslyn, S. (1980). *Language and Mind*. Cambridge, Mass: Harvard University press.

Osherson, D. (1978). Three conditions on conceptual naturalness. *Cognition, 6*, 263-289.

Pylyshyn, Z.W. (1979) Validating computational models: A critique of Anderson's indeterminism of representation claim. *Psychological Review, 86*, 383-394.

Pylyshyn, Z.W. (1984) *Computation and Cognition: Toward a Foundation for Cognitive Science*. Cambridge, Mass: MIT Press (Bradford Books).

Pylyshyn, Z.W. (forthcoming). The role of location-indexes in spatial perception.

Stabler. E. (in press). *The Logical Approach to Syntax.* Cambridge, Mass: MIT Press (A Bradford Book).